NYU CDS, SPRING 2018

INTRODUCTION TO DATA SCIENCE

OFFICE HOURS AND CONTACT INFORMATION

- CDS, 60 5th Avenue, 6th floor, room 620
- Wednesday 12-2pm: Lecturer, Iddo Drori, idrori@nyu.edu
- Tuesday 11am-1pm: Section Leader, Datta Sainath Dwarampudi, ddattasainath@nyu.edu
- Friday 2-4pm: Grader, Samhita Damotharan, sd2941@nyu.edu
- Thursday 2-4pm: Grader, Sai Anirudh Kondaveeti, sak797@nyu.edu

NYU CLASSES

- Resources: lecture and lab notes
- Assignments: homework
- Forums: discussions

SUPERVISED LEARNING: CLASSIFICATION, FITTING

- Logistic regression & support vector machines: DSB Ch 4
- Overfitting: DSB Ch 5
- Bias and variance
- Regularization
- Classification in data science example

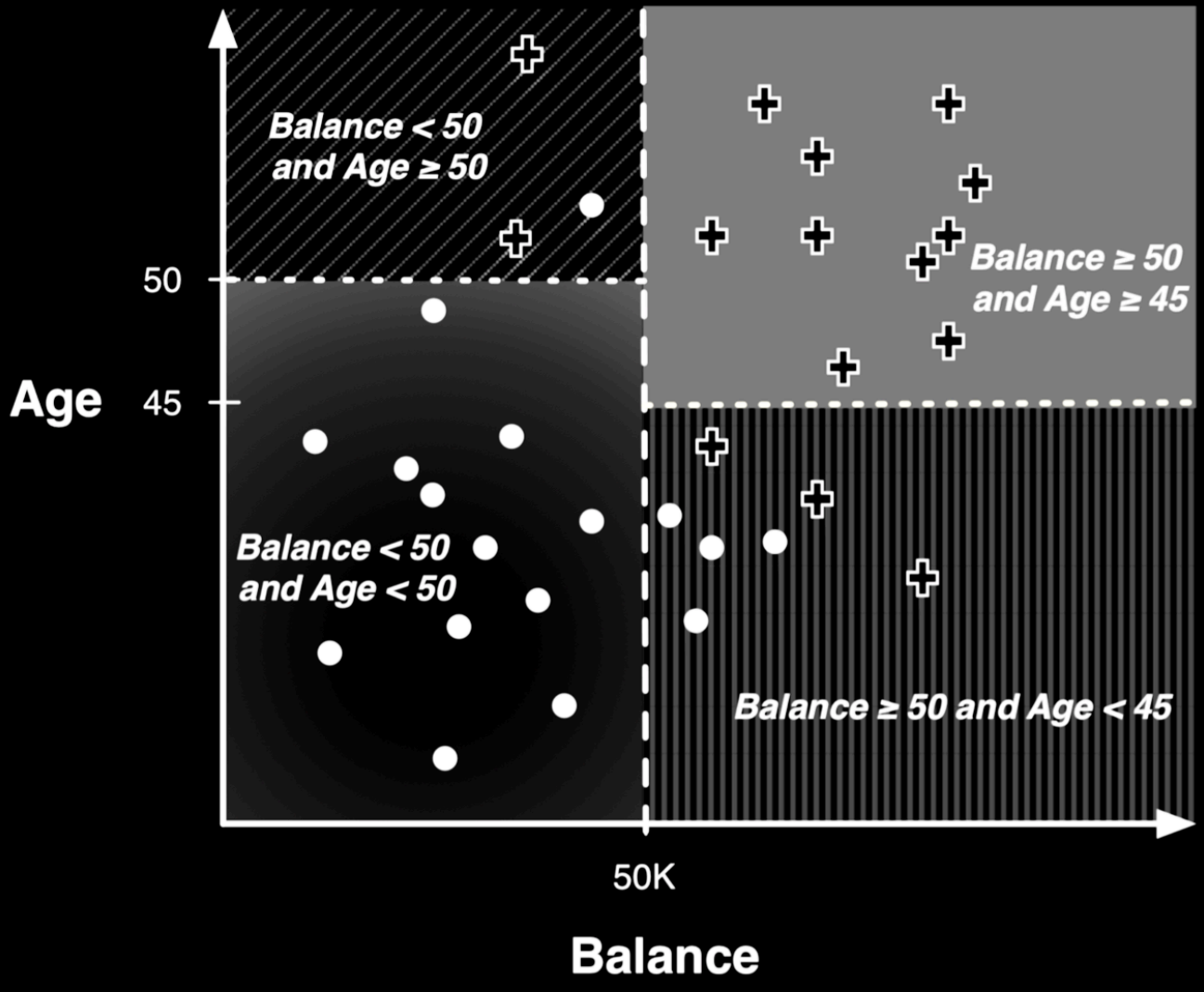
CLASSIFICATION

- Binary: 0/1
- Categorical: multiple classes

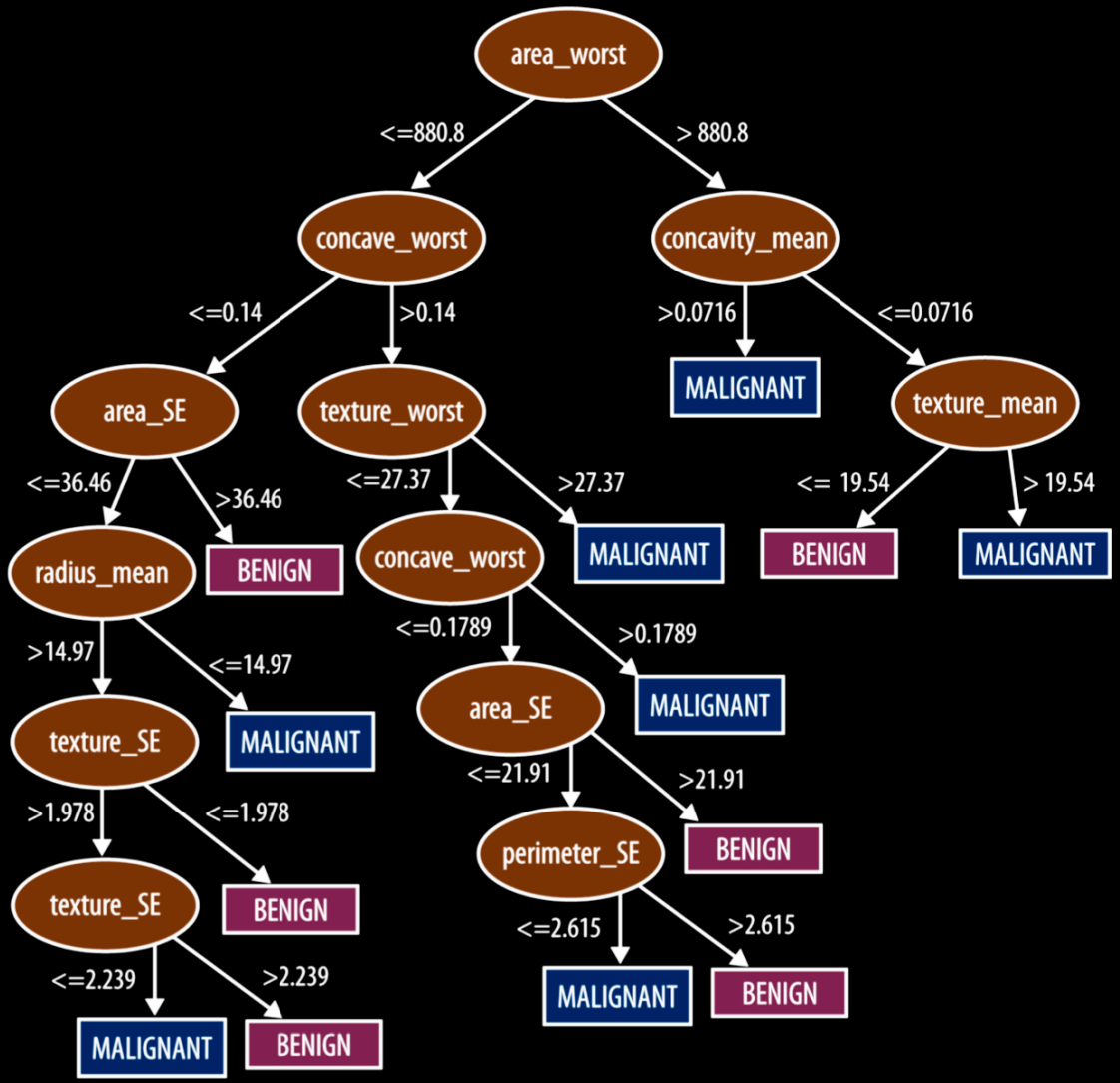
REGRESSION

- Estimate numerical value

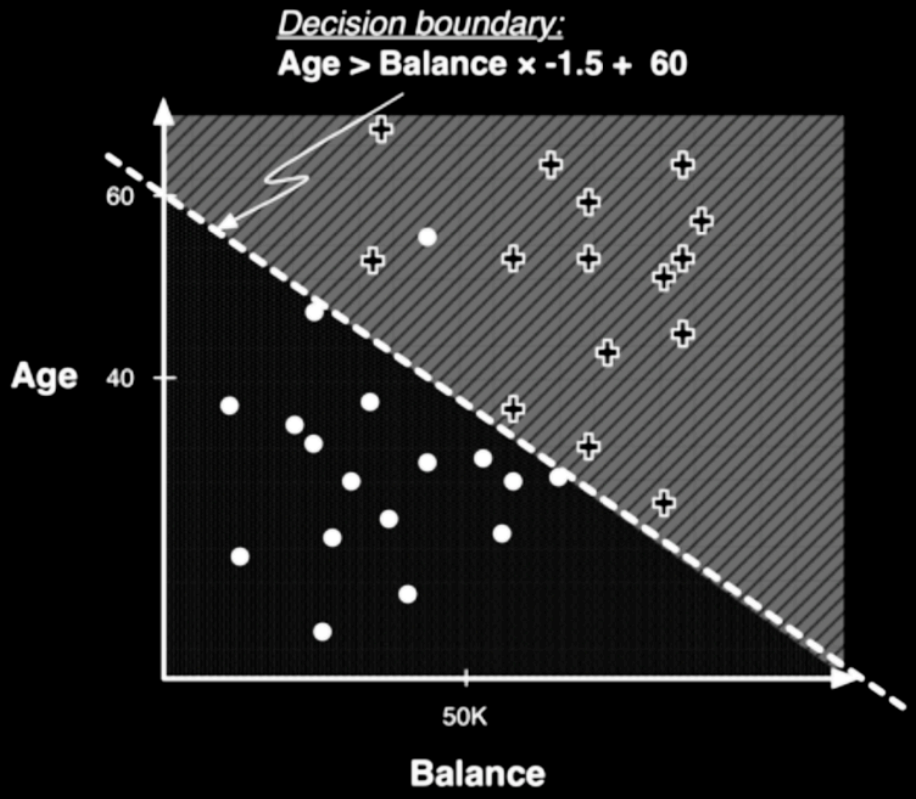
DECISION TREE



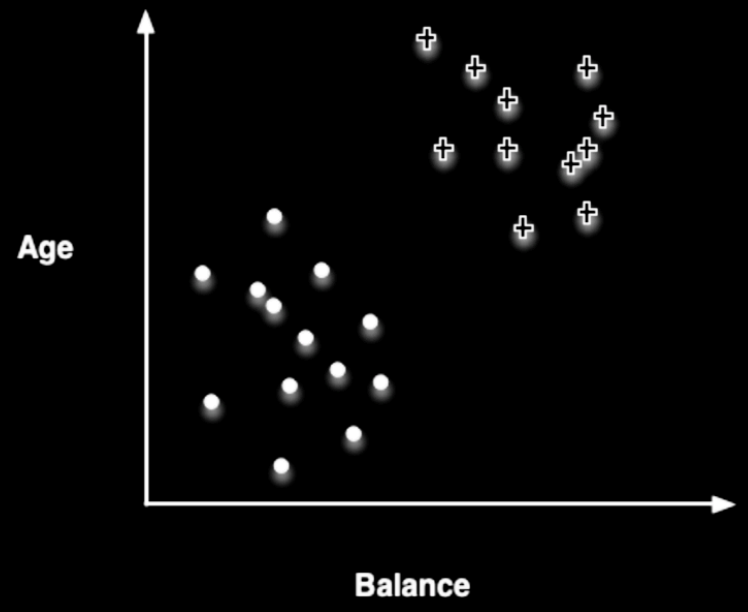
DECISION TREE



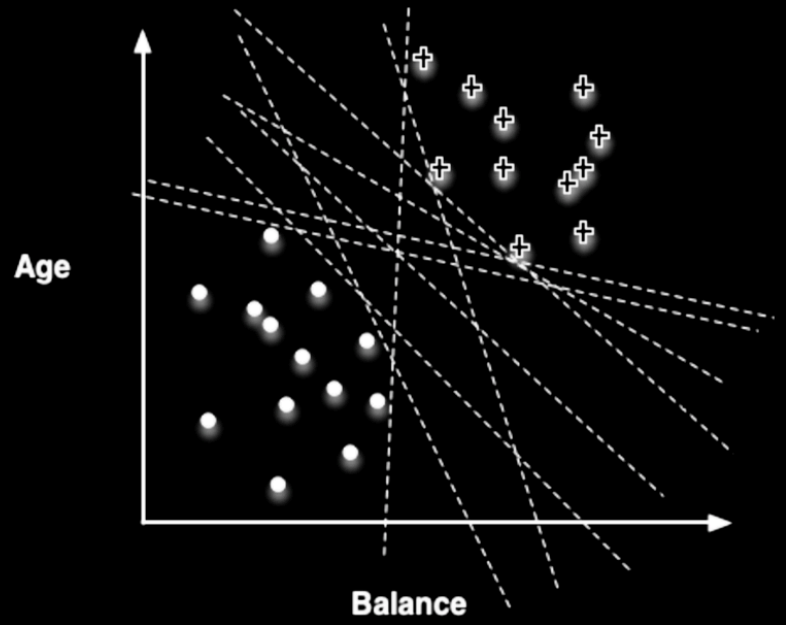
LOGISTIC REGRESSION



