



Lecture 8: Classifiers II

Intro to Data Science for Public Policy
Spring 2017

Jeff Chen + Dan Hammer

Roadmap

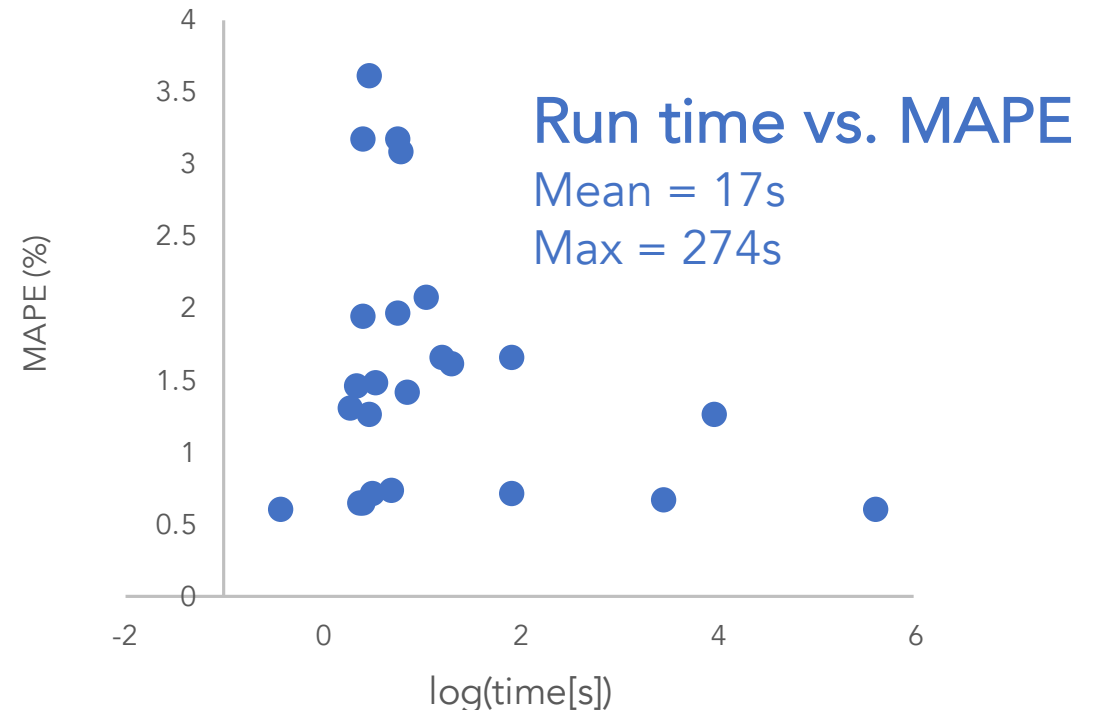
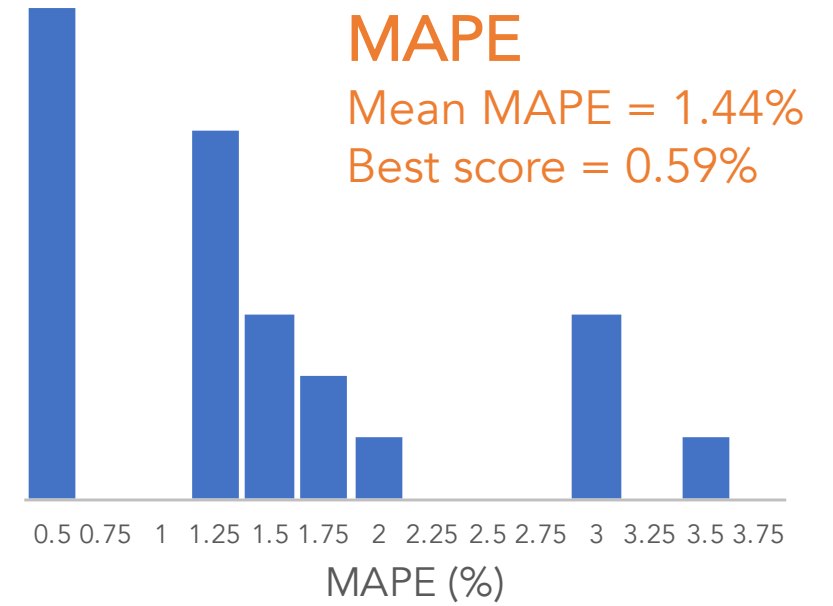
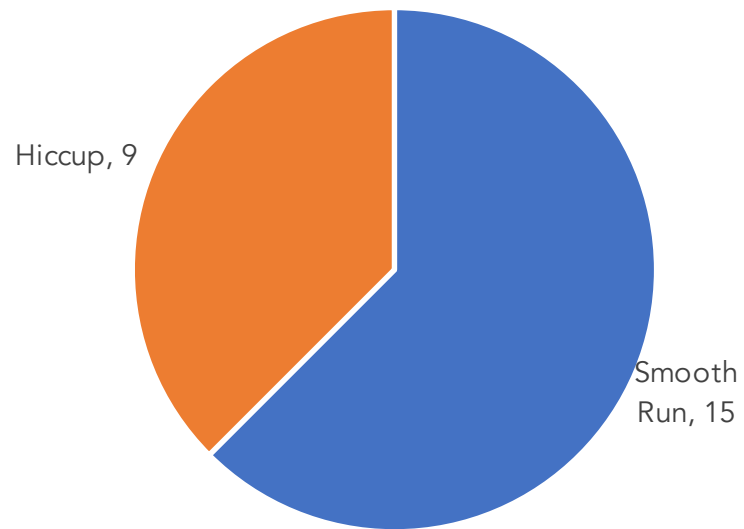
- Homework #2
- How to think about applying data science
- Logistic Regression
- <break>
- Support Vector Machines
- Neural Networks (if we have time)
- Homework assignment

