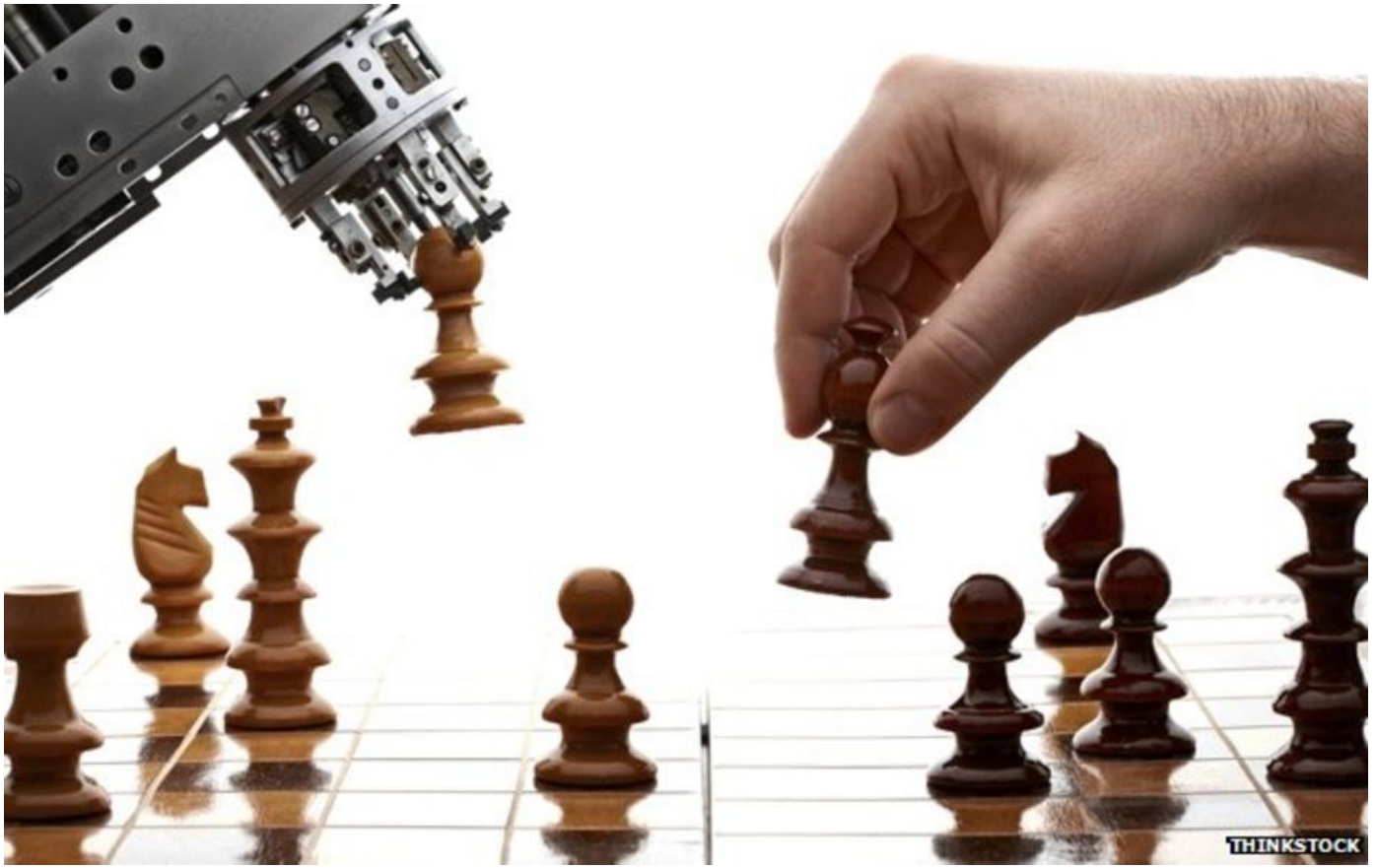


Introduction to Artificial Intelligence

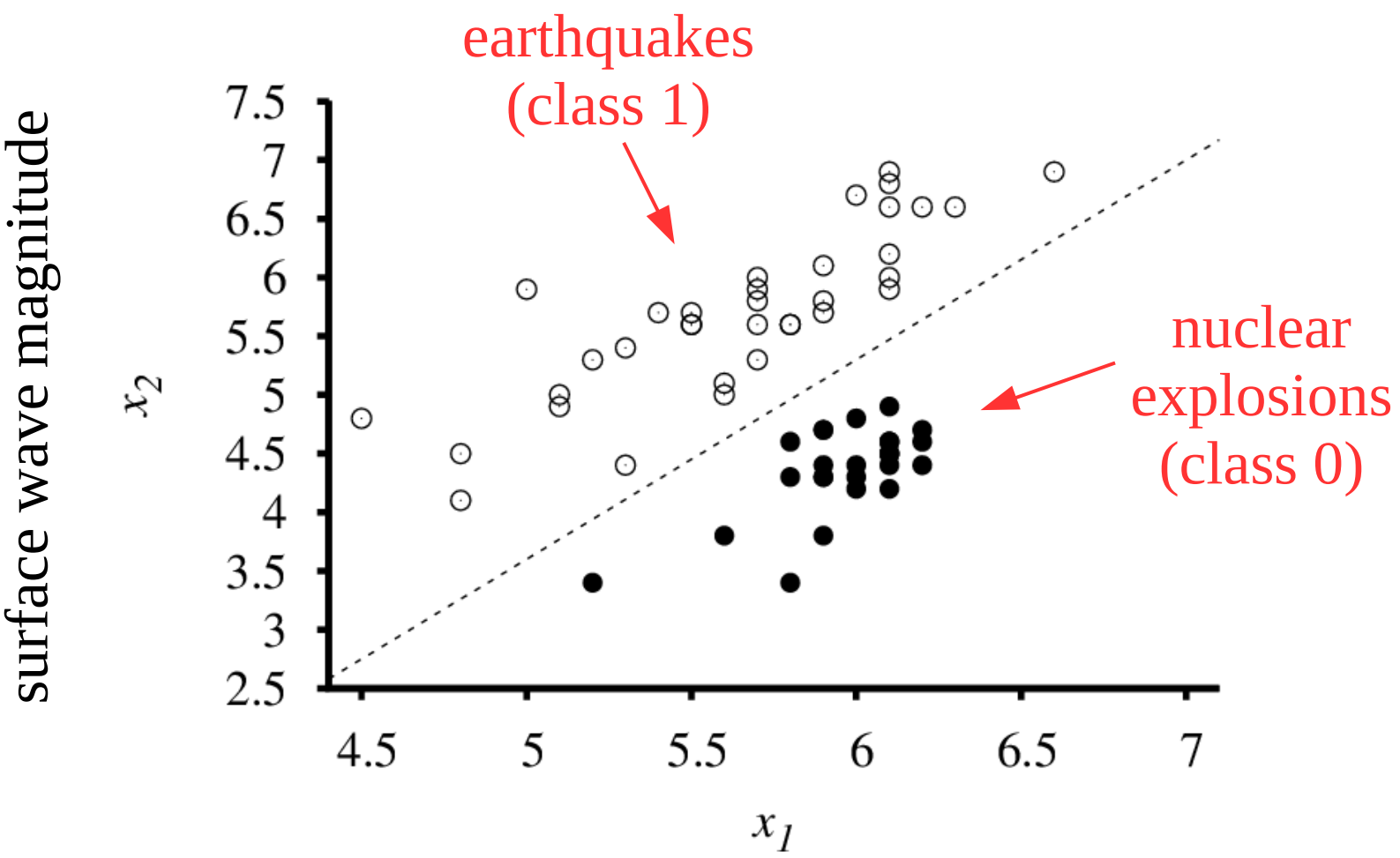
COSC 4550 / COSC 5550

Professor Cheney

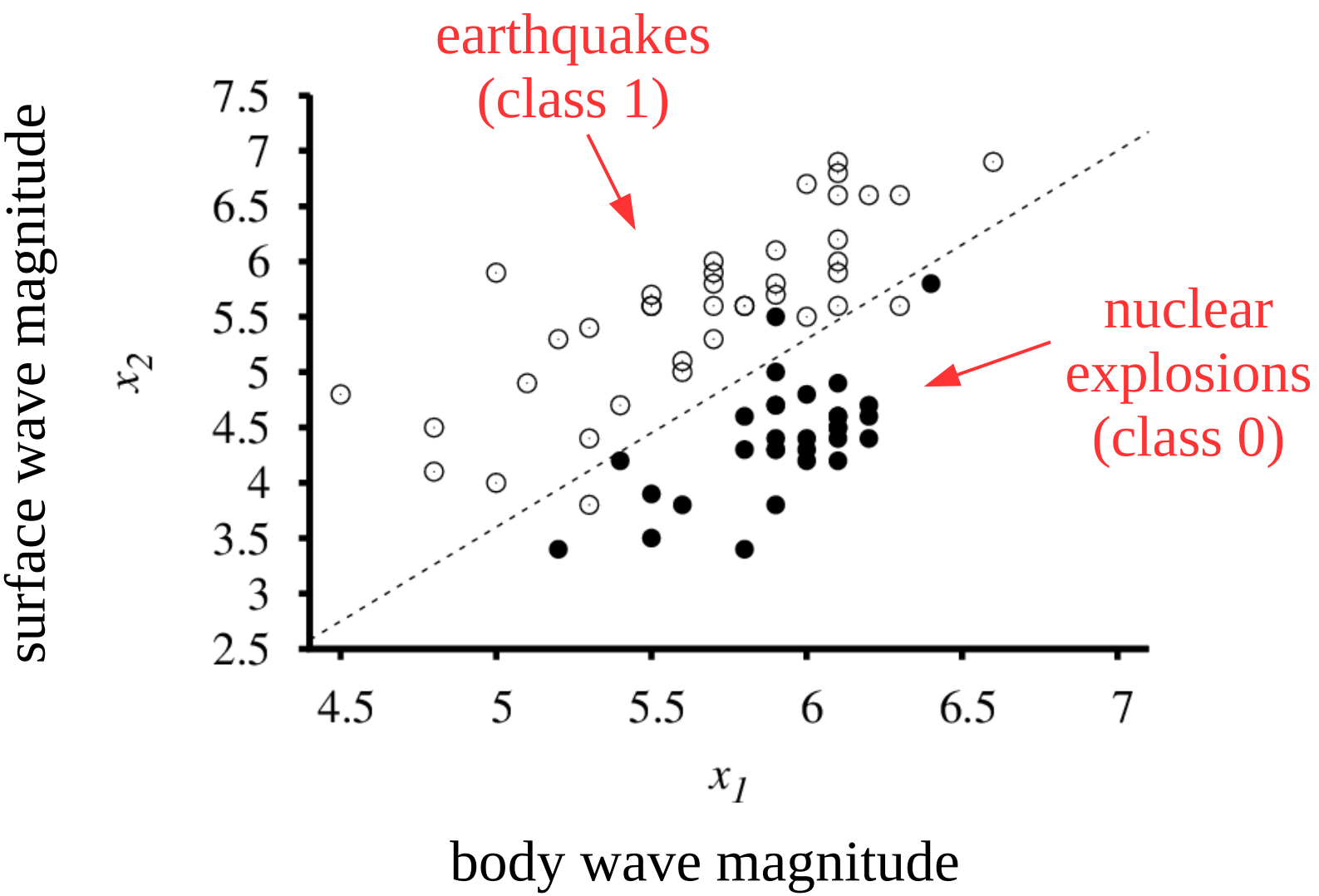
10/16/17

linear classifiers

Seismic events from 1982 to 1990 in Asia and the Middle East



Seismic events from 1982 to 1990 in Asia and the Middle East



just like before, we are still trying to find the parameters for a straight line that best “fits” the data