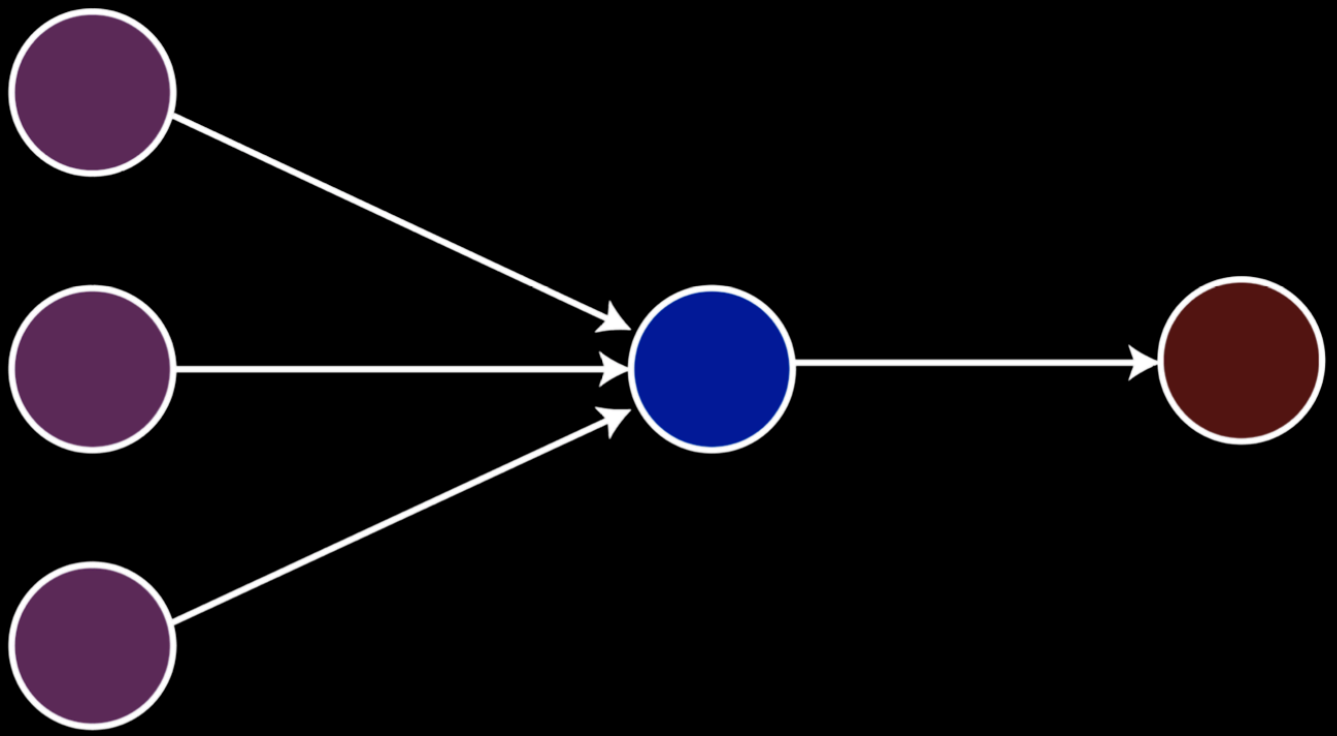
LOGISTIC REGRESSION



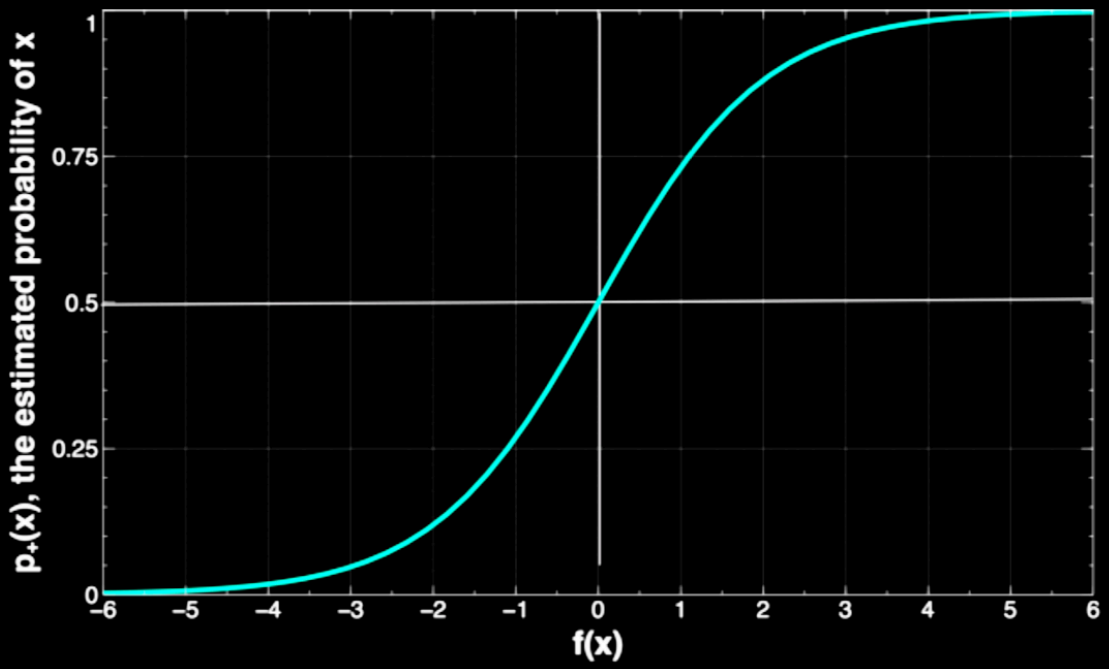
LOGISTIC REGRESSION



LOGISTIC REGRESSION

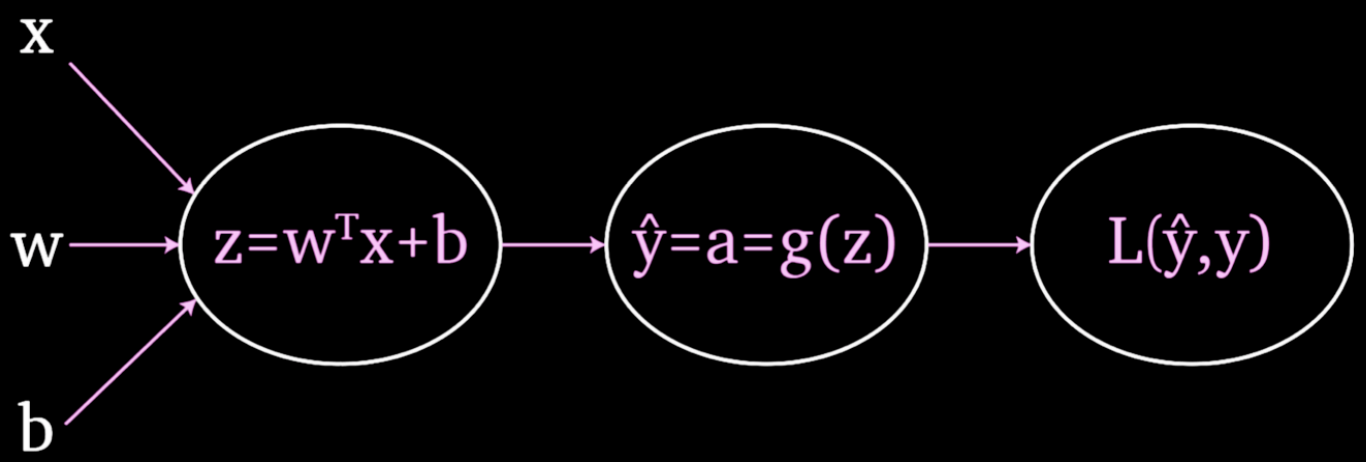


SIGMOID

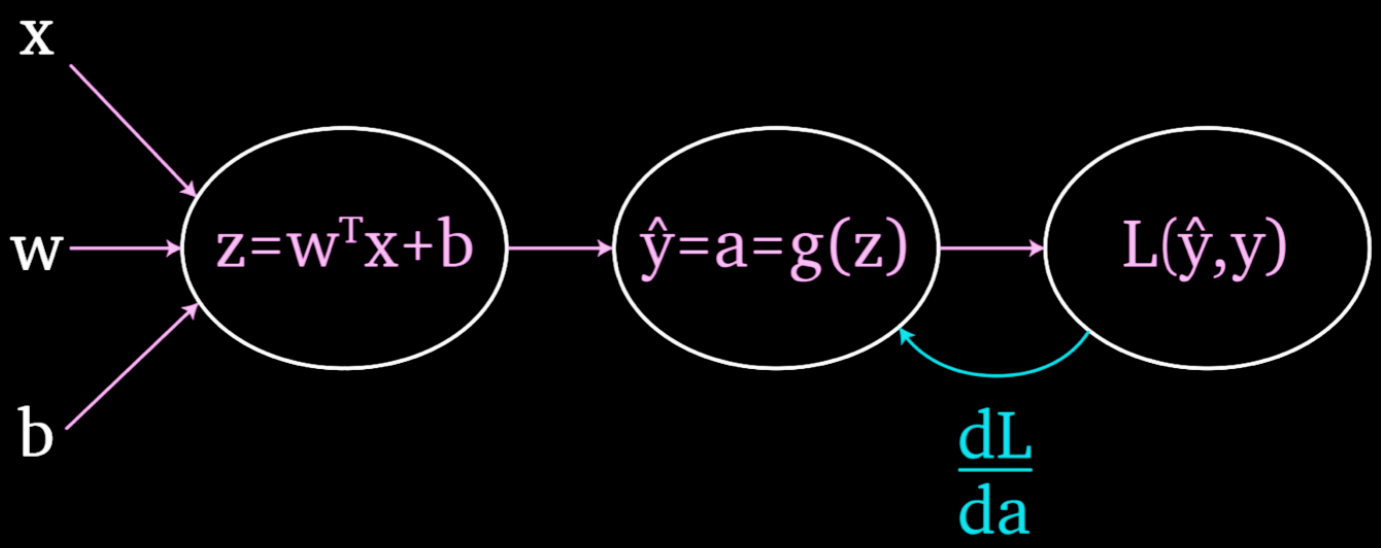


$$g(z) = \frac{1}{1 + e^{-z}}$$

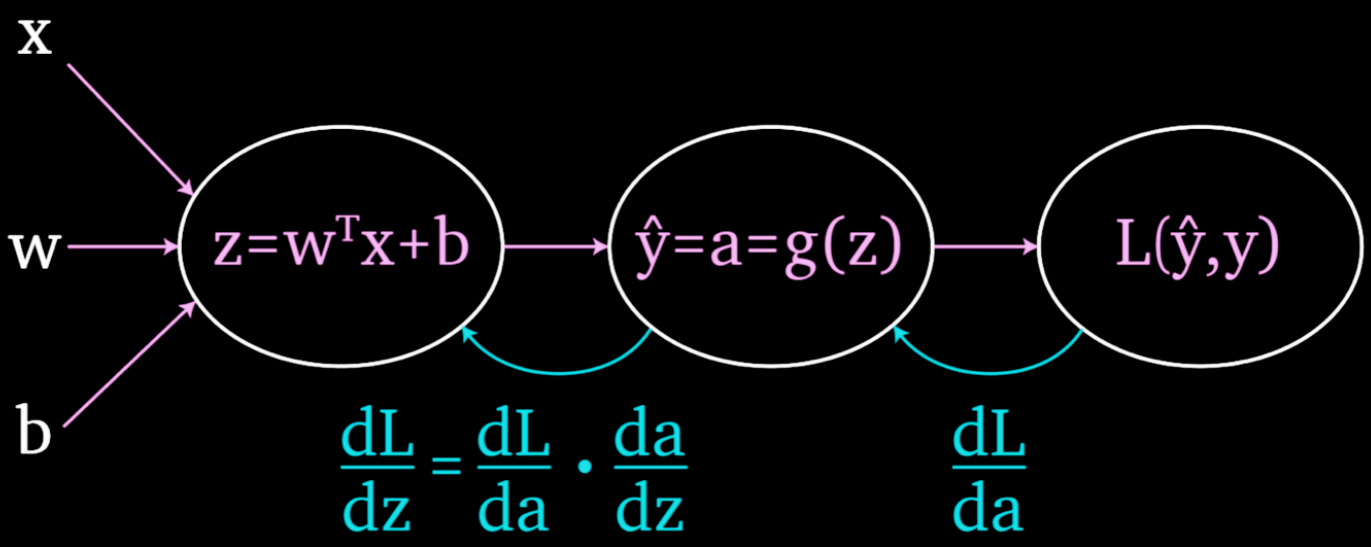
LOGISTIC REGRESSION: COMPUTATION GRAPH



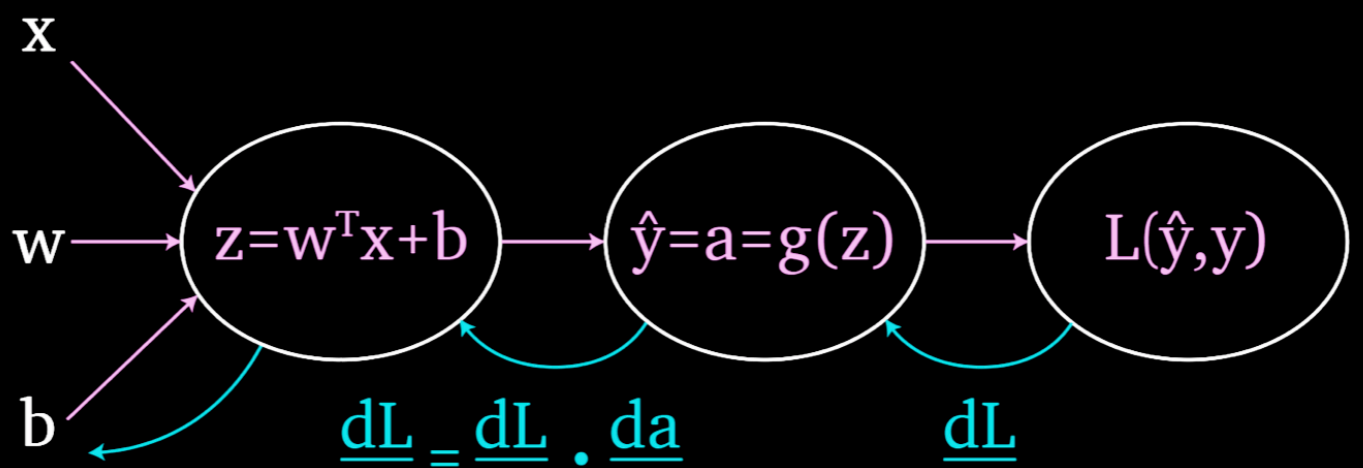
LOGISTIC REGRESSION: OPTIMIZATION



LOGISTIC REGRESSION: OPTIMIZATION



LOGISTIC REGRESSION: OPTIMIZATION



$$\frac{dL}{dz} = \frac{dL}{da} \cdot \frac{da}{dz} \qquad \frac{dL}{da}$$

$$\frac{dL}{dw} = \frac{dL}{da} \cdot \frac{da}{dz} \cdot \frac{dz}{dw} = \frac{dL}{dz} \cdot x$$

$$\frac{dL}{db} = \frac{dL}{da} \cdot \frac{da}{dz} \cdot \frac{dz}{db} = \frac{dL}{dz}$$

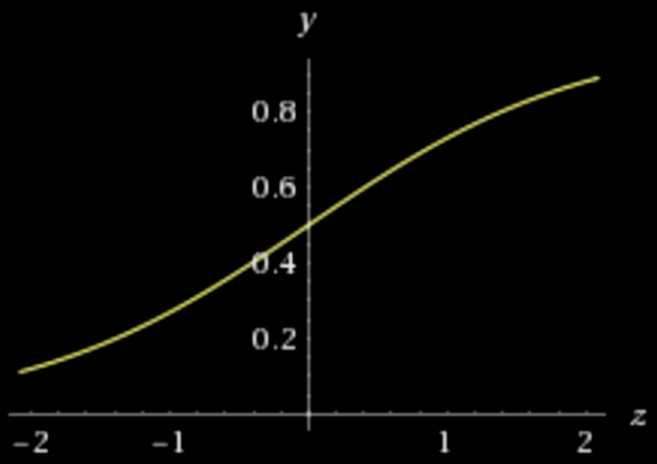
LOGISTIC REGRESSION

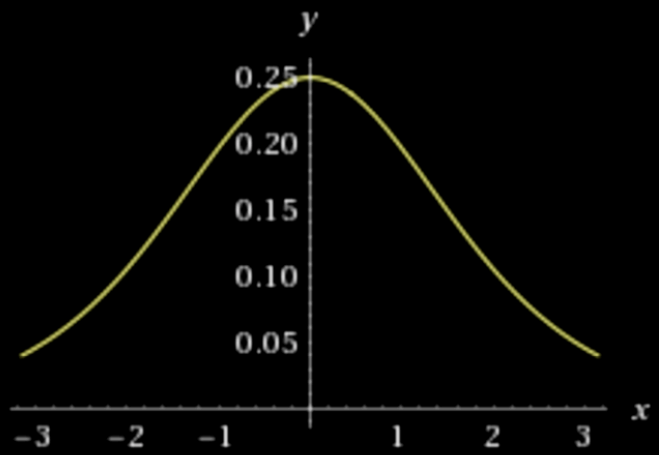
$p(x)$		$\log\left(\frac{p(x)}{1-p(x)}\right)$
Probability	Odds	Log-odds
0.5	50:50 or 1	0
0.9	90:10 or 9	2.19
0.999	999:1 or 999	6.9
0.01	1:99 or 0.0101	–4.6
0.001	1:999 or 0.001001	–6.9