Homework #2: Scores

Scores	Tabulation	Proportion
7	2	8.3%
8	3	12.5%
9	6	25.0%
10	13	54.2%

Homework #2: Performance

Ran without Error



Homework #2: Agreement

GEOID	MSA	Cnt	% of Class
48660	Wichita Falls, TX (Metro Area)	15	63%
27980	Kahului-Wailuku, HI (Micro Area)	12	50%
10780	Alexandria, LA (Metro Area)	10	42%
35380	New Orleans-Metairie-Kenner, LA (Metro Area)	9	38%
17980	Columbus, GA-AL (Metro Area)	8	33%
19100	Dallas-Fort Worth-Arlington, TX (Metro Area)	8	33%
29860	Laurel, MS (Micro Area)	8	33%
46980	Vicksburg, MS (Micro Area)	8	33%

Homework #2: Why Did We Do This?

- Focus on the technique of dividing data and scoring
- Develop skill in testing models at scale

Roadmap

- Homework #2
- Pause to think about applying types of data analytics
- Logistic Regression
- <break>
- Support Vector Machines
- Neural Networks (if we have time)
- Homework assignment

descriptive analytics

using data to describe what happened in the past

explanatory analytics

using data and statistical methods to test a hypothesis pertaining to a theory

predictive analytics

using data to predictive the future

prescriptive analytics

using predictions to drive specific targeted actions

Analytics in this context refers to data analysis and science.

Note that in tech industry, web development, and marketing, analytics often refers to web analytics (e.g. click, views, buys).

summary stats from car sensors

Descriptive.

tests that indicate the existence of
relationships among sensors

Explanatory.

the algorithms that predicts relationships

Predictive.

driverless car that can sense danger,
navigate, and drive using predictions

Prescriptive.

summary stats of fires by building type

Descriptive.

logistic model -- odds ratios tell stories

Explanatory.

random forest and SVM to produce low
error, reliable predictions

Predictive.

predictions used to auto-schedule risk-
based inspections

Prescriptive.

mean, median, standard deviation

Descriptive. simple measures and cross tab to provide context

hypothesis tests, QED, evaluations

Explanatory. Testing if phenomena exist, find a relationship, tell a story.

low-error models, algorithms

Predictive. Machine learning techniques to reliably anticipate and replicate phenomena

policies, auto-behavior

Prescriptive. Software integration, new policies, standard operating protocols that use predictions to make automated decisions

Lectures 1 to 3

Descriptive. simple measures and cross tab to provide context

Lecture 4, most of quant. econ.

Explanatory. Testing if phenomena exist, focus on finding a relationship, not a prediction.

Lectures 5 to 9

Predictive. Machine learning techniques to reliably anticipate and replicate phenomena

Lectures 10 to 13

Prescriptive. Software integration and standard operating protocols that use predictions to make automated decisions

Where does experimental design play into this?
How about feature selection?
How predictive accuracy?
How about interpreting coefficients?
How about application of the models?

Where does
experiment
design play
into this?

How about
interpreting
coefficients?

How about
feature
selection?