now “fit” is the line that best separates the classes, rather than tracks the data points (“linear separator”)

more generally called a “decision boundary”

we need a function to turn the outputs of our linear function into class labels

$$\text{Threshold}(z) = \begin{cases} 1 & \text{if } z >= 0 \\ 0 & \text{if } z < 0 \end{cases}$$

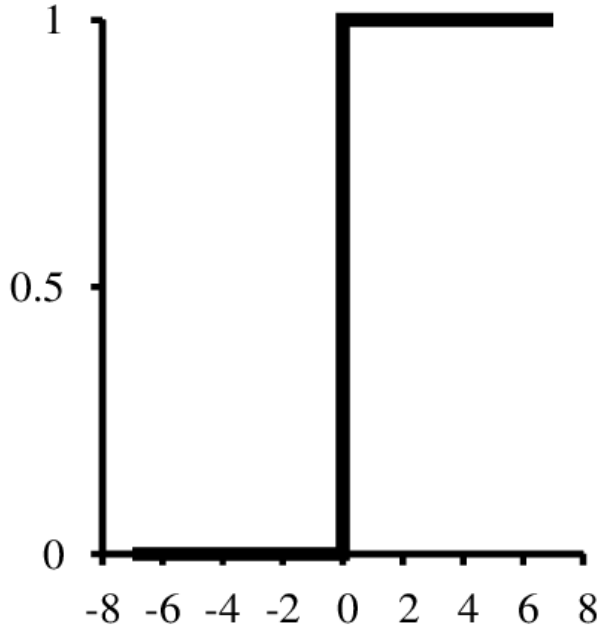
$$w_i \leftarrow w_i + \alpha\,(y - h(x)) * x_i$$