LOGISTIC REGRESSION

- Linear classifier $f_w(x) = w^T x$
- Logistic regression $f_w(x) = g(w^T x)$
  - Absorb bias b in weight w and x
- Sigmoid $g(z) = \frac{1}{1+e^{-z}}$
  - Q: why sigmoid function?
  - A: set log-odds to be a linear classifier:
    - Setting $\log\left(\frac{g(x)}{1-g(x)}\right) = w^T x$
    - Results in sigmoid function g

LOGISTIC REGRESSION: SIGMOID AND DERIVATIVE





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

$$g'(z) = \frac{e^{z}}{\left(1+e^{z}\right)^{2}}$$

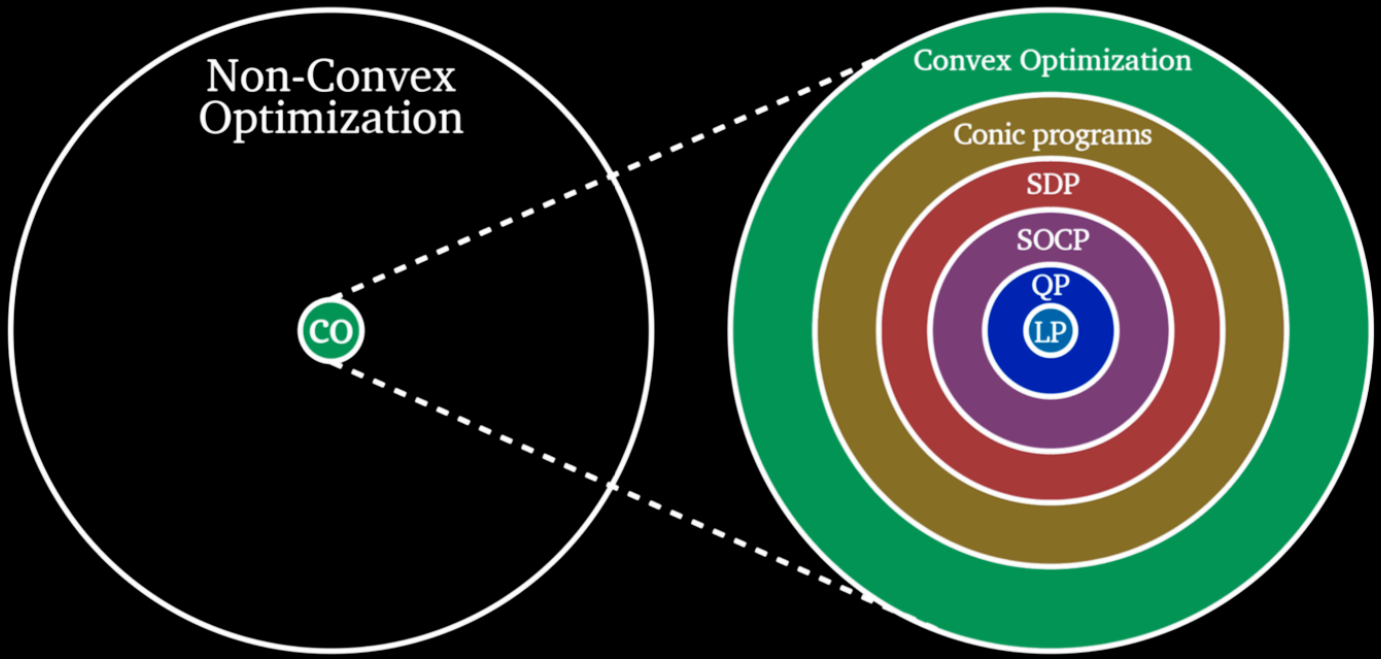
LOGISTIC REGRESSION

- Prediction $a = f_w(x) = g(w^T x)$
- Loss function L(y, a) between ground truth label y and prediction a
  - Log-loss $L(y,a) = L(y, f_w(x)) = -y\log(a) - (1-y)\log(1-a)$
- Cost: average loss over all examples i=1..n

$$J(w) = \frac{1}{n}\sum_{i=1}^{n} L(y^i, f_w(x^i))$$

- Find weights w to minimize cost J(w)

OPTIMIZATION



GRADIENT DESCENT

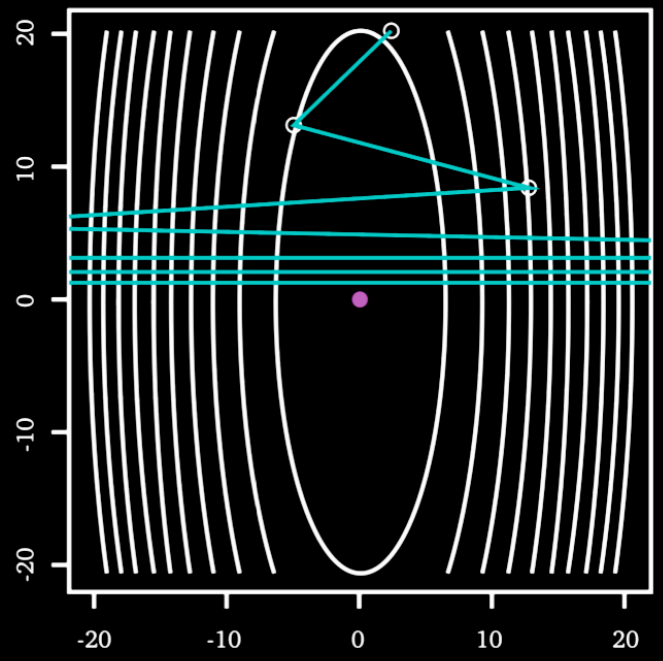
$$\underset{\mathbf{x}}{\textbf{minimize}} f(x) \tag{3.1}$$

Algorithm 1 Gradient descent

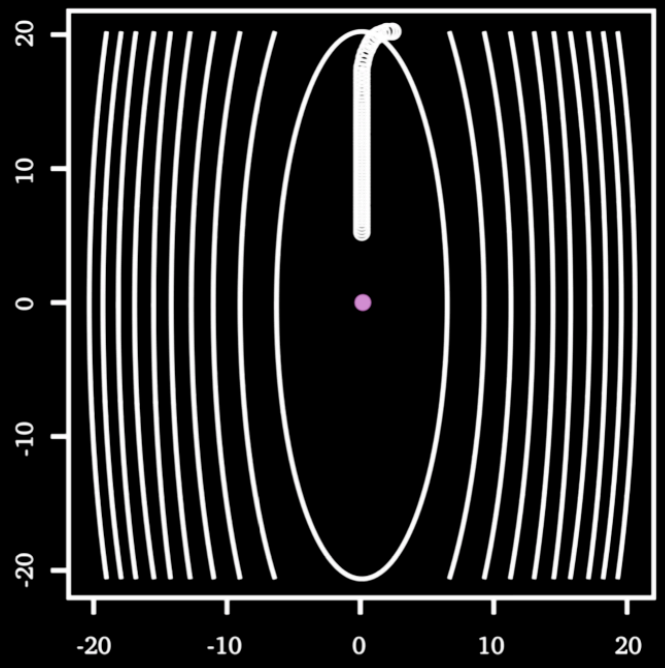
given a starting point $x \in \textbf{dom} f$
repeat
  determine descent direction $\Delta x = -\nabla f(x)$
  choose step size α
  update $x = x + \alpha \Delta x$
until stopping criterion is satisfied.

Simple, fast, well conditioned for convex problems.
Local minimum for convex problems.