How predictive
accuracy?

descriptive analytics

using data to describe what happened in the past

explanatory analytics

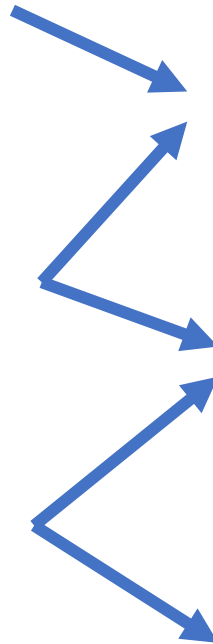
using data and statistical methods to test a hypothesis pertaining to a theory

predictive analytics

using data to predictive the future

prescriptive analytics

using predictions to drive specific targeted actions



How do we interpret results of other model types if there aren't linear coefficients?

Context. Note that prediction models might not be the same model used for explanation.

- Marginal effects (like in regression): Score a set of data where all features except one are held at the mean or mode. The one feature that is held out will illustrate relative effects.
- Accuracy benchmarks in public policy will translate into people and dollars. 1% regression error can mean \$1 billion. 1% in TPR may be 10 people.

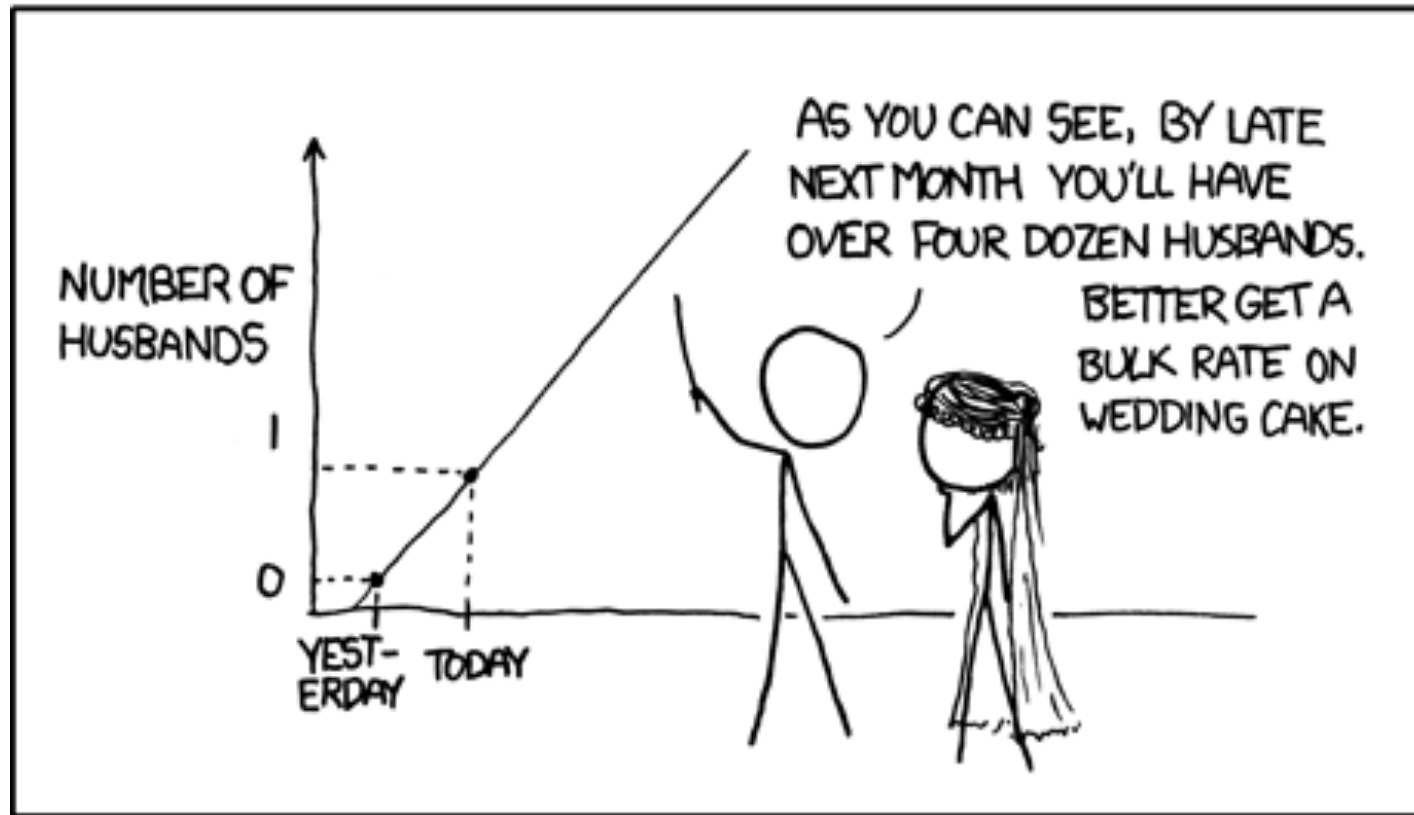
Why should I use non-linear methods?