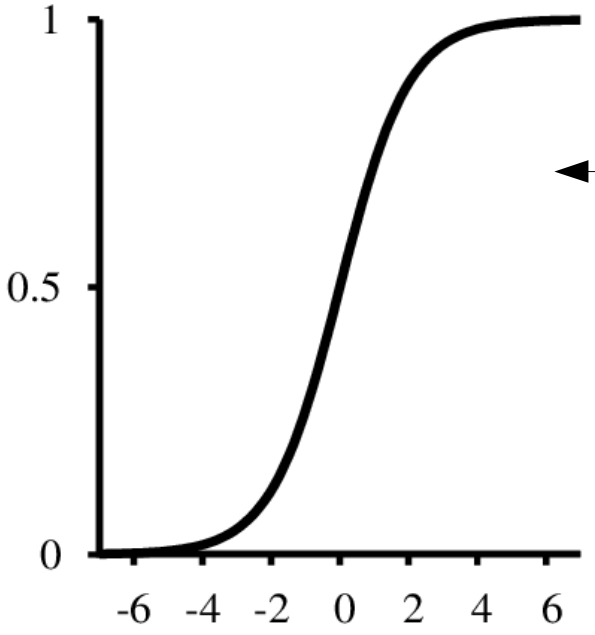
where $h(x) = \text{Threshold}(w \circ x)$

what's the problem?

the (step) threshold function has no non-zero derivatives!

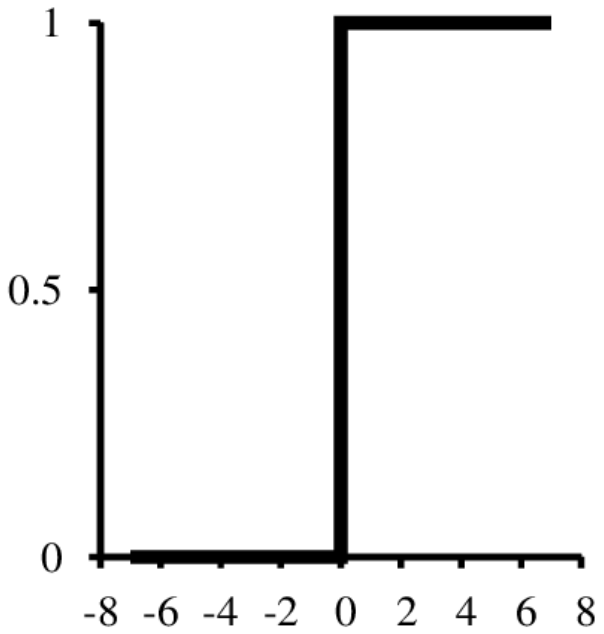
(we know what direction to change, but not by how much, learning is unstable!)

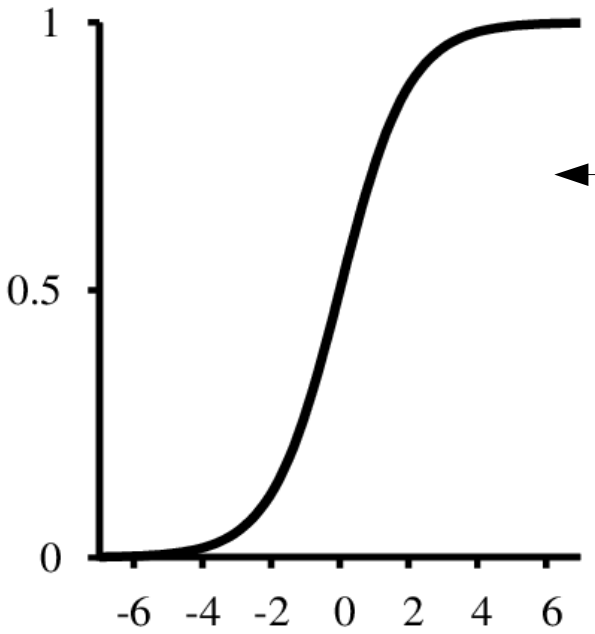




logistic function

$$\frac{1}{1+e^{-w \circ x}}$$





logistic function

$$\frac{1}{1+e^{-w \circ x}}$$

only outputs either 0 or 1

completely confident

outputs any value between 0 or 1

allow for uncertainty by giving probability of a certain class

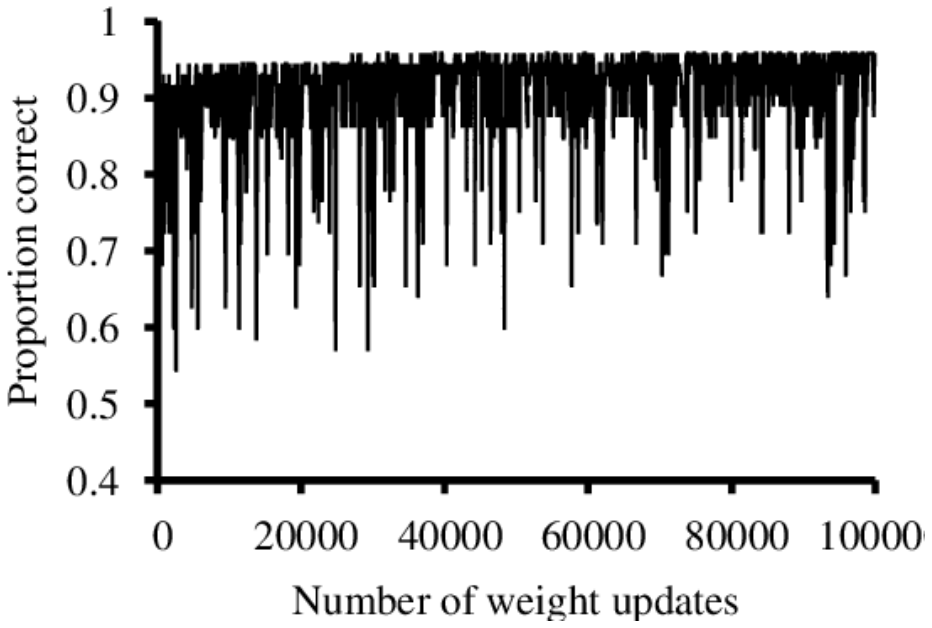
$$
\begin{aligned}
g'_{\text{logistic}}(z) &= \frac{\partial}{\partial z}\left(\frac{1}{1+e^{-z}}\right) \\
&= \frac{e^{-z}}{(1+e^{-z})^2}(\text{chain rule}) \\
&= \frac{1+e^{-z}-1}{(1+e^{-z})^2} \\
&= \frac{1+e^{-z}}{(1+e^{-z})^2} - \left(\frac{1}{1+e^{-z}}\right)^2 \\
&= \frac{1}{(1+e^{-z})} - \left(\frac{1}{1+e^{-z}}\right)^2 \\
&= g_{\text{logistic}}(z) - g_{\text{logistic}}(z)^2 \\
&= g_{\text{logistic}}(z)(1 - g_{\text{logistic}}(z))
\end{aligned}
$$