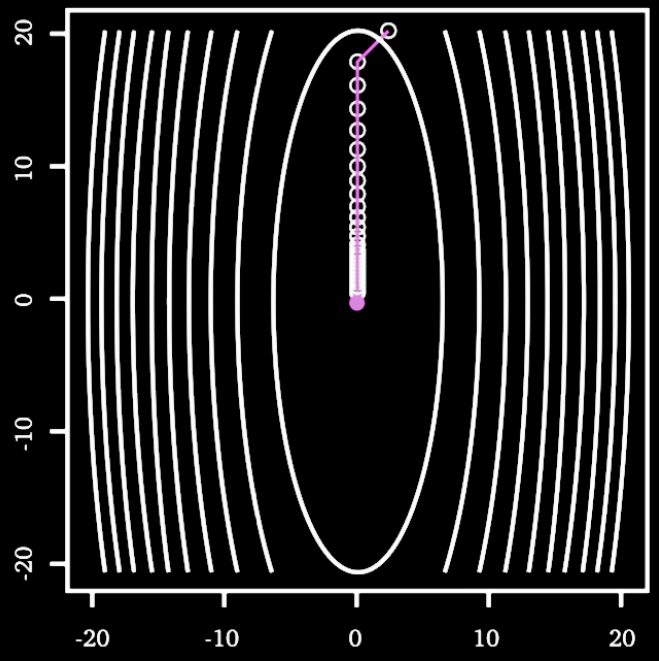
GRADIENT DESCENT



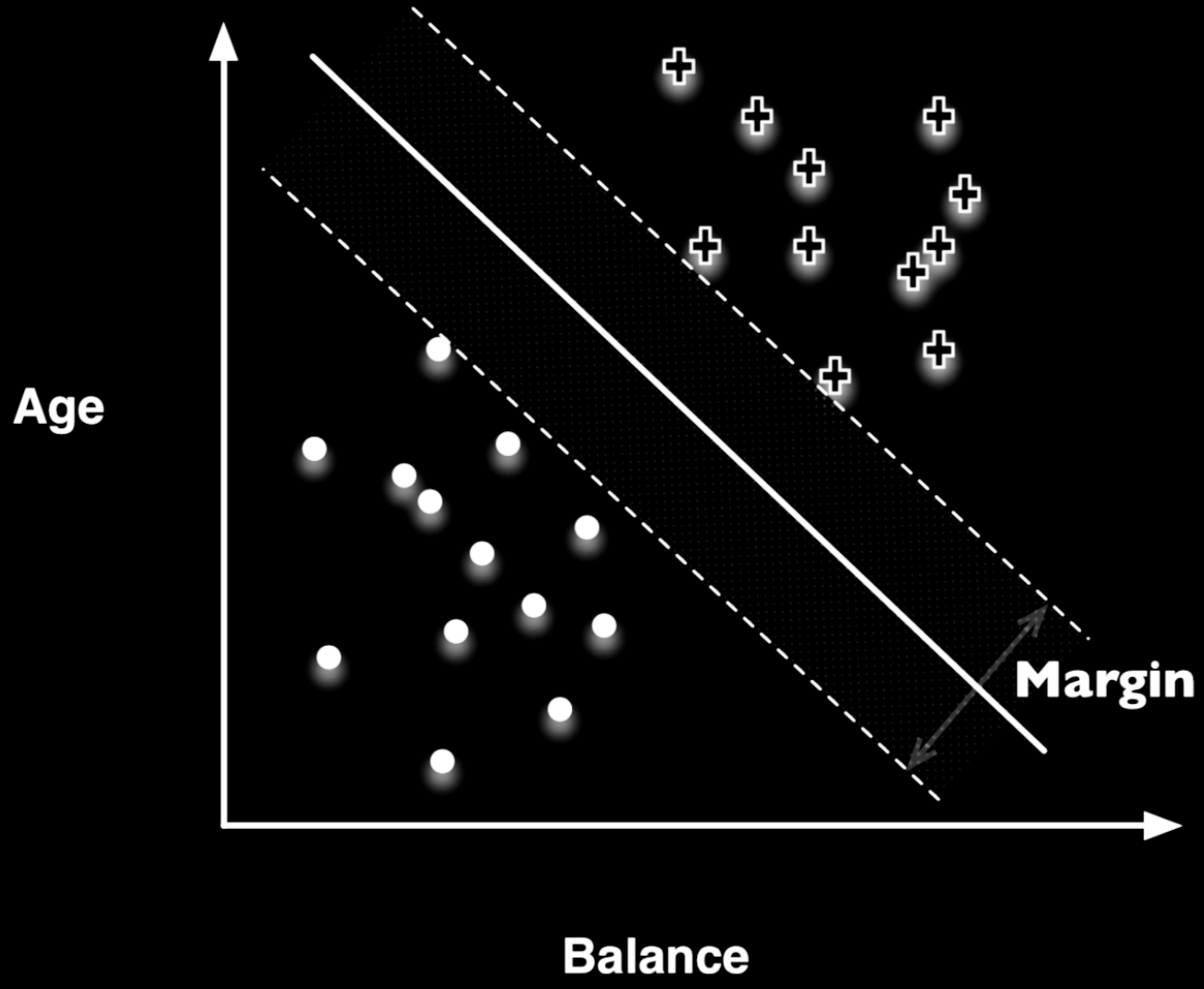
GRADIENT DESCENT



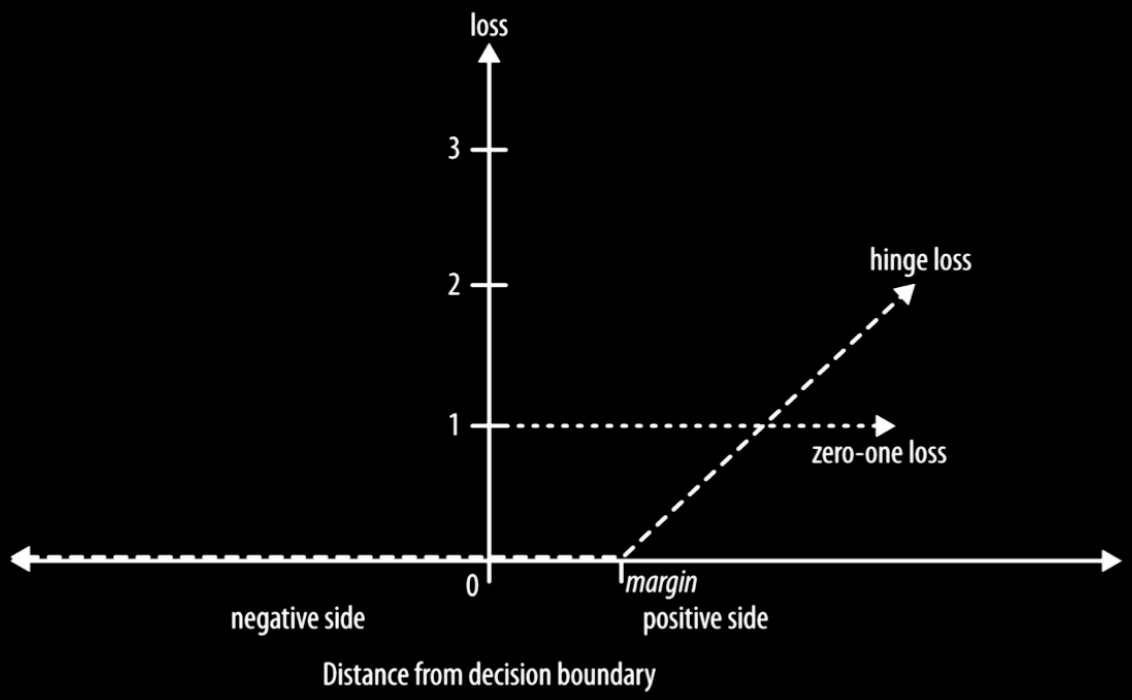
GRADIENT DESCENT



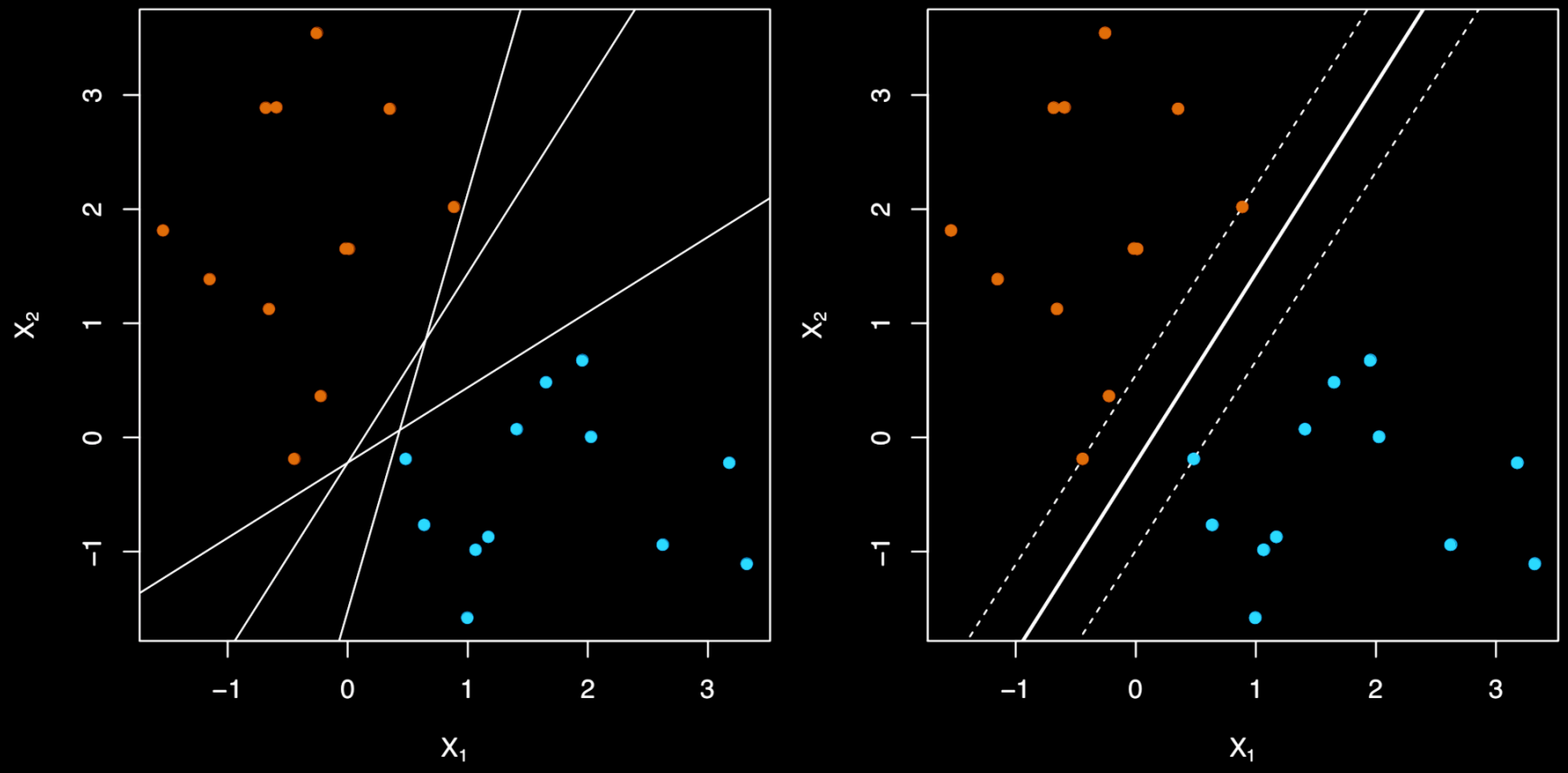
SUPPORT VECTOR MACHINES



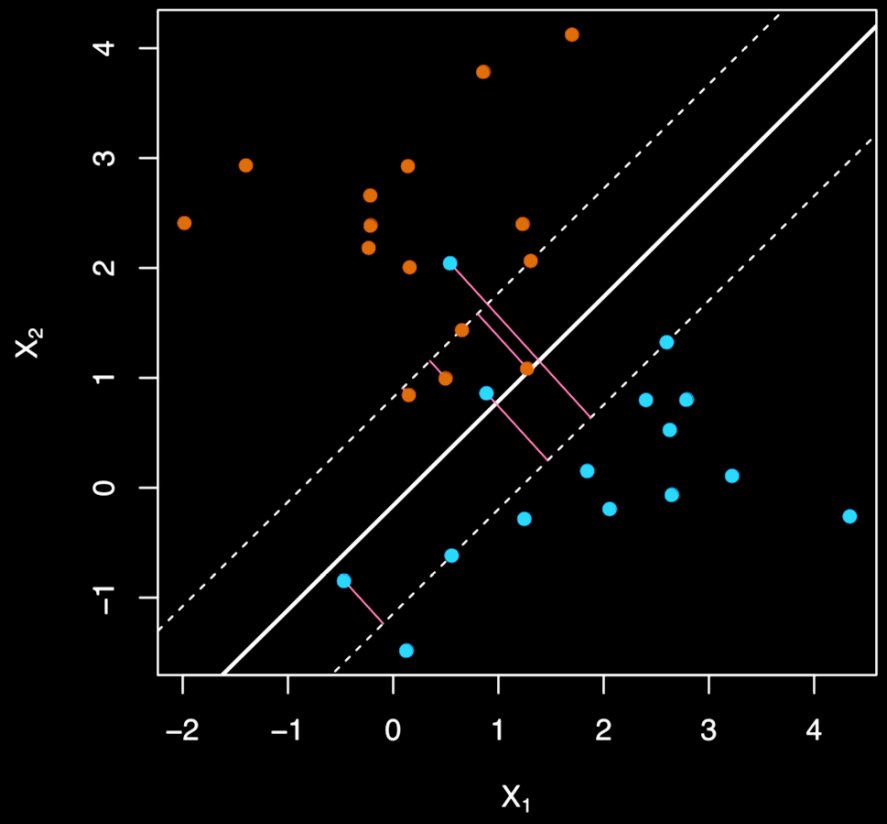
SUPPORT VECTOR MACHINE: HINGE LOSS

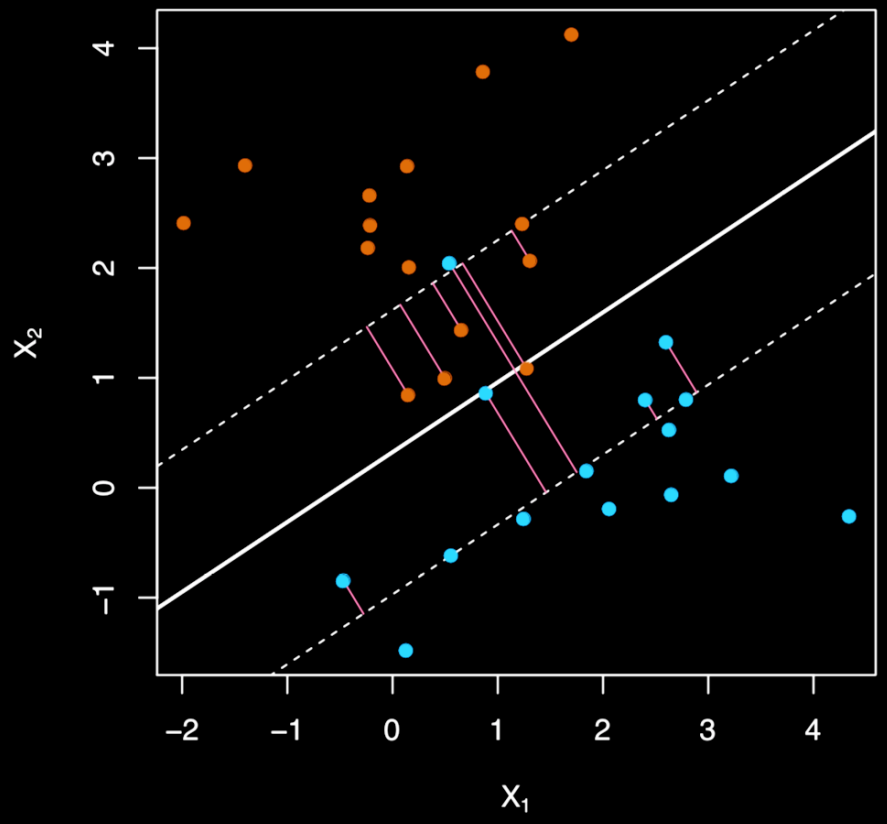


SUPPORT VECTOR MACHINES



SUPPORT VECTOR MACHINES





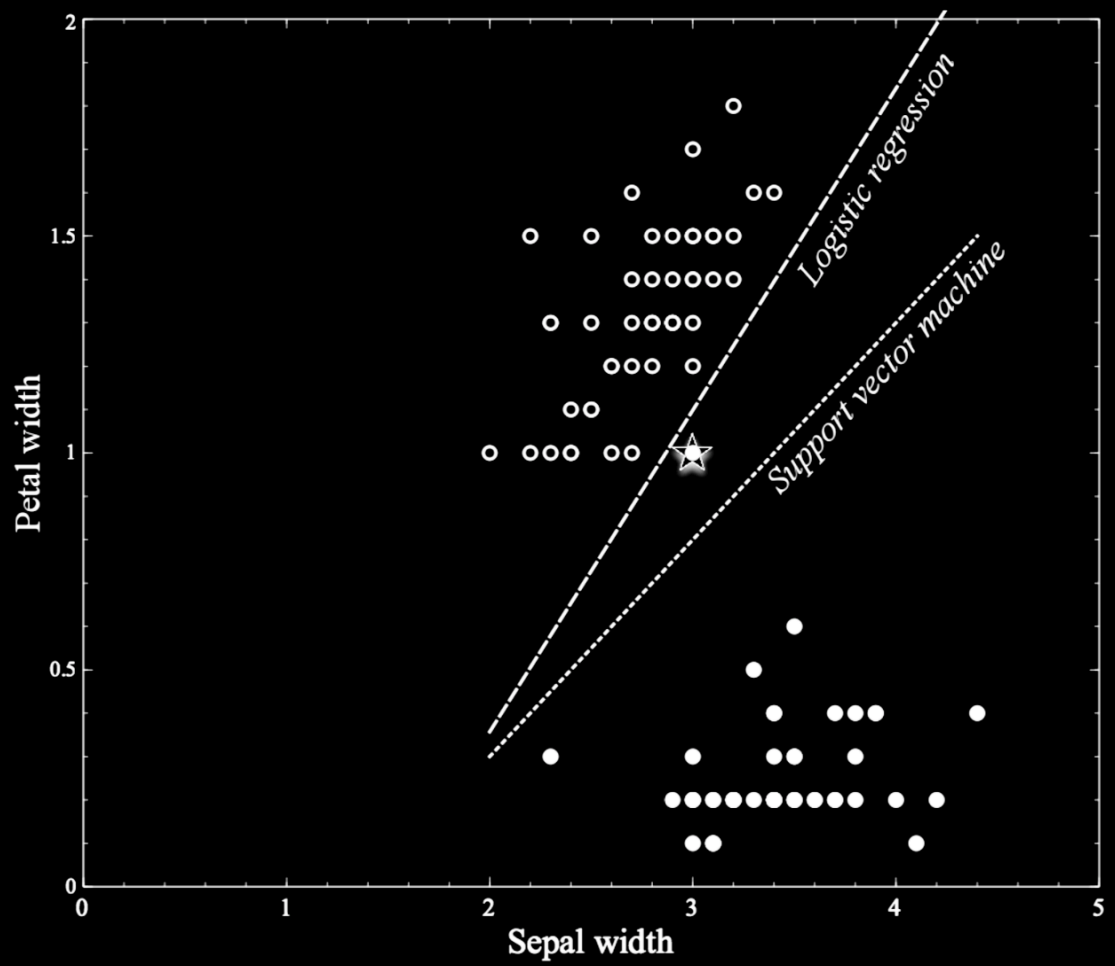
SUPPORT VECTOR MACHINES

- Different loss function: hinge loss

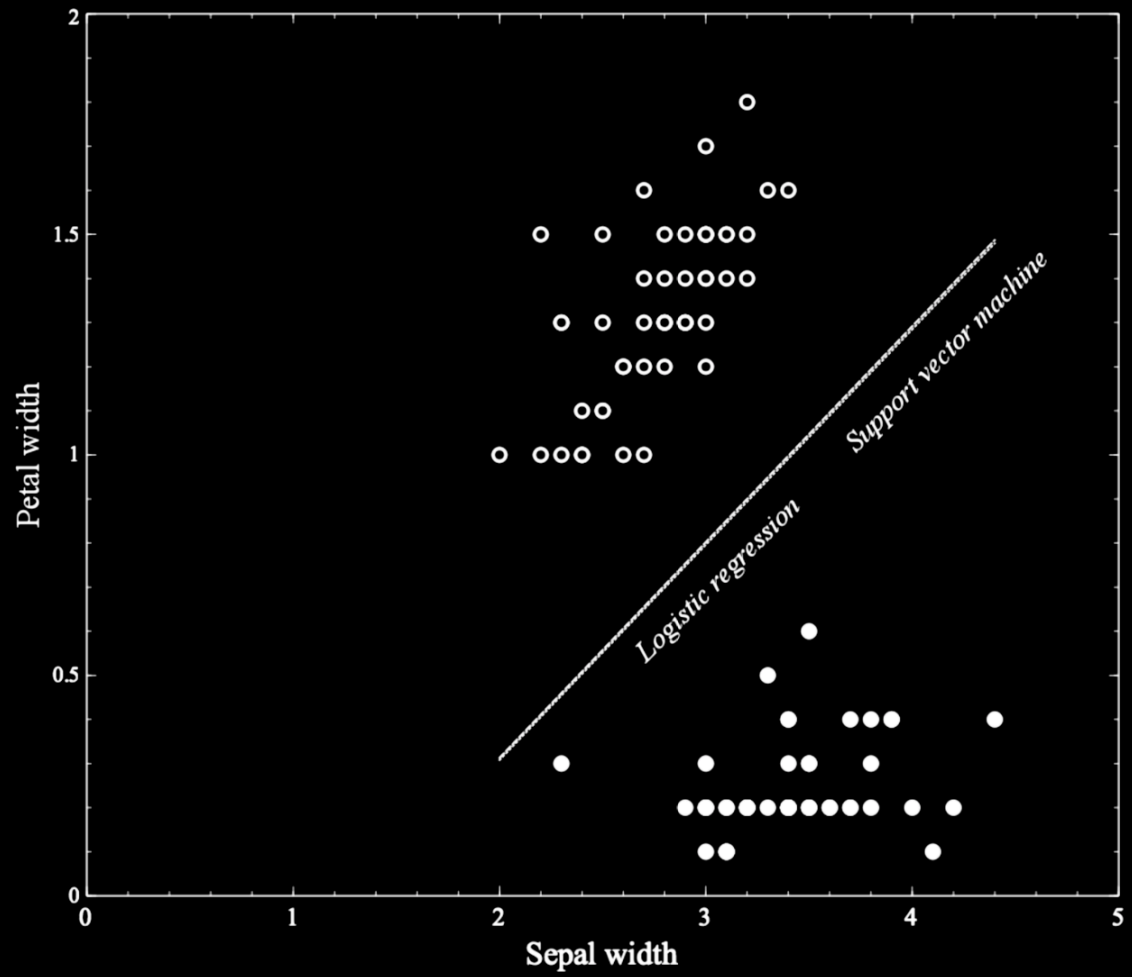
$$L\left(y, f\left(x\right)\right) = \max\left(0, 1 - yf\left(x\right)\right)$$

- Generalize to soft margin: allow points to violate margin.
  - Total violation allowed is a tunable parameter.