Many times, non-linear methods can capture more of the signal.

```
| "GLM : 0.729213526935792"  
| "SVM : 0.745670551667047"  
| "Decision Tree : 0.760410180902045"  
| "rf : 0.759588238892838"  
| "KNN : 0.490571415299874"
```

Most academics and practitioners conflate explanatory, predictive and prescriptive. There isn't much discussion or distinction as few outside of tech industry have implemented prescriptive systems.

MY HOBBY: EXTRAPOLATING

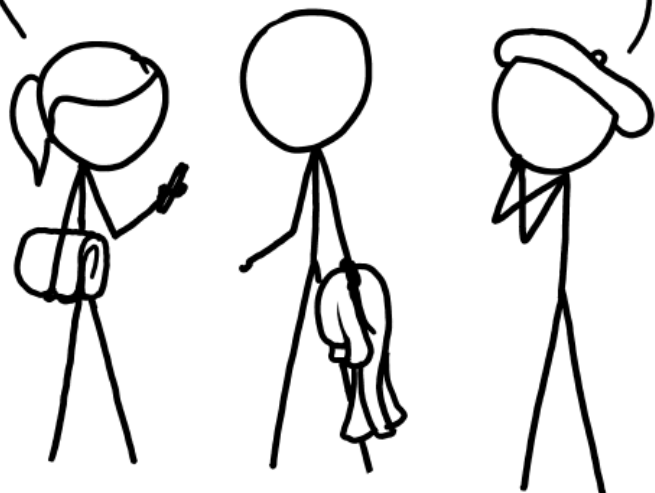


Never use simple descriptive or explanatory analytics for prescriptive tasks. Prescriptive analytics requires reliable predictions.

WE SHOULD GO TO THE NORTH BEACH.
SOMEONE SAID THE SOUTH BEACH HAS
A 20% HIGHER RISK OF SHARK ATTACKS.

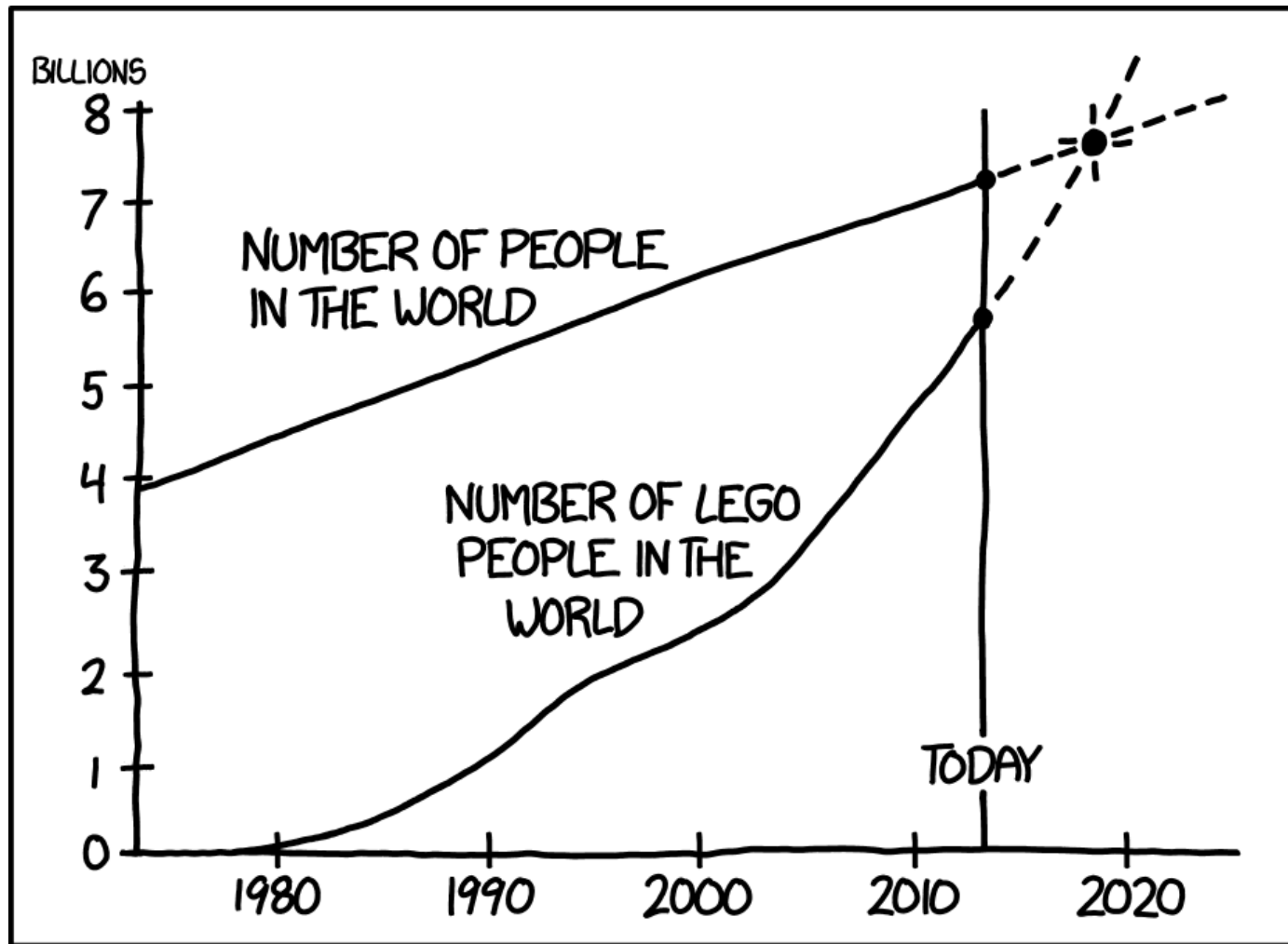
YEAH, BUT STATISTICALLY, TAKING
THREE BEACH TRIPS INSTEAD OF TWO
INCREASES OUR ODDS OF GETTING
SHOT BY A SWIMMING DOG CARRYING
A HANDGUN IN ITS MOUTH BY **50%**!