$$\text{Loss} = (y - h_w(x))^2$$

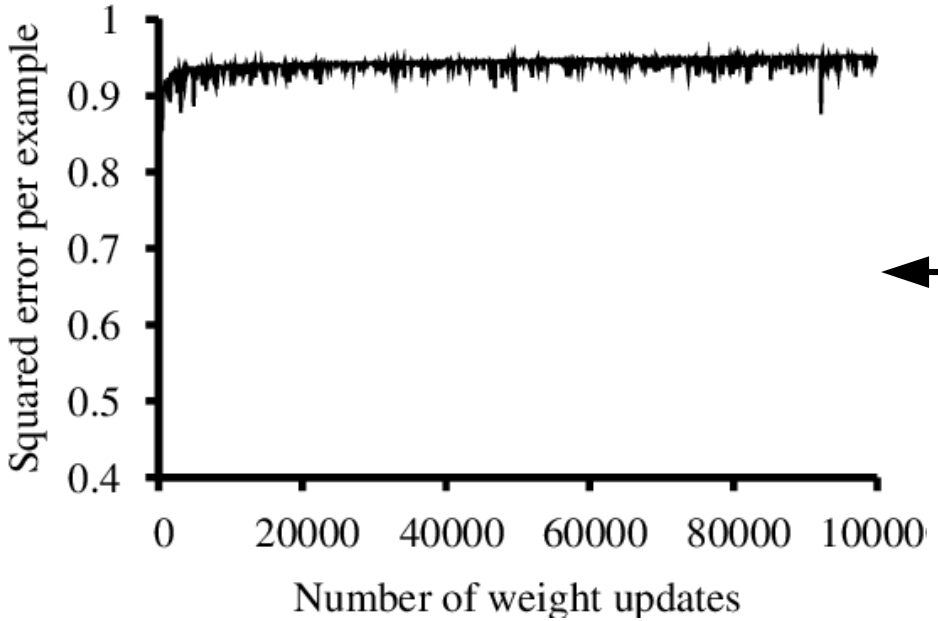
$$\begin{aligned}\delta/\delta w_i \text{ Loss} &= 2\,(y - h_w(x)) * \delta/\delta w_i\,(y - h_w(x)) \\ &= 2\,(y - h_w(x)) * h_w(x) * (1-h_w(x)) * \delta/\delta w_i\,(w \circ x) \\ &= 2\,(y - h_w(x)) * h_w(x) * (1-h_w(x)) * x\end{aligned}$$

$$w_i \leftarrow w_i + \alpha\,((y - h_w(x)) * h_w(x) * (1-h_w(x)) * x_i)$$

I know it looks complicated, but it's easy to calculate!
all we need to know is x, y, and h(x)



unstable learning
with hard threshold



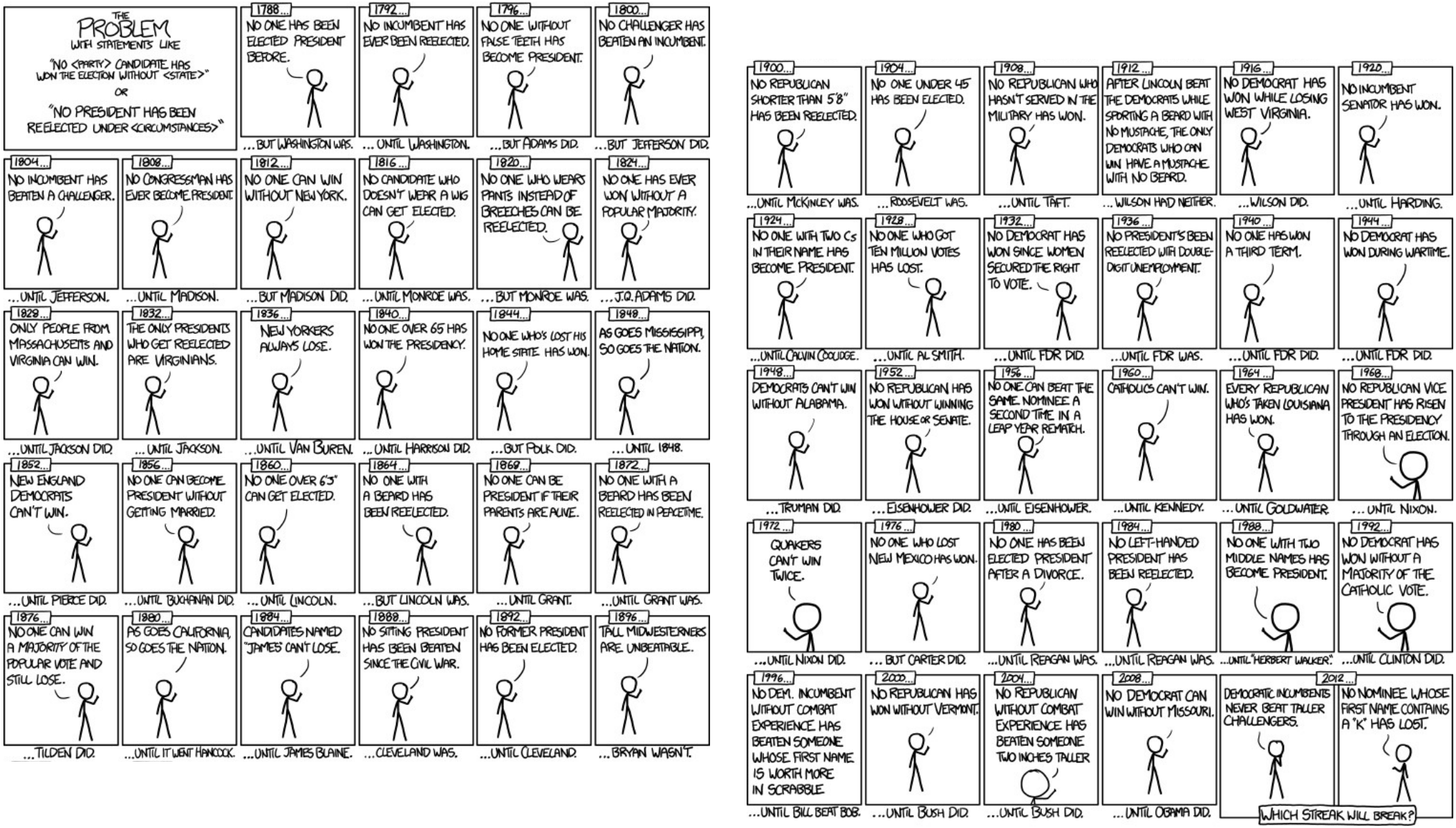
smoothed learning curve
using logistic threshold
(and nice derivatives)