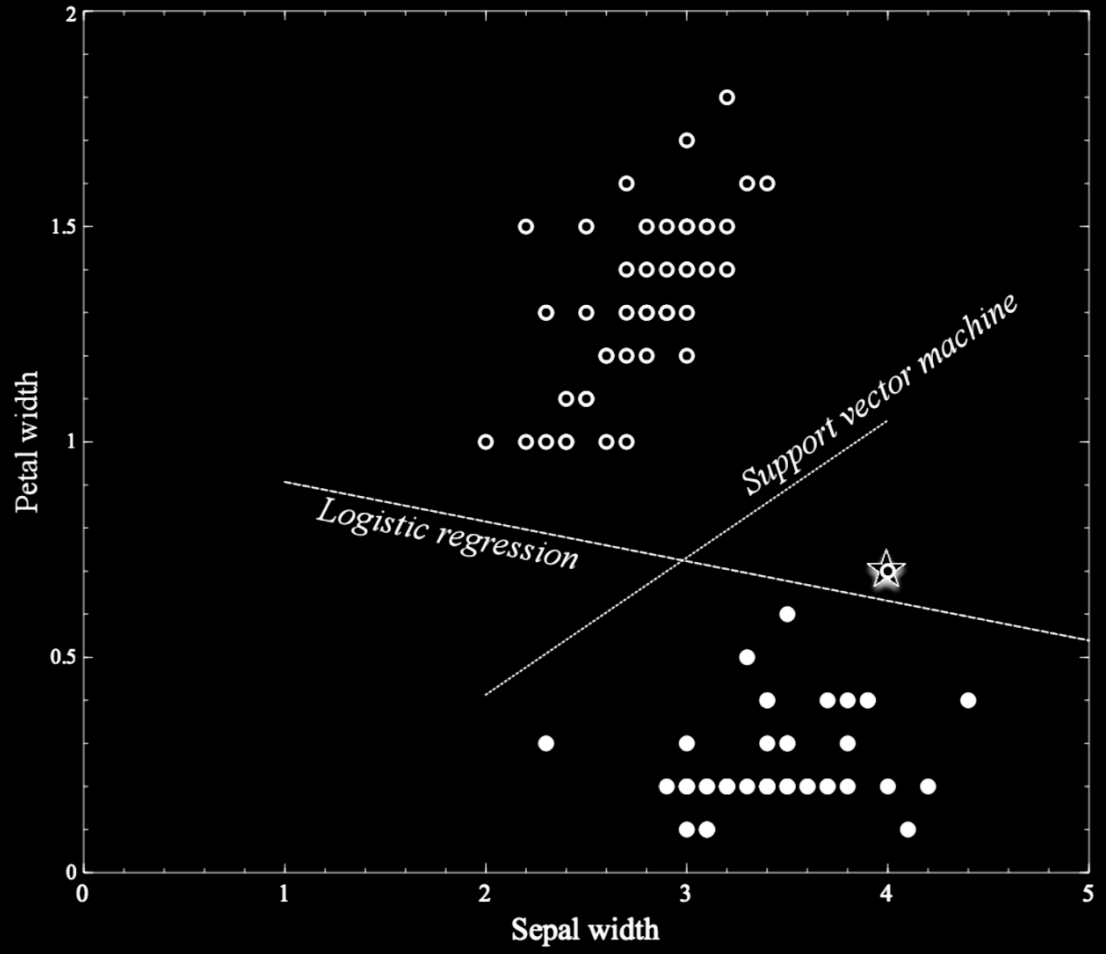
LOGISTIC REGRESSION VS. SUPPORT VECTOR MACHINE



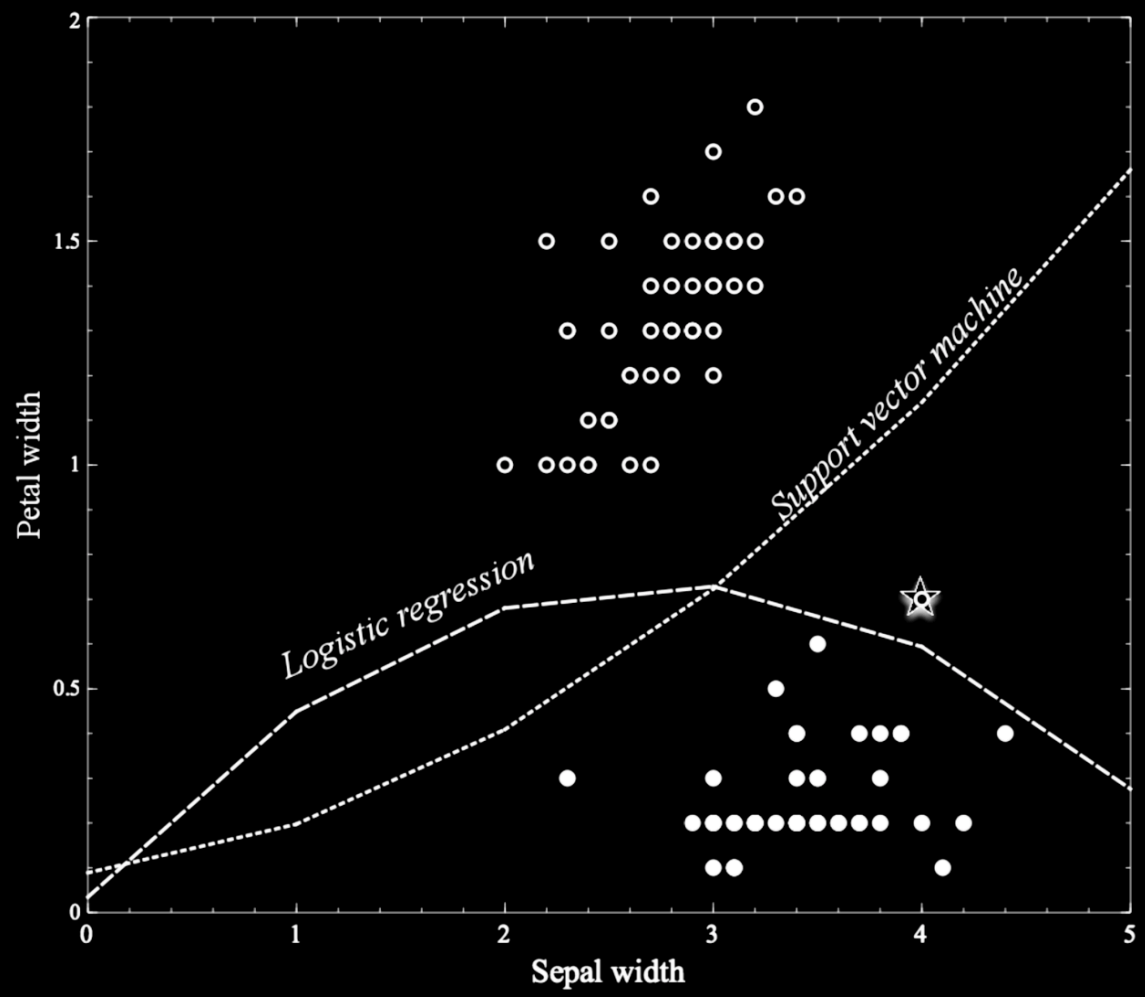
LOGISTIC REGRESSION VS. SUPPORT VECTOR MACHINE



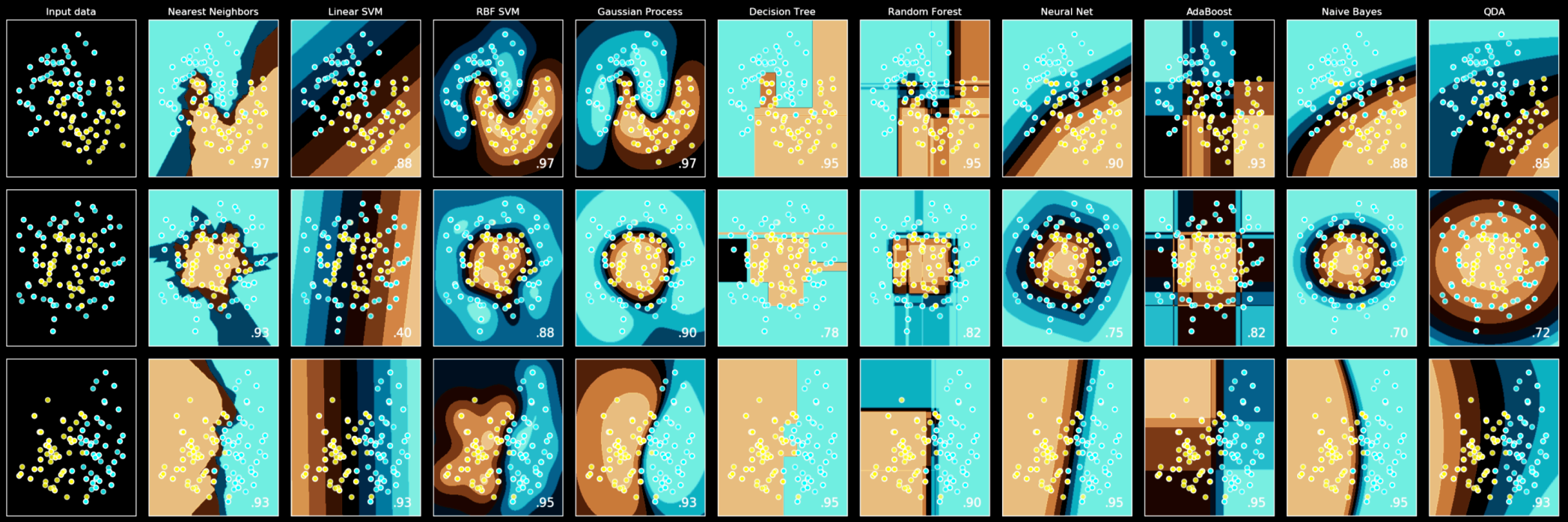
LOGISTIC REGRESSION VS. SUPPORT VECTOR MACHINE



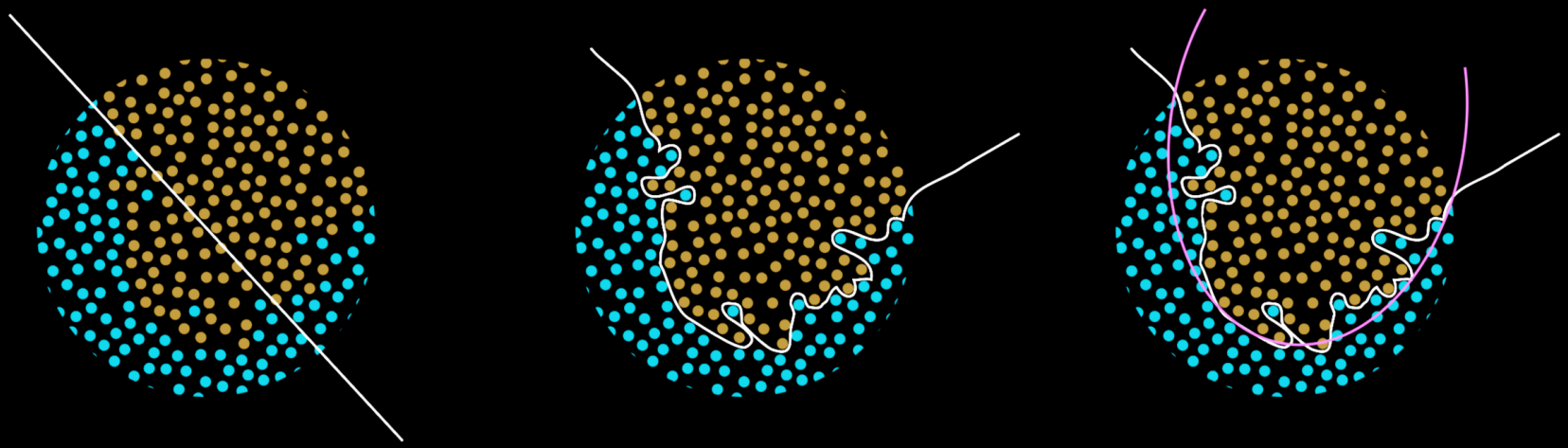
LOGISTIC REGRESSION VS. SUPPORT VECTOR MACHINE