OH NO! THIS IS
OUR THIRD TRIP!



REMINDER: A 50% INCREASE
IN A TINY RISK IS **STILL TINY**.

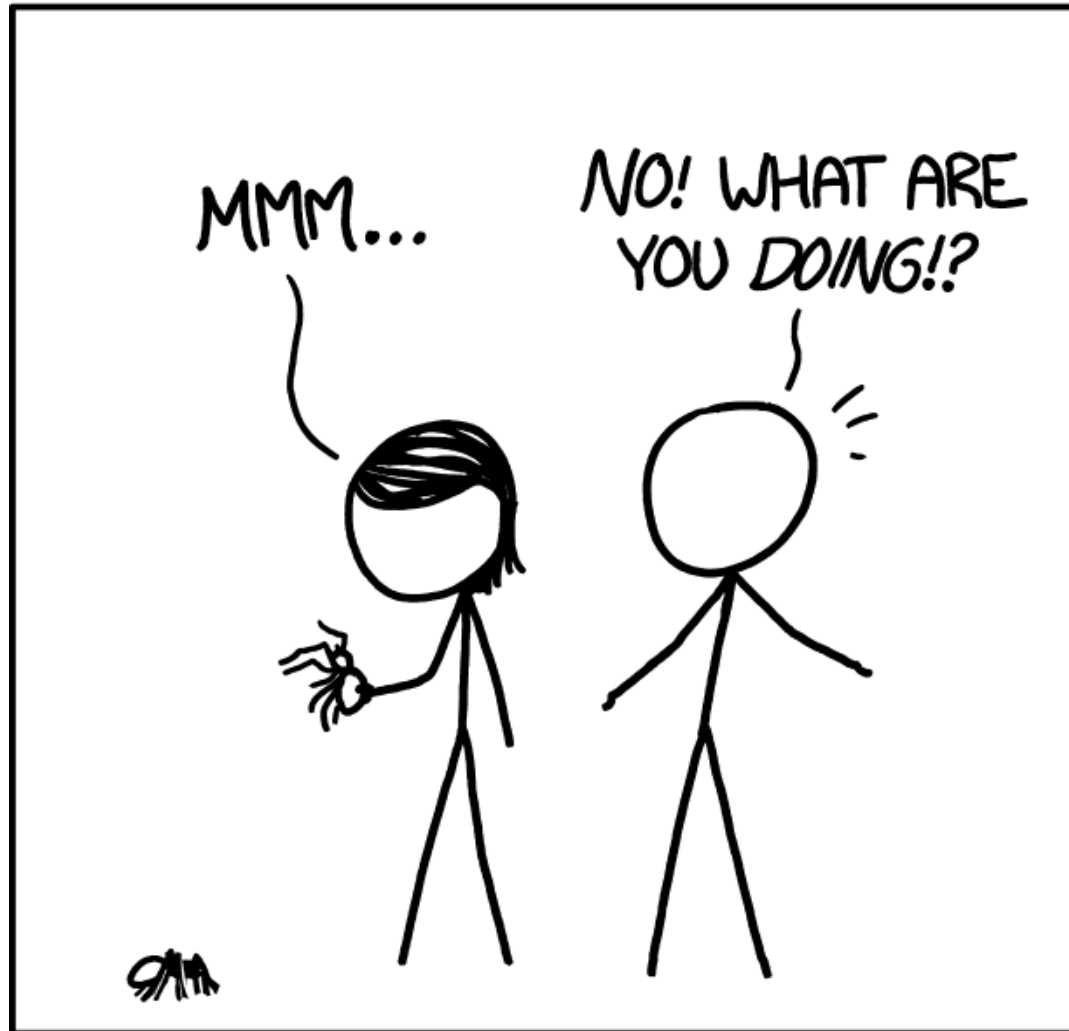
Remember, just cause
your odds ratio in a
logistic regression is
statistically significant
does not mean you can
predict a phenomenon.