"logistic regression"

overfitting

we've said in previous lectures that learning a model from data assumes that your data is i.i.d (independent and identically distributed)

this means your data is produced by the same underlying (true) model/trend, and variations from that trend are simply due to random noise


THE PROBLEM WITH STATEMENTS LIKE
"NO <PARTY> CANDIDATE HAS WON THE ELECTION WITHOUT <STATE>"
OR
"NO PRESIDENT HAS BEEN REELECTED UNDER <CIRCUMSTANCES>"
1788... NO ONE HAS BEEN ELECTED PRESIDENT BEFORE. ...BUT WASHINGTON WAS.
1792... NO INCUMBENT HAS EVER BEEN REELECTED. ...UNTIL WASHINGTON.
1796... NO ONE WITHOUT FALSE TEETH HAS BECOME PRESIDENT. ...BUT ADAMS DID.
1800... NO CHALLENGER HAS BEATEN AN INCUMBENT. ...BUT JEFFERSON DID.
1804... NO INCUMBENT HAS BEATEN A CHALLENGER. ...UNTIL JEFFERSON.
1808... NO CONGRESSMAN HAS EVER BECOME PRESIDENT. ...UNTIL MADISON.
1812... NO ONE CAN WIN WITHOUT NEW YORK. ...BUT MADISON DID.
1816... NO CANDIDATE WHO DOESN'T WEAR A WIG CAN GET ELECTED. ...UNTIL MONROE WAS.
1820... NO ONE WHO WEARS PANTS INSTEAD OF BREECHES CAN BE REELECTED. ...BUT MONROE WAS.
1824... NO ONE HAS EVER WON WITHOUT A POPULAR MAJORITY. ...J.Q. ADAMS DID.
1828... ONLY PEOPLE FROM MASSACHUSETTS AND VIRGINIA CAN WIN. ...UNTIL JACKSON DID.
1832... THE ONLY PRESIDENTS WHO GET REELECTED ARE VIRGINIANS. ... UNTIL JACKSON.
1836... NEW YORKERS ALWAYS LOSE. ...UNTIL VAN BUREN.
1840... NO ONE OVER 65 HAS WON THE PRESIDENCY. ... UNTIL HARRISON DID.
1844... NO ONE WHO'S LOST HIS HOME STATE HAS WON. ...BUT POLK DID.
1848... AS GOES MISSISSIPPI, SO GOES THE NATION. ... UNTIL 1848.
1852... NEW ENGLAND DEMOCRATS CAN'T WIN. ...UNTIL PIERCE DID.
1856... NO ONE CAN BECOME PRESIDENT WITHOUT GETTING MARRIED. ...UNTIL BUCHANAN DID.
1860... NO ONE OVER 6'3" CAN GET ELECTED. ... UNTIL LINCOLN.
1864... NO ONE WITH A BEARD HAS BEEN REELECTED. ...BUT LINCOLN WAS.
1868... NO ONE CAN BE PRESIDENT IF THEIR PARENTS ARE ALIVE. ... UNTIL GRANT.
1872... NO ONE WITH A BEARD HAS BEEN REELECTED IN PEACETIME. ...UNTIL GRANT WAS.
1876... NO ONE CAN WIN A MAJORITY OF THE POPULAR VOTE AND STILL LOSE. ...TILDEN DID.
1880... AS GOES CALIFORNIA, SO GOES THE NATION. ...UNTIL IT WENT HANCOCK.
1884... CANDIDATES NAMED "JAMES" CAN'T LOSE. ...UNTIL JAMES BLAINE.
1888... NO SITTING PRESIDENT HAS BEEN BEATEN SINCE THE CIVIL WAR. ...CLEVELAND WAS.
1892... NO FORMER PRESIDENT HAS BEEN ELECTED. ...UNTIL CLEVELAND.
1896... TALL MIDWESTERNERS ARE UNBEATABLE. ...BRYAN WASN'T.
1900... NO REPUBLICAN SHORTER THAN 5'8" HAS BEEN REELECTED. ...UNTIL MCKINLEY WAS.
1904... NO ONE UNDER 45 HAS BEEN ELECTED. ...ROOSEVELT WAS.
1908... NO REPUBLICAN WHO HASN'T SERVED IN THE MILITARY HAS WON. ...UNTIL TAFT.
1912... AFTER LINCOLN BEAT THE DEMOCRATS WHILE SPORTING A BEARD WITH NO MUSTACHE, THE ONLY DEMOCRATS WHO CAN WIN HAVE A MUSTACHE WITH NO BEARD. ... WILSON HAD NEITHER.
1916... NO DEMOCRAT HAS WON WHILE LOSING WEST VIRGINIA. ...WILSON DID.
1920... NO INCUMBENT SENATOR HAS WON. ...UNTIL HARDING.
1924... NO ONE WITH TWO Cs IN THEIR NAME HAS BECOME PRESIDENT. ...UNTIL CALVIN COOLIDGE.
1928... NO ONE WHO GOT TEN MILLION VOTES HAS LOST. ...UNTIL AL SMITH.
1932... NO DEMOCRAT HAS WON SINCE WOMEN SECURED THE RIGHT TO VOTE. ...UNTIL FDR DID.
1936... NO PRESIDENT'S BEEN REELECTED WITH DOUBLE-DIGIT UNEMPLOYMENT. ...UNTIL FDR WAS.
1940... NO ONE HAS WON A THIRD TERM. ...UNTIL FDR DID.
1944... NO DEMOCRAT HAS WON DURING WARTIME. ...UNTIL FDR DID.
1948... DEMOCRATS CAN'T WIN WITHOUT ALABAMA. ...TRUMAN DID.
1952... NO REPUBLICAN HAS WON WITHOUT WINNING THE HOUSE OR SENATE. ...EISENHOWER DID.
1956... NO ONE CAN BEAT THE SAME NOMINEE A SECOND TIME IN A LEAP YEAR REMATCH. ...UNTIL EISENHOWER.
1960... CATHOLICS CAN'T WIN. ...UNTIL KENNEDY.
1964... EVERY REPUBLICAN WHO'S TAKEN LOUISIANA HAS WON. ... UNTIL GOLDWATER.
1968... NO REPUBLICAN VICE PRESIDENT HAS RISEN TO THE PRESIDENCY THROUGH AN ELECTION. ... UNTIL NIXON.
1972... QUAKERS CAN'T WIN TWICE. ...UNTIL NIXON DID.
1976... NO ONE WHO LOST NEW MEXICO HAS WON. ... BUT CARTER DID.
1980... NO ONE HAS BEEN ELECTED PRESIDENT AFTER A DIVORCE. ...UNTIL REAGAN WAS.
1984... NO LEFT-HANDED PRESIDENT HAS BEEN REELECTED. ...UNTIL REAGAN WAS.
1988... NO ONE WITH TWO MIDDLE NAMES HAS BECOME PRESIDENT. ...UNTIL "HERBERT WALKER."
1992... NO DEMOCRAT HAS WON WITHOUT A MAJORITY OF THE CATHOLIC VOTE. ...UNTIL CLINTON DID.
1996... NO DEM. INCUMBENT WITHOUT COMBAT EXPERIENCE HAS BEATEN SOMEONE WHOSE FIRST NAME IS WORTH MORE IN SCRABBLE. ...UNTIL BILL BEAT BOB.
2000... NO REPUBLICAN HAS WON WITHOUT VERMONT. ...UNTIL BUSH DID.
2004... NO REPUBLICAN WITHOUT COMBAT EXPERIENCE HAS BEATEN SOMEONE TWO INCHES TALLER ...UNTIL BUSH DID.
2008... NO DEMOCRAT CAN WIN WITHOUT MISSOURI. ... UNTIL OBAMA DID.
2012...
DEMOCRATIC INCUMBENTS NEVER BEAT TALLER CHALLENGERS.
NO NOMINEE WHOSE FIRST NAME CONTAINS A "K" HAS LOST.
WHICH STREAK WILL BREAK?