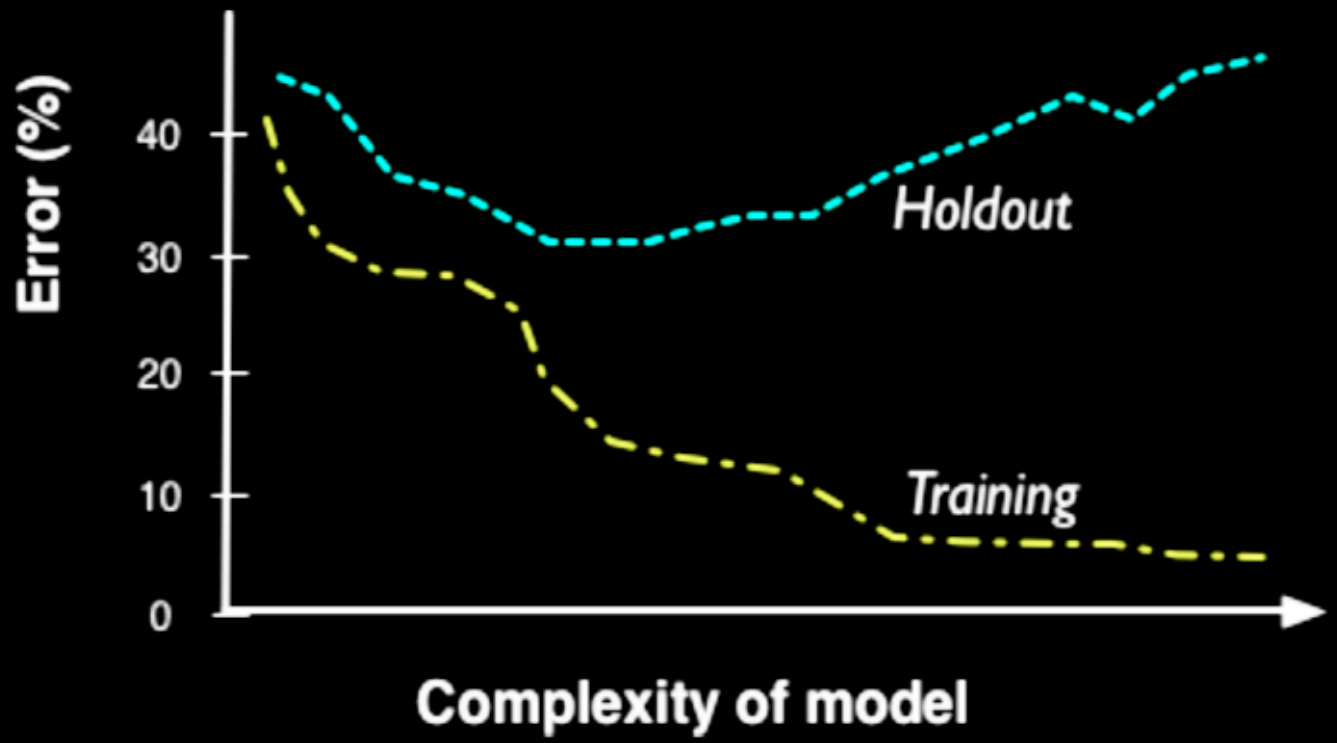
CLASSIFICATION: SCIKIT-LEARN



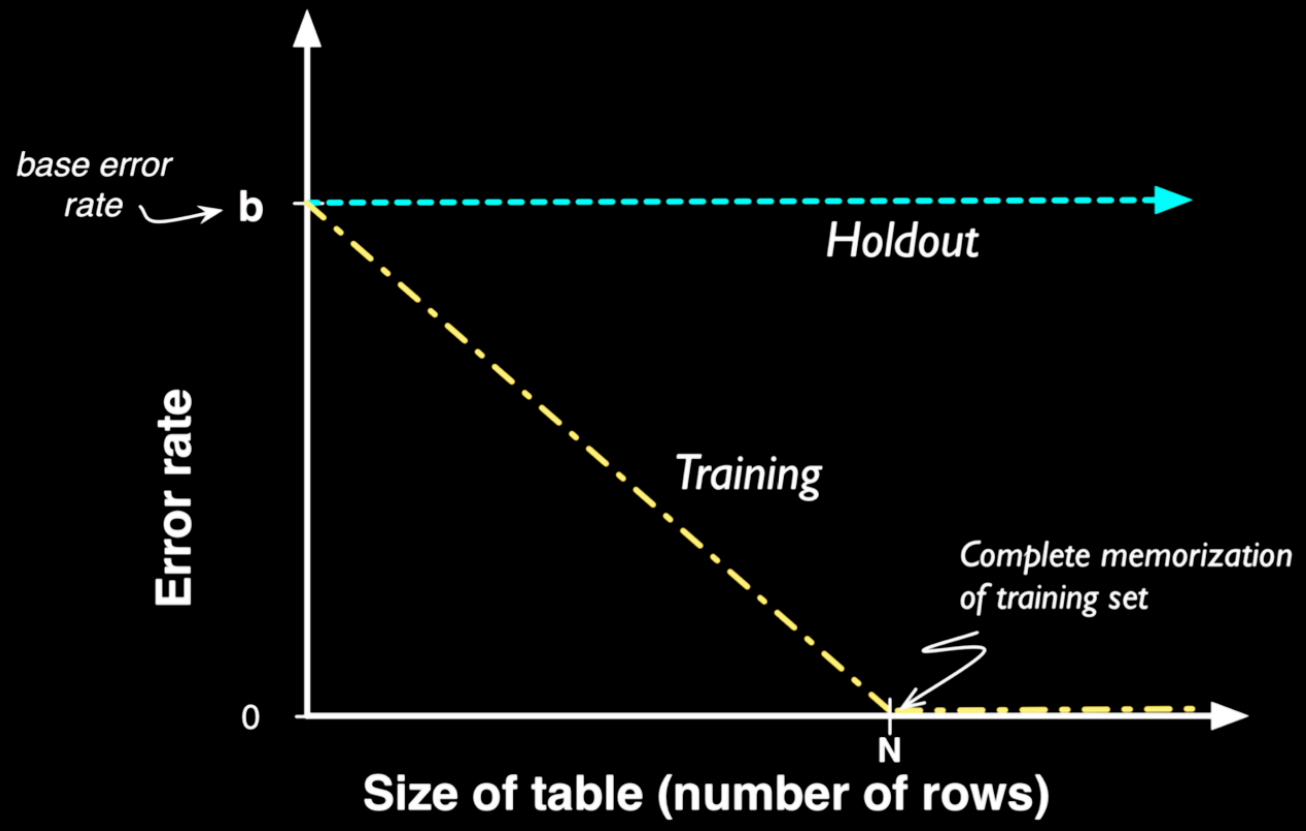
OVERFITTING



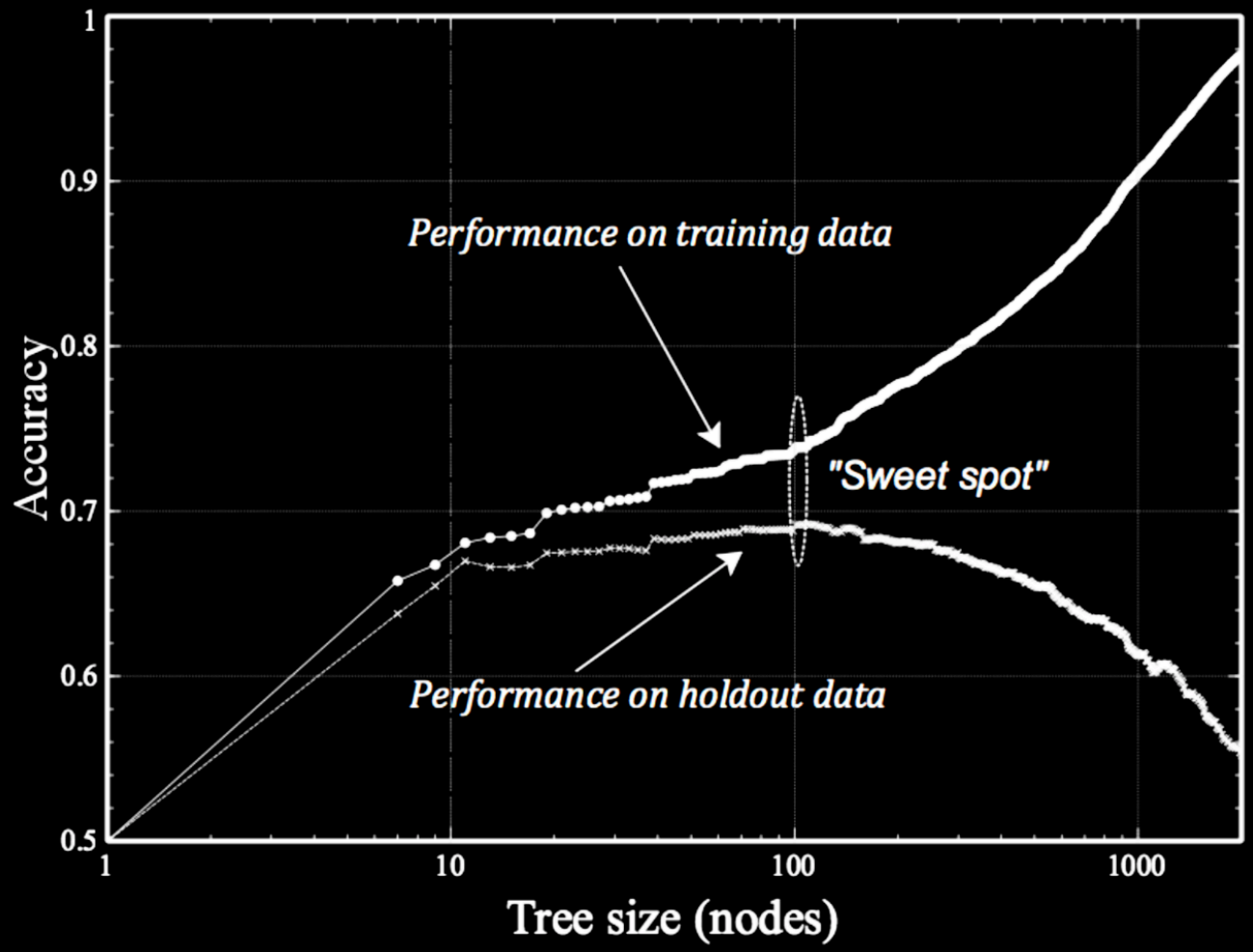
FITTING CURVES



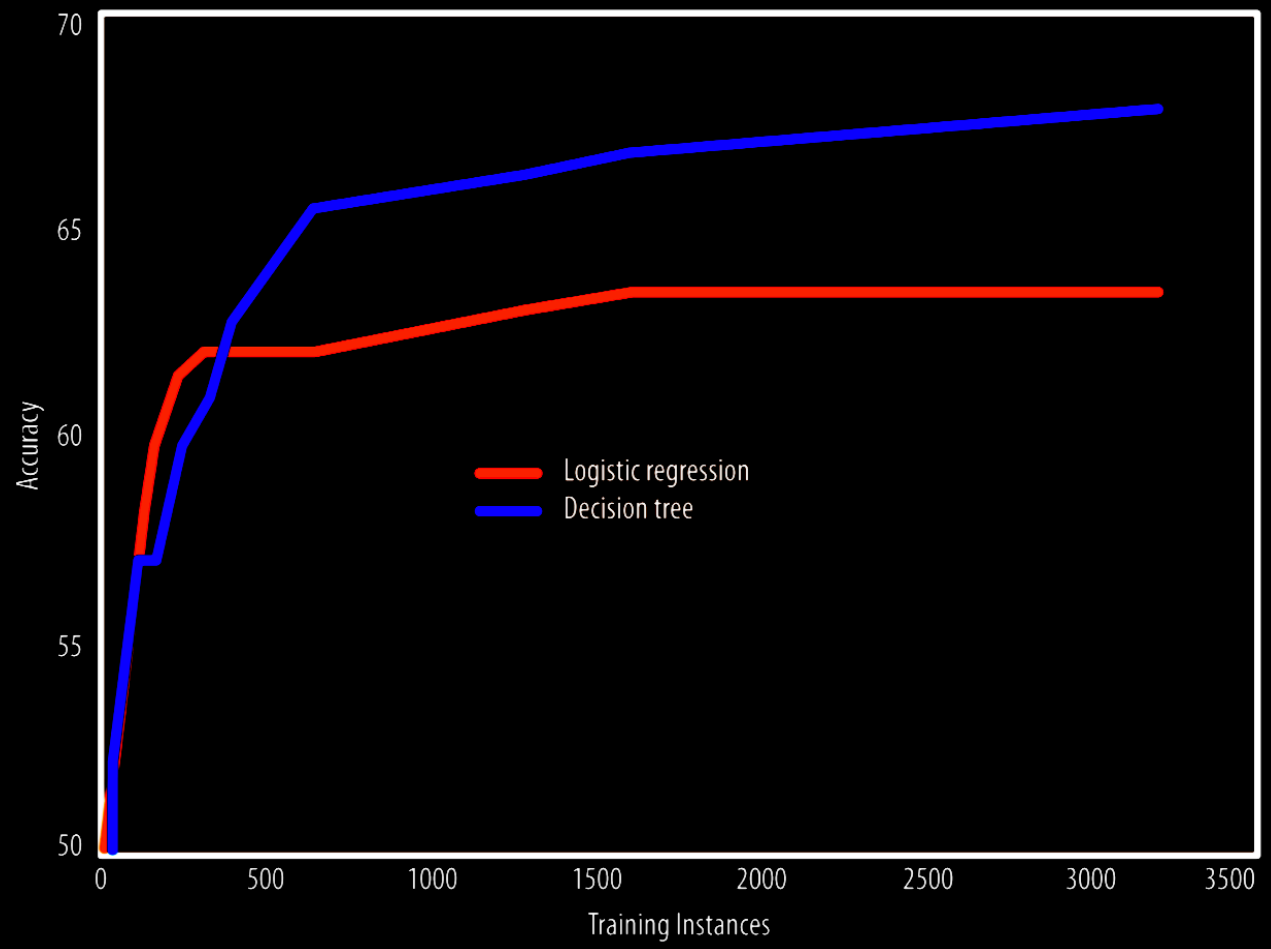
FITTING CURVES



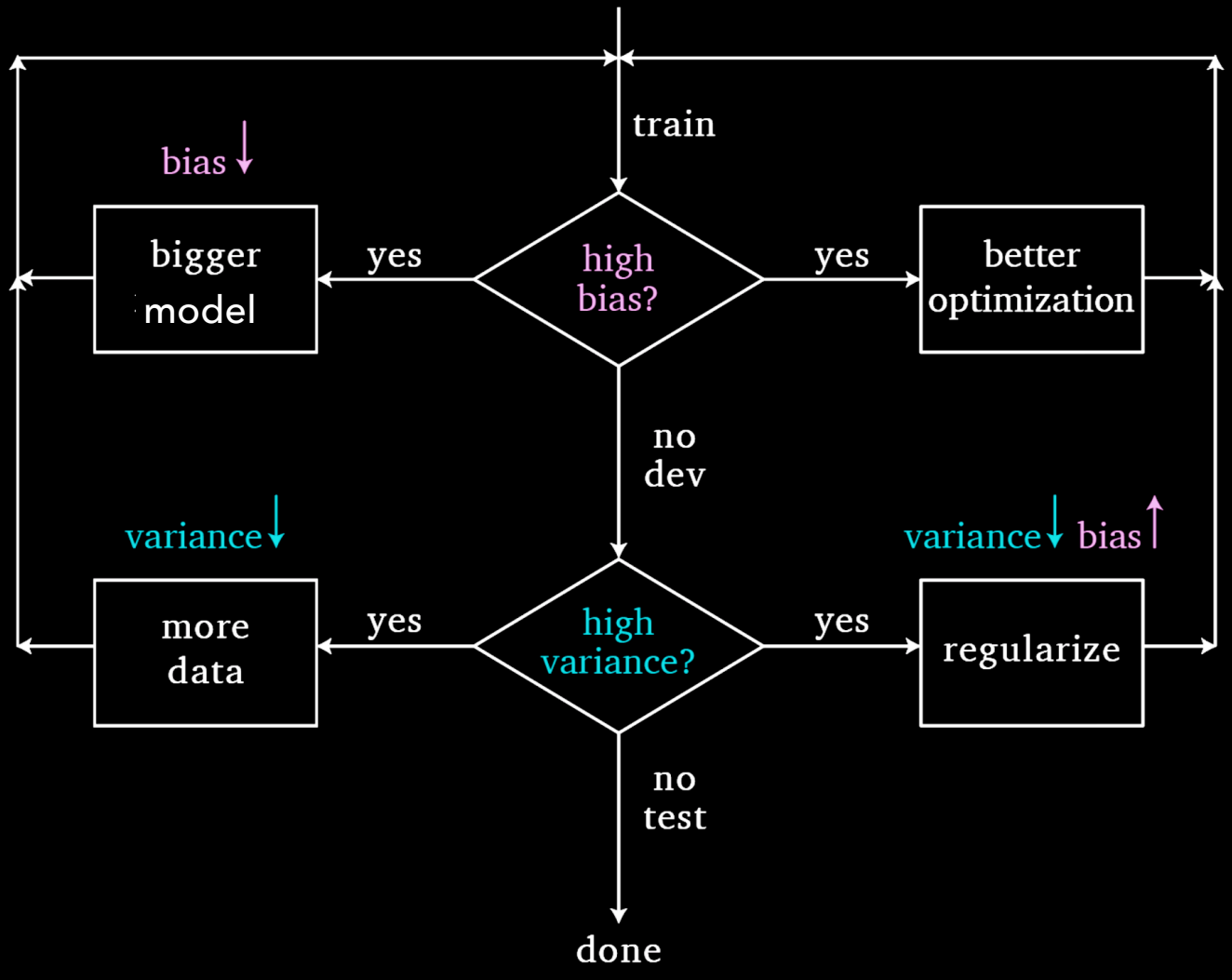
FITTING CURVES



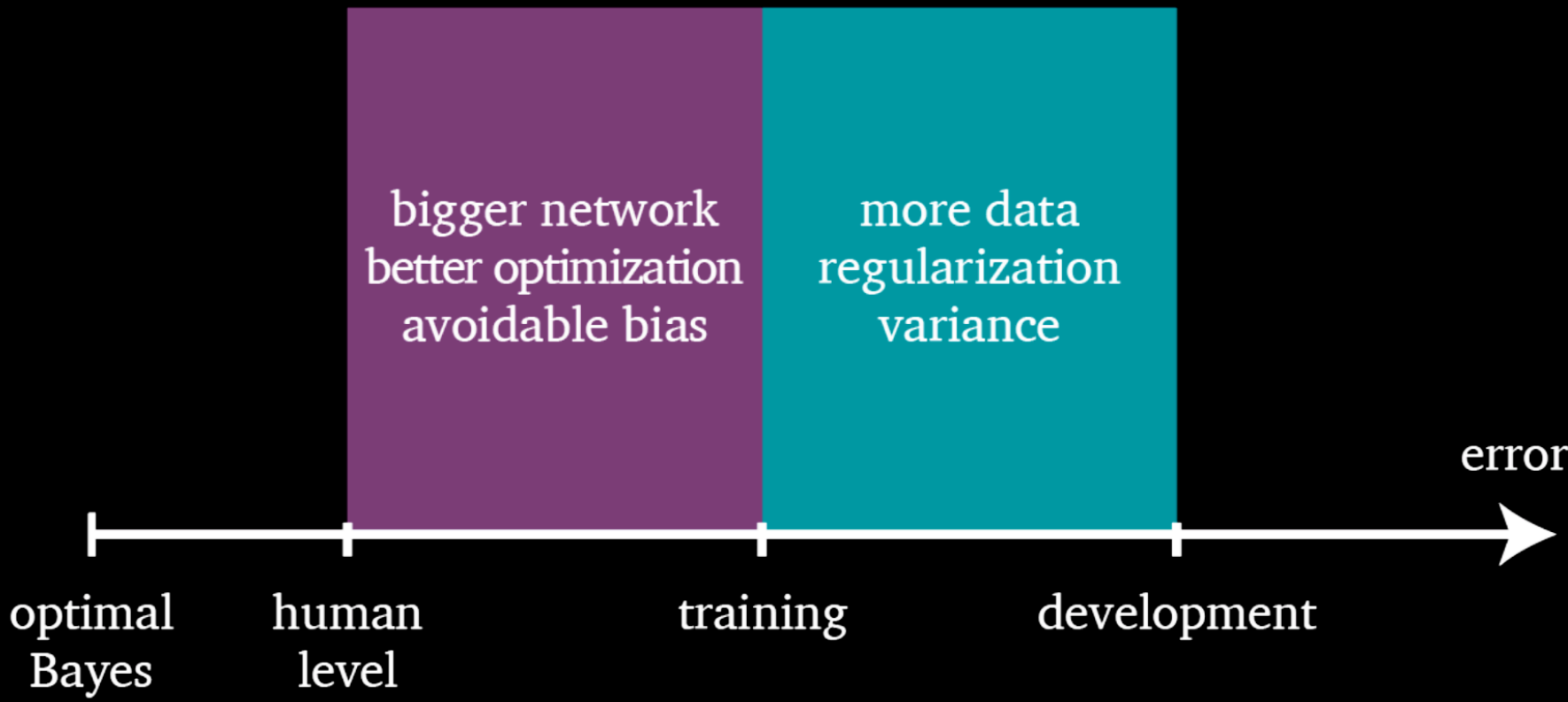
LEARNING CURVES