BY 2019, HUMANS WILL BE OUTNUMBERED.

Predictive Analytics only works if the paradigm of the past continues into the future.

IMAGINE YOU WERE TRANSPORTED TO AN
ALTERNATE UNIVERSE JUST LIKE YOUR OWN,
EXCEPT PEOPLE OCCASIONALLY ATE SPIDERS.
YOU CAN'T CONVINCE ANYONE THIS IS WEIRD.

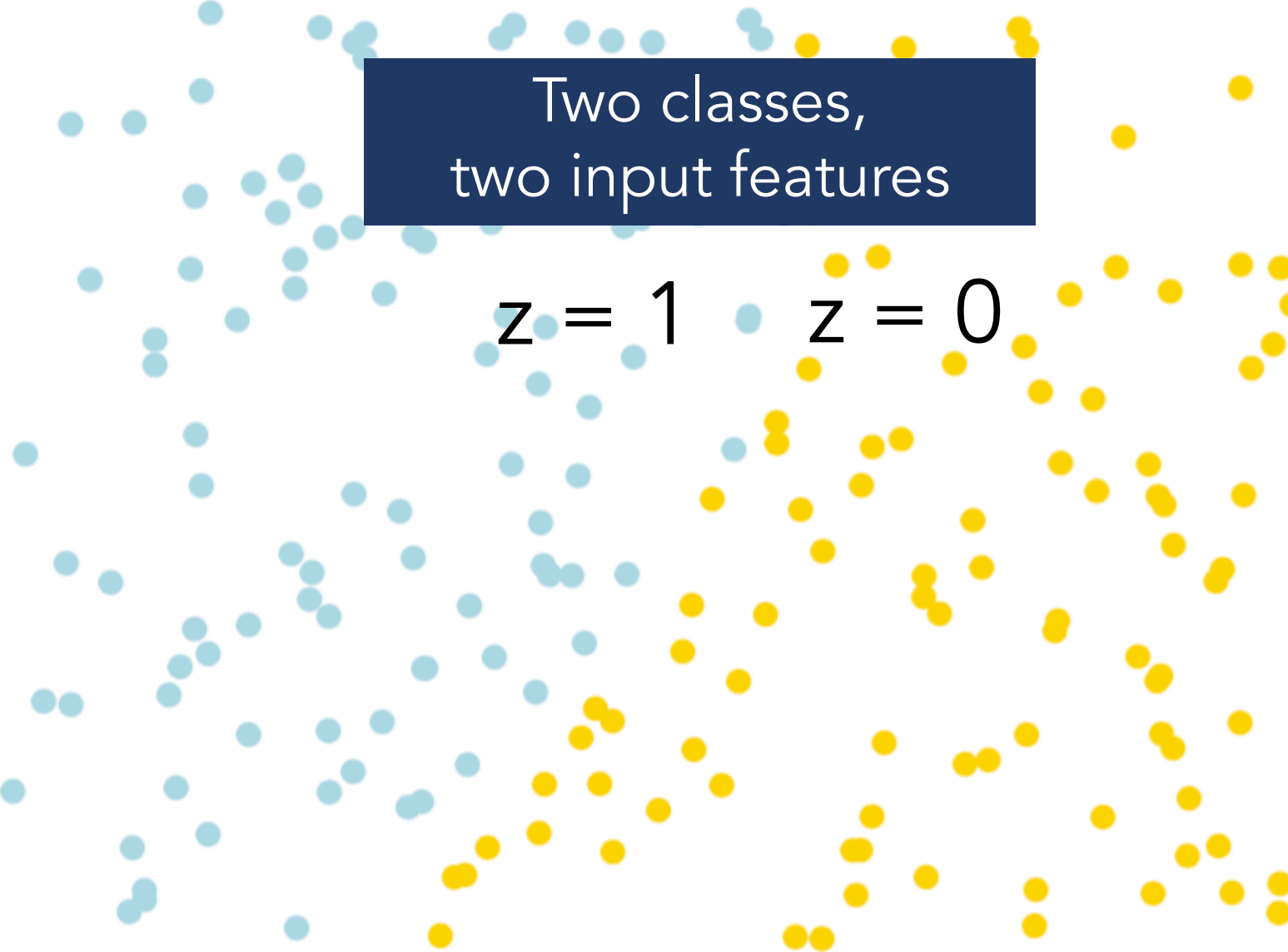


And this is how I feel when
policymakers choose to
use descriptive analytics
rather than letting the
process run its course –