we want to be able to capture the true underlying trends in our data, but we do not want to be modeling any artifacts due to noise (if we can help it)

(iPython example!)

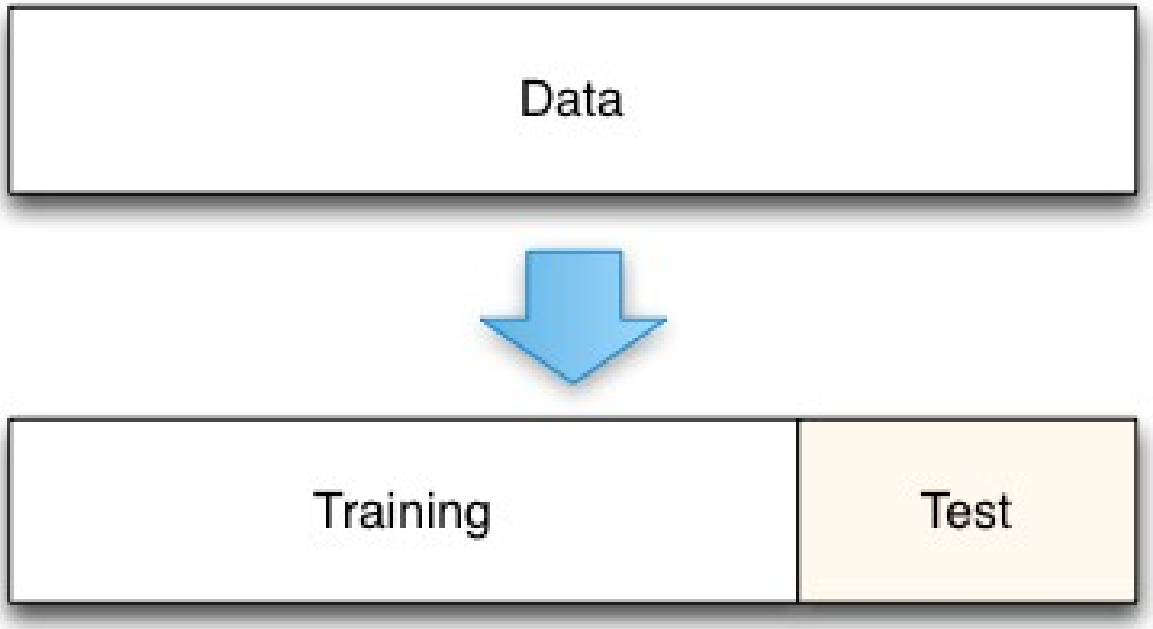
we could add noise to our observations
(or our features/parameters) to prevent the model from fitting exactly to the (imperfect) observations we have

noise would change every optimization iteration
(meaning it would be impossible to fit to)
unlike noise in our data set (which is fixed)

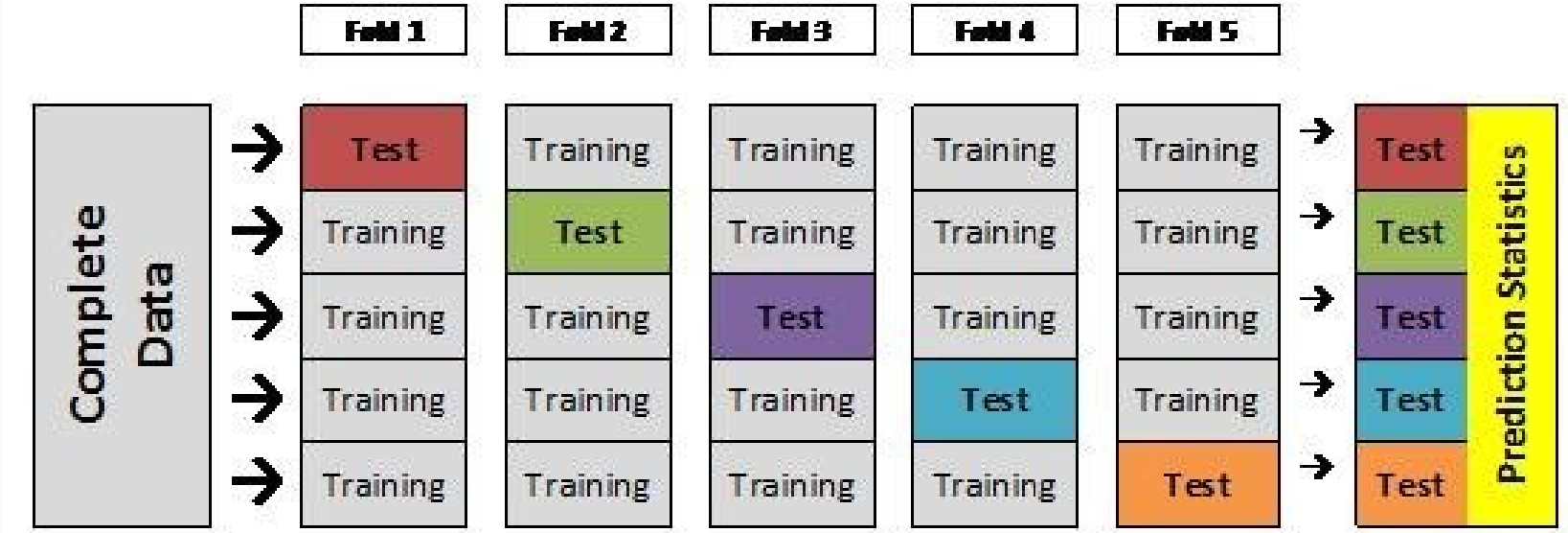
unfortunately adding noise also distorts our underlying data
(but is simple and can be effective in preventing overfitting)