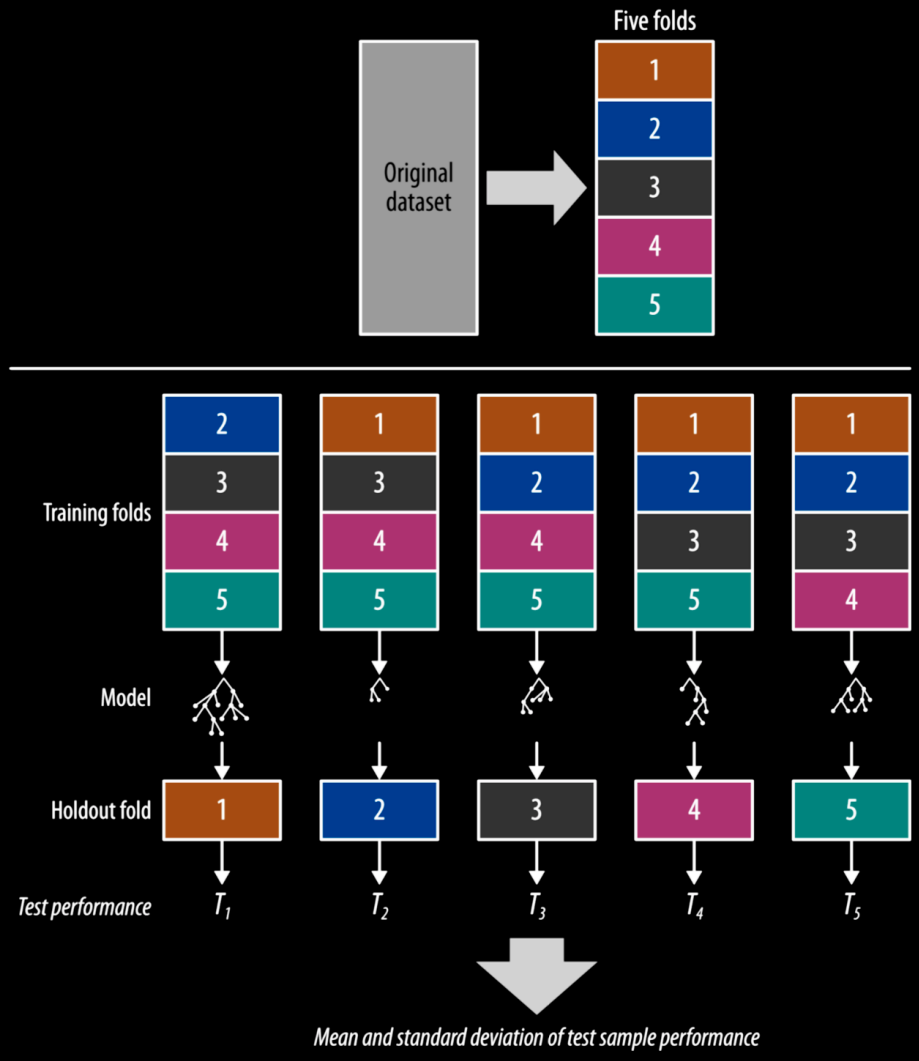
BIAS AND VARIANCE



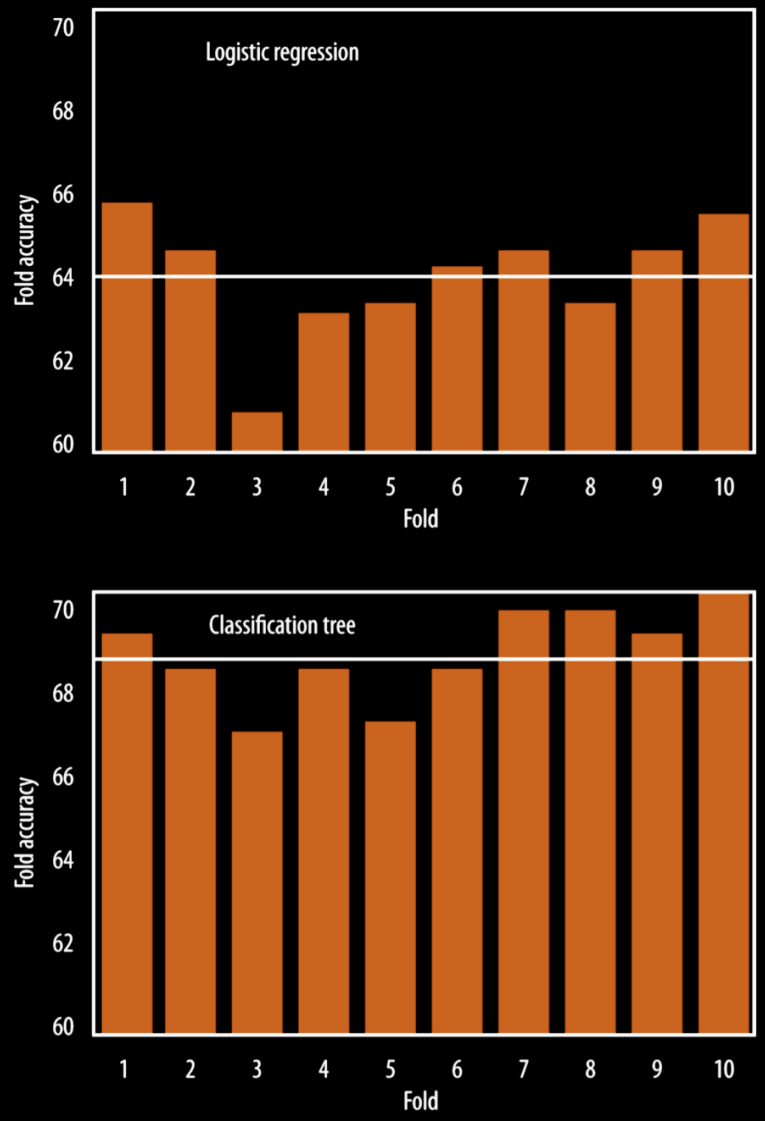
BIAS AND VARIANCE



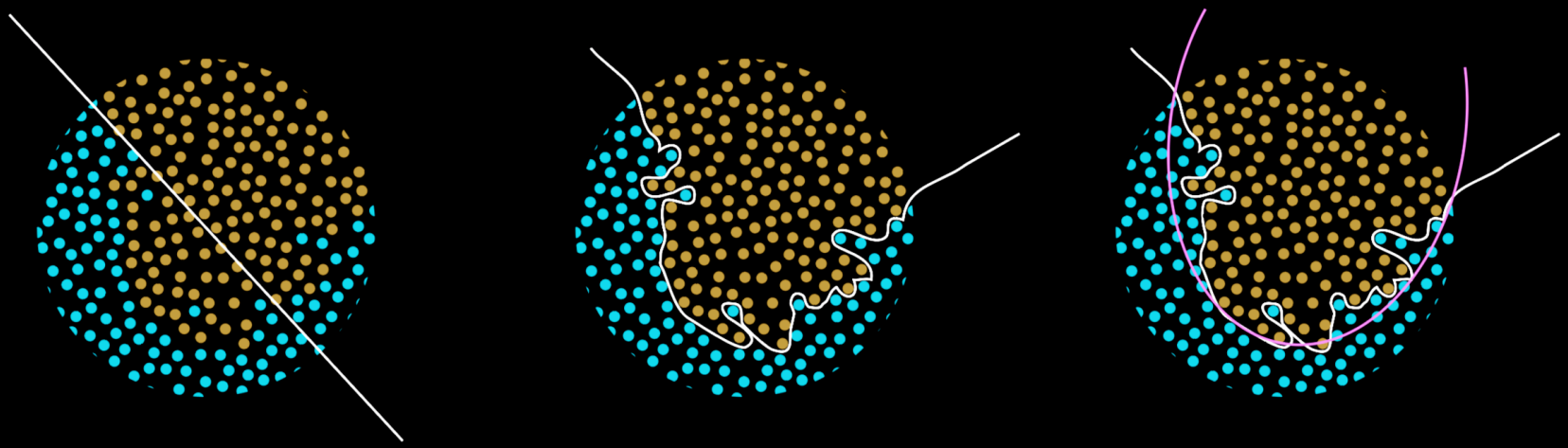
CROSS VALIDATION



CROSS VALIDATION



REGULARIZATION

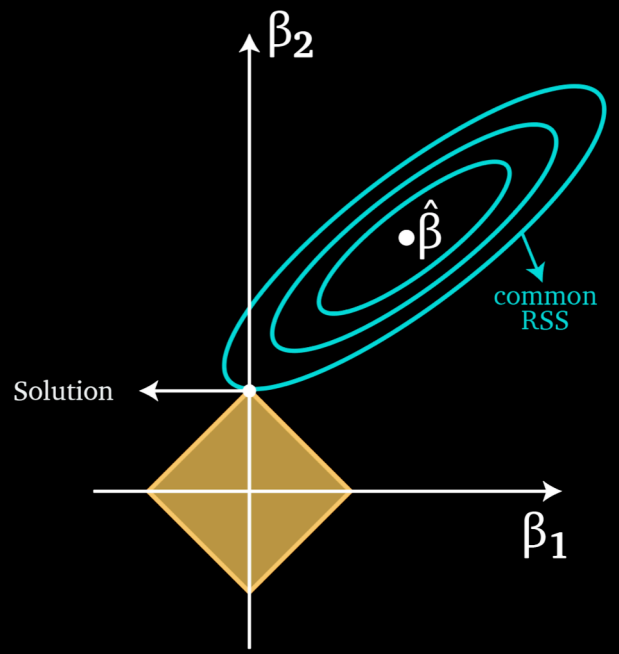


REGULARIZATION

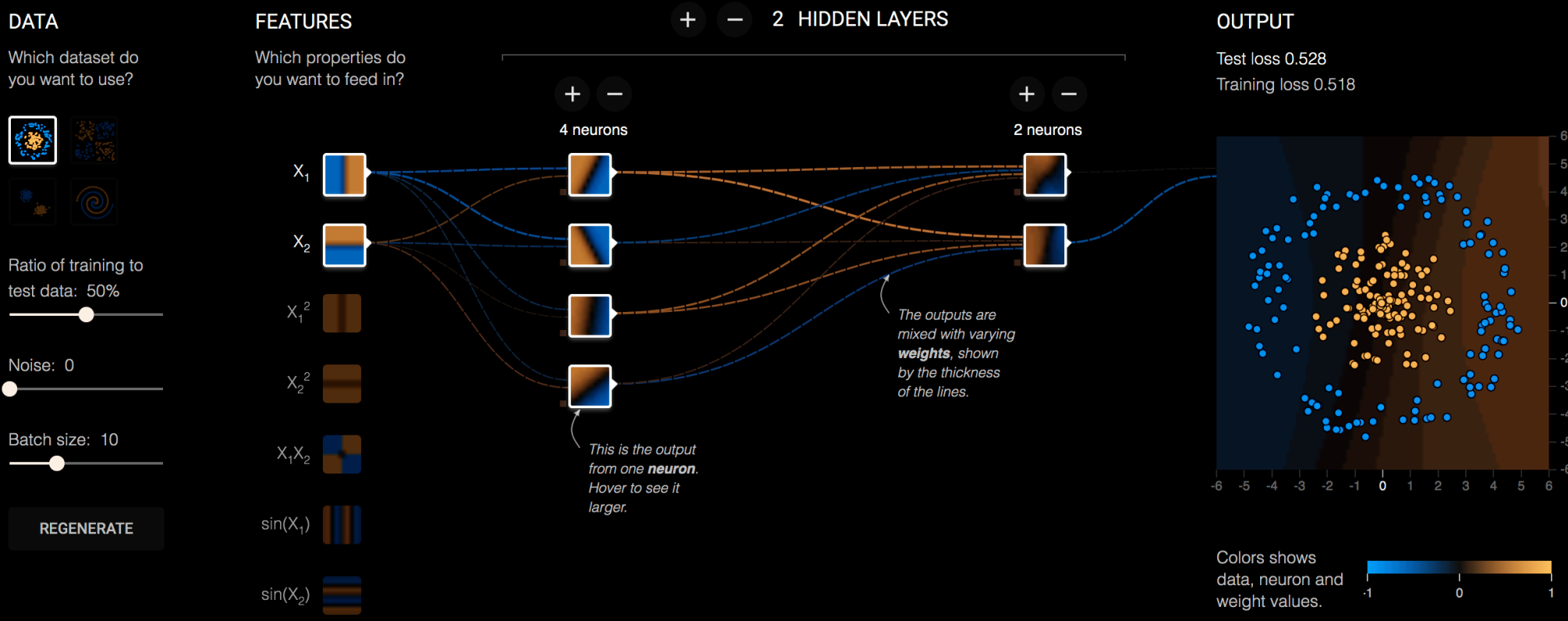
- Add regularization function R of w to cost function