No! What are you doing?!

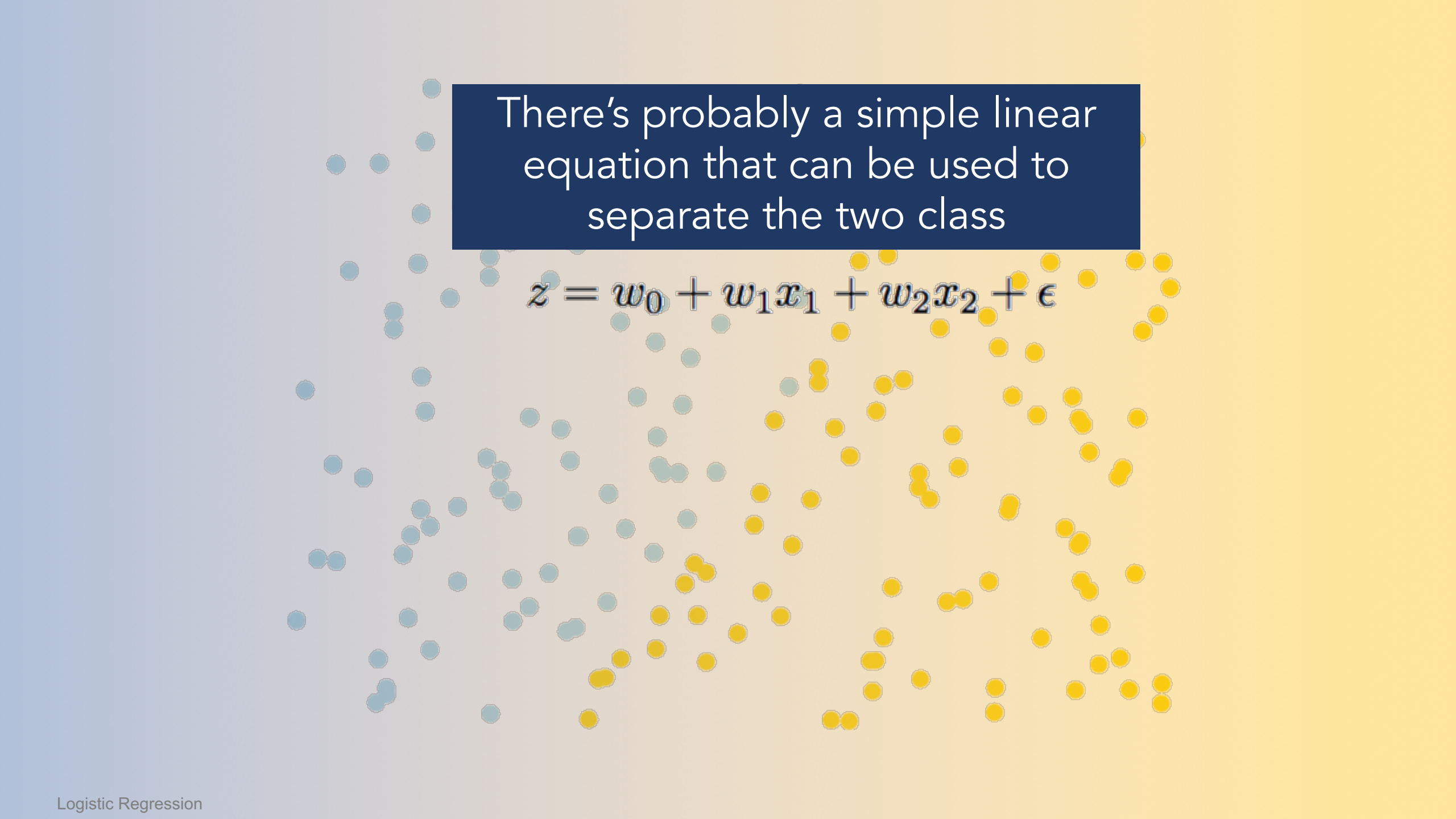
Roadmap

- Homework #2
- Pause to think about applying types of data analytics
- **Logistic Regression**
- <break>
- Support Vector Machines
- Neural Networks (if we have time)
- Homework assignment

A scatter plot showing two classes of data points on a 2D plane. The first class is represented by light blue dots, and the second class is represented by yellow dots. The points are distributed across the plot, with some overlap between the two classes. A dark blue rectangular box is positioned in the upper-middle part of the plot, containing white text. Below the box, the labels $z = 1$ and $z = 0$ are placed, with a light blue dot under $z = 1$ and a yellow dot under $z = 0$.