cross-validation

hold-out method



k-fold cross validation



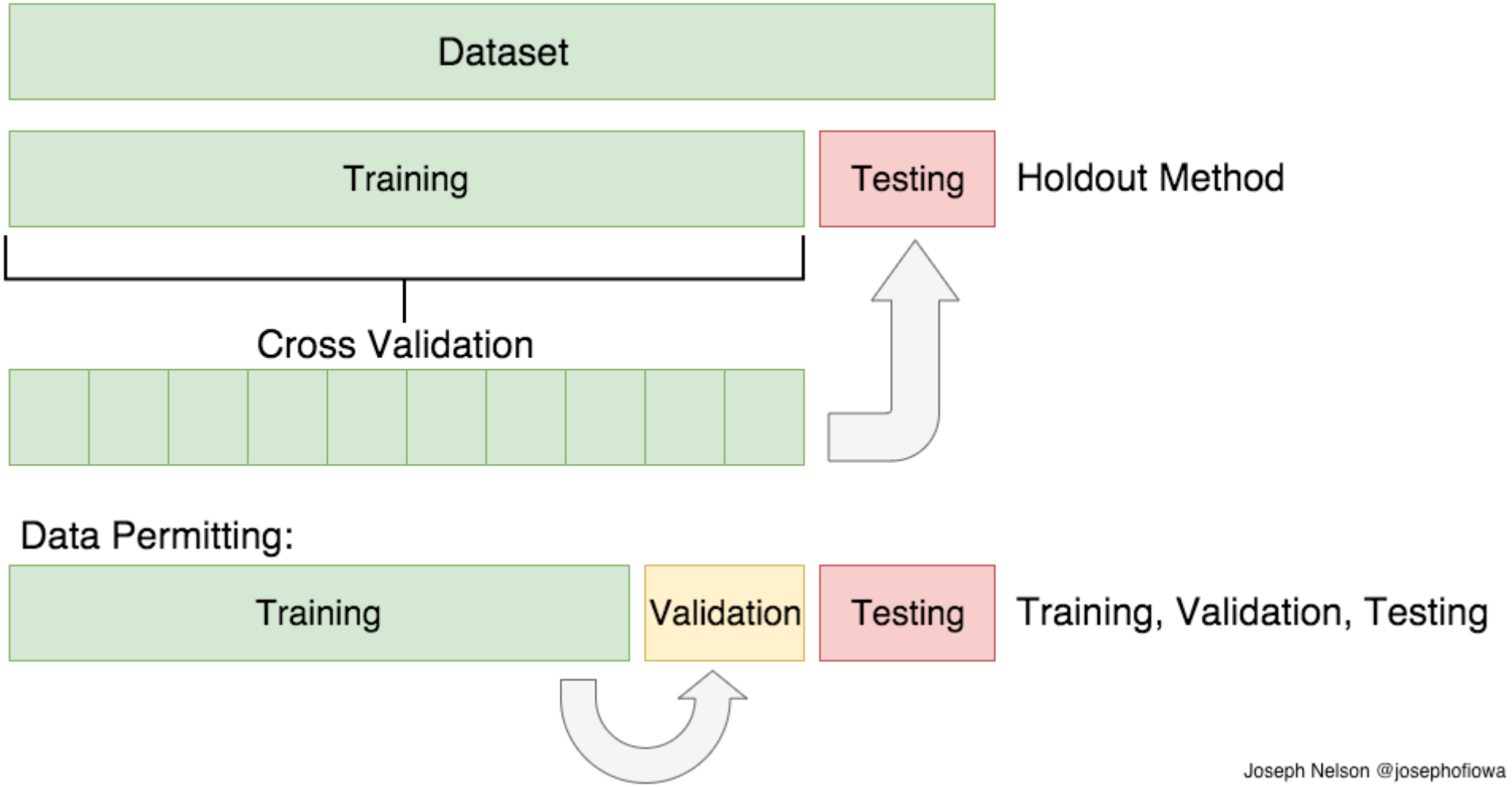
adds noise to training without introducing noisy data!

creates both mean and variance of prediction score!

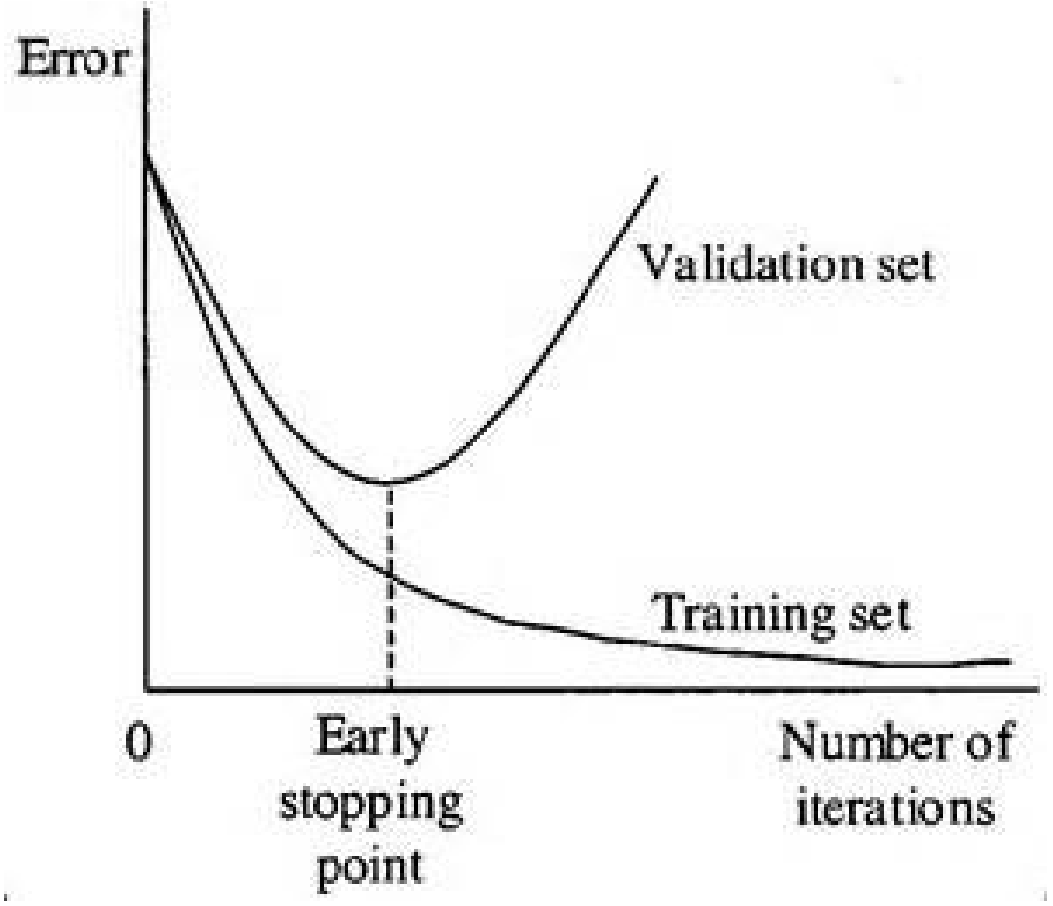
generally awesome... and standard practice

(see “bootstrapping” for more delicious goodness)

can also use a distinct validation set if you have extra data...