$$J(w) = \frac{1}{n}\sum_{i=1}^{n} L\left(y^i, f\left(x^i\right)\right) + R(w)$$

- L2 regularization $R(w) = \lambda \sum_i w_i^2$
- L1 regularization $R(w) = \lambda \sum_i |w_i|$



CLASSIFICATION: NEURAL NETWORK



CLASSIFICATION IN DATA SCIENCE PROJECT

Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States

Timnit Gebru[a,1], Jonathan Krause[a], Yilun Wang[a], Duyun Chen[a], Jia Deng[b], Erez Lieberman Aiden[c,d,e], and Li Fei-Fei[a]

[a]Artificial Intelligence Laboratory, Computer Science Department, Stanford University, Stanford, CA 94305; [b]Vision and Learning Laboratory, Computer Science and Engineering Department, University of Michigan, Ann Arbor, MI 48109; [c]The Center for Genome Architecture, Department of Genetics, Baylor College of Medicine, Houston, TX 77030; [d]Department of Computer Science, Rice University, Houston, TX 77005; and [e]The Center for Genome Architecture, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved October 16, 2017 (received for review January 4, 2017)

The United States spends more than $250 million each year on the American Community Survey (ACS), a labor-intensive door-to-door study that measures statistics relating to race, gender, education, occupation, unemployment, and other demographic factors. Although a comprehensive source of data, the lag between demographic changes and their appearance in the ACS can exceed several years. As digital imagery becomes ubiquitous and machine vision techniques improve, automated data analysis may become an increasingly practical supplement to the ACS. Here, we present a method that estimates socioeconomic characteristics of regions spanning 200 US cities by using 50 million images of street scenes gathered with Google Street View cars. Using deep learning-based computer vision techniques, we determined the make, model, and year of all motor vehicles encountered in particular neighborhoods. Data from this census of motor vehicles, which enumerated 22 million automobiles in total (8% of all automobiles in the United States), were used to accurately estimate income, race, education, and voting patterns at the zip code and precinct level. (The average US precinct contains ∼1,000 people.) The resulting associations are surprisingly simple and powerful. For instance, if the number of sedans encountered during a drive through a city is higher than the number of pickup trucks, the city is likely to vote for a Democrat during the next presidential election (88% chance); otherwise, it is likely to vote Republican (82%). Our results suggest that automated systems for monitoring demographics may effectively complement labor-intensive approaches, with the potential to measure demographics with fine spatial resolution, in close to real time.

of analyzing demographic trends in great detail, in real time, and at a fraction of the cost.

Recently, Naik et al. (7) used publicly available imagery to quantify people's subjective perceptions of a neighborhood's physical appearance. They then showed that changes in these perceptions correlate with changes in socioeconomic variables (8). Our work explores a related theme: whether socioeconomic statistics can be inferred from objective characteristics of images from a neighborhood.

Here, we show that it is possible to determine socioeconomic statistics and political preferences in the US population by combining publicly available data with machine-learning methods. Our procedure, designed to build upon and complement the ACS, uses labor-intensive survey data for a handful of cities to train a model that can create nationwide demographic estimates. This approach allows for estimation of demographic variables with high spatial resolution and reduced lag time.

Specifically, we analyze 50 million images taken by Google Street View cars as they drove through 200 cities, neighborhood-by-neighborhood and street-by-street. In Google Street View images, only the exteriors of houses, landscaping, and vehicles on the street can be observed. Of these objects, vehicles are among the most personalized expressions of American culture: Over 90% of American households own a motor vehicle (9), and their choice of automobile is influenced by disparate demographic factors including household needs, personal preferences, and economic wherewithal (10). (Note that, in principle, other factors such as spacing between houses, number of stories, and extent of shrubbery could also be integrated into such models.) Such street scenes are a natural data type to explore: They already cover

TEXT

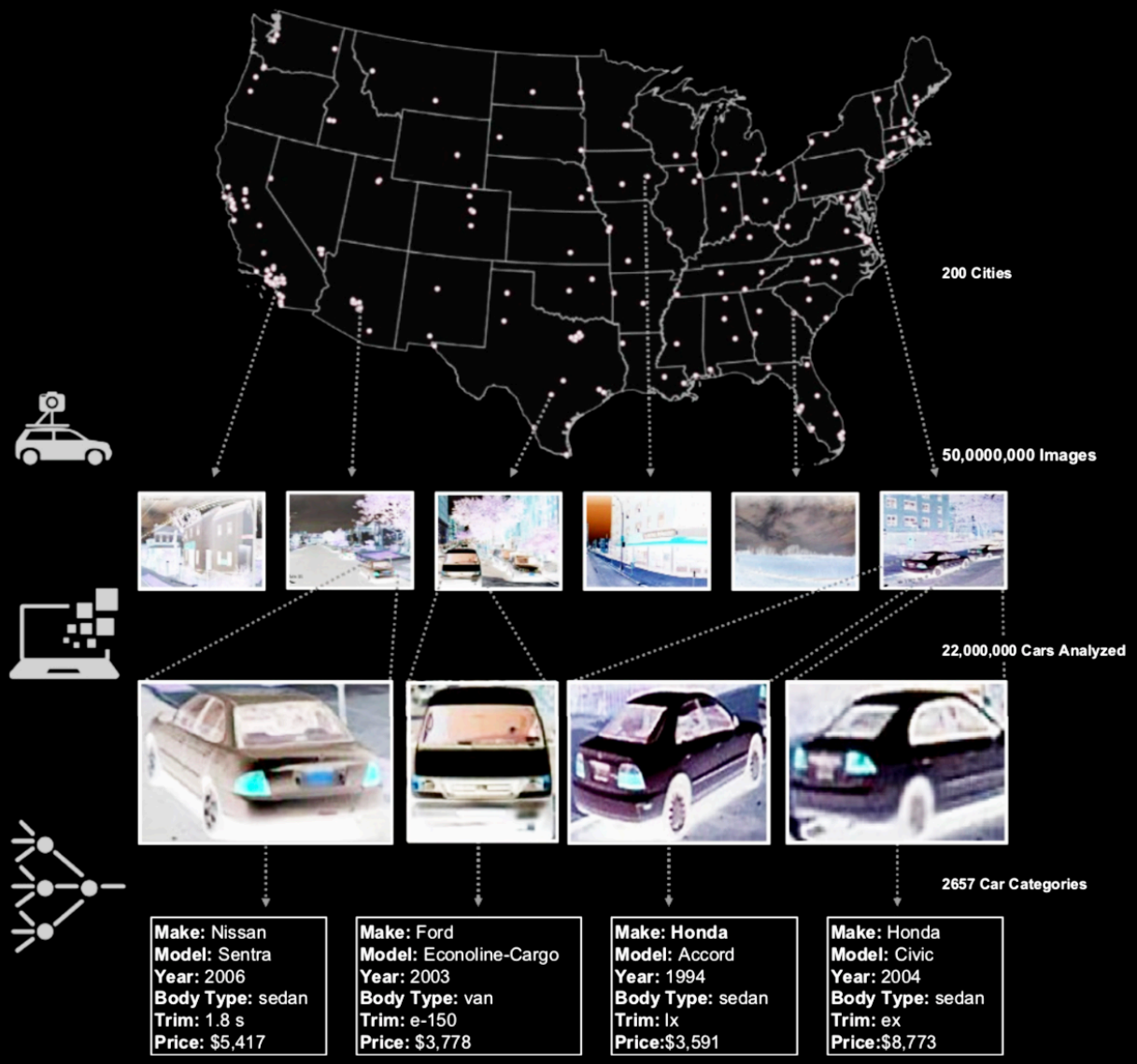


Fig. 1. We perform a vehicular census of 200 cities in the United States using 50 million Google Street View images. In each image, we detect cars with computer vision algorithms based on DPM and count an estimated 22 million cars. We then use CNN to categorize the detected vehicles into one of 2,657 classes of cars. For each type of car, we have metadata such as the make, model, year, body type, and price of the car in 2012. Images courtesy of Google Maps/Google Earth.

Fig. 2. We use all of the cities in counties starting with A, B, and C (shown in purple on the map) to train a model estimating socioeconomic data from car attributes. Using this model, we estimate demographic variables at the zip code level for all of the cities shown in green. We show actual vs. predicted maps for the percentage of Black, Asian, and White people in Seattle, WA (*i–iii*); the percentage of people with less than a high school degree in Milwaukee, WI (*iv*); and the percentage of people with graduate degrees in Milwaukee, WI (*v*). (*vi*) Maps the median household income in Tampa, FL. The ground truth values are mapped on *Left*, and our estimated results are on *Right*. We accurately localize zip codes with the highest and lowest concentrations of each demographic variable such as the three zip codes in Eastern Seattle with high concentrations of Caucasians, one Northern zip code in Milwaukee with highly educated inhabitants, and the least wealthy zip code in Southern Tampa.

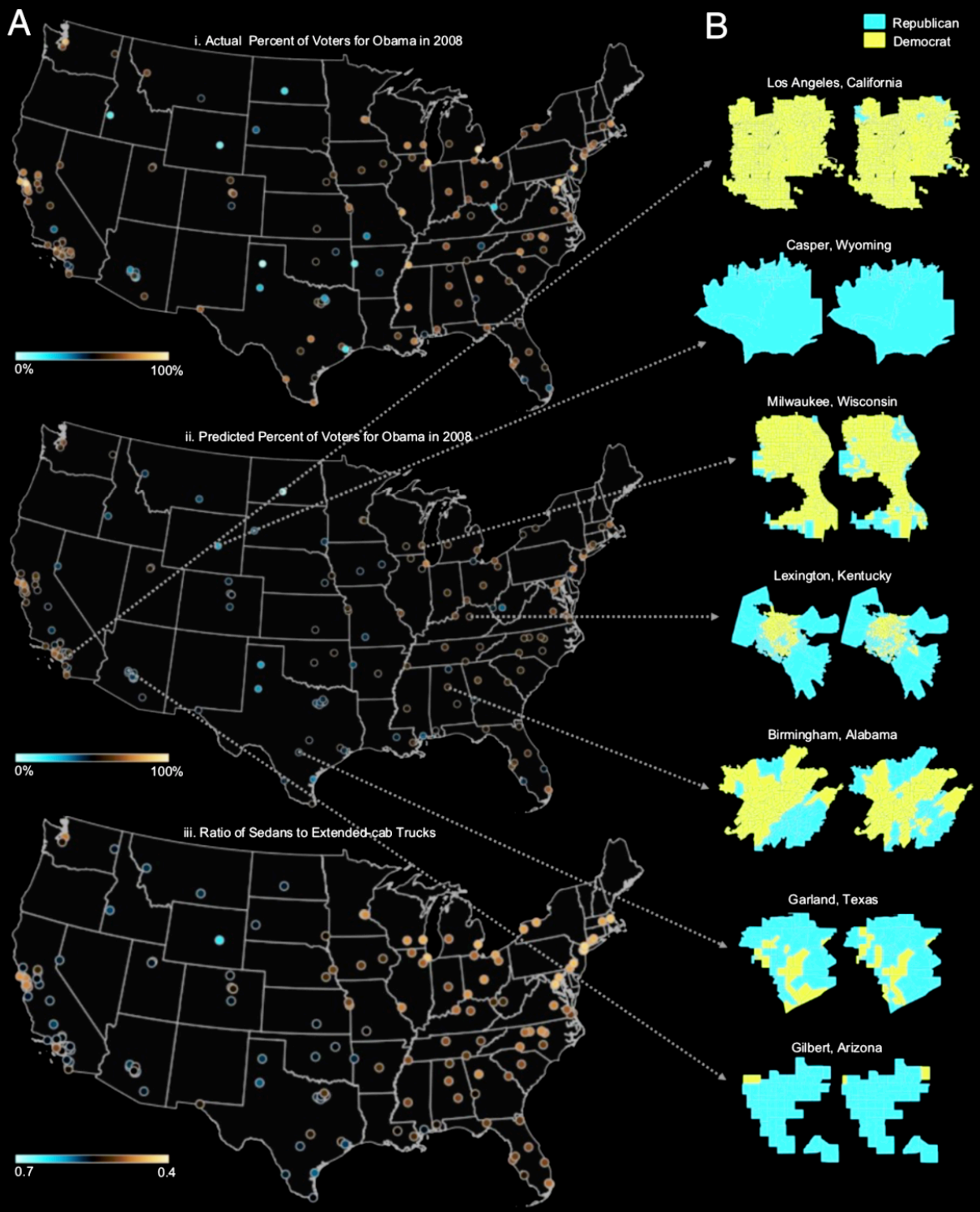


Fig. 3. Actual and inferred voting patterns. *A*, *i* and *ii* map the actual and predicted percentage of people who voted for Barack Obama in the 2008 presidential election (r = 0.74). *iii* maps the ratio of detected pickup trucks to sedans in the 165 cities in our test set. As can be seen from the map, the ratio is very low in Democratic cities such as those in the East Coast and high in Republican cities such as those in Texas and Wyoming. (*B*) Shows actual vs. predicted voter affiliations for various cities in our test set at the precinct level using our full model. Democratic precincts are shown in blue, and Republican precincts are shown in red. Our model correctly classifies Casper, WY as a Republican city and Los Angeles, CA as a Democratic city. We accurately predict that Milwaukee, WI is a Democratic city except for a few Republican precincts in the southern, western, and northeastern borders of the city.