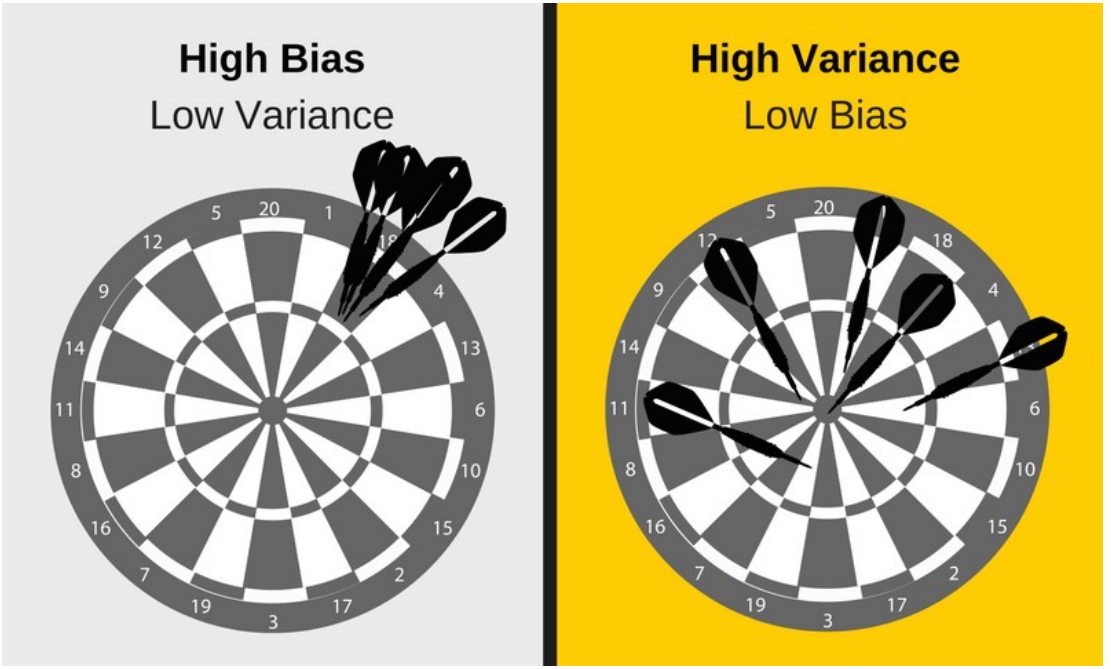


early stopping with a validation error



(iPython example revisited)

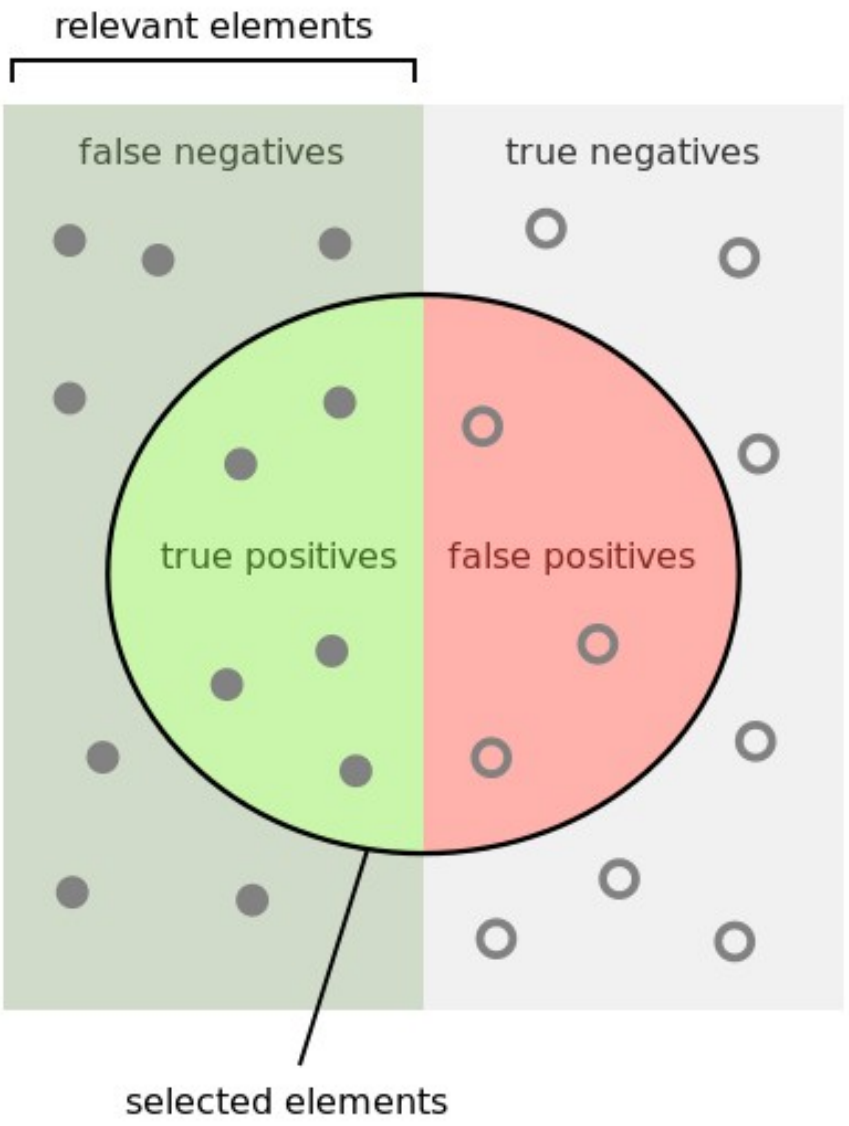
bias vs. variance trade-off



high bias models haven't fit the data well (often due to too limited flexibility of the model – i.e. underfitting)

high variance models are too sensitive to the differences in each instance of data (i.e. overfitting) – accurate on average, but inconsistent

precision vs. recall



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

when in doubt, choose the simplest model possible