Two classes,
two input features

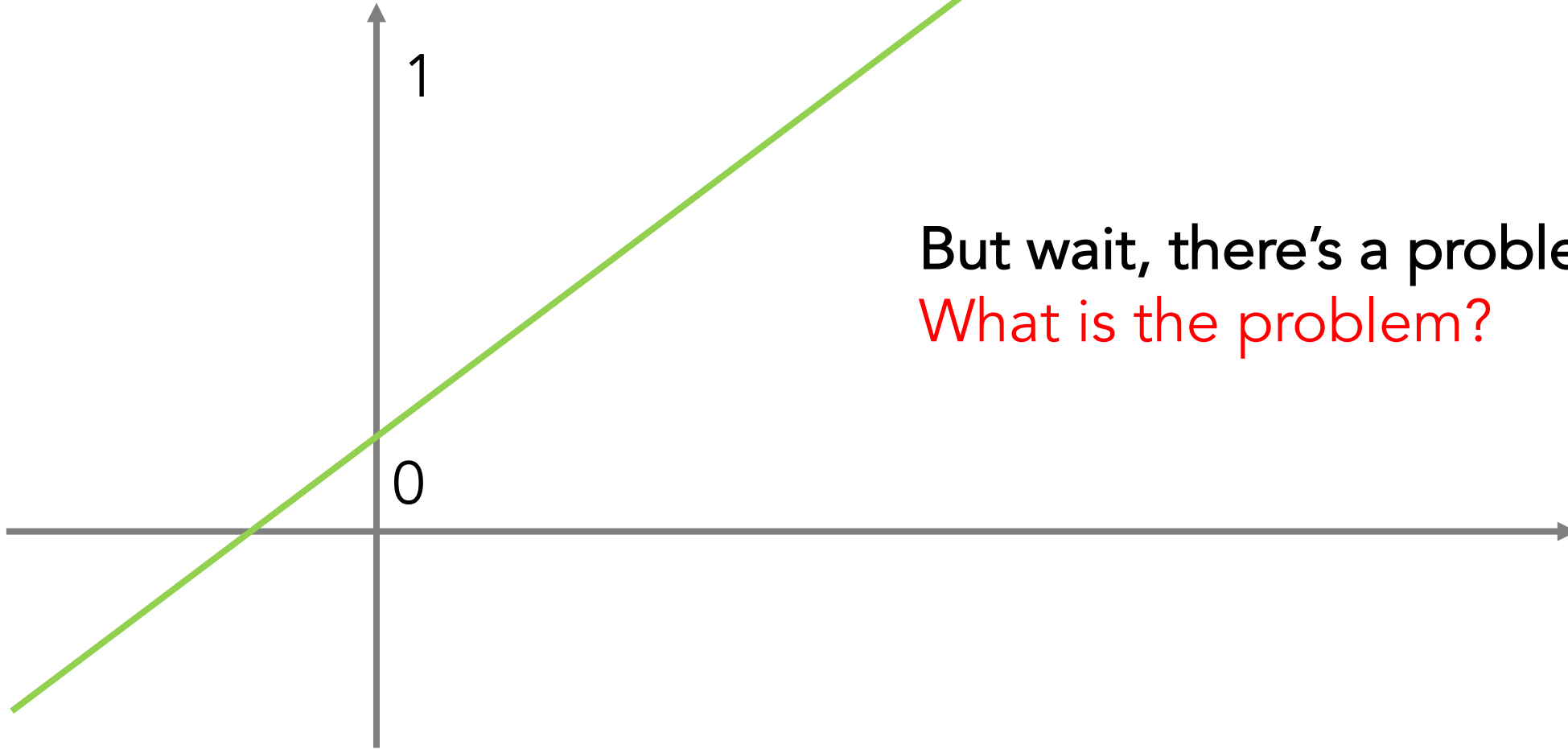
$z = 1$ $z = 0$

A scatter plot on a light blue to yellow gradient background. The plot contains two classes of data points: blue circles and yellow circles. The blue circles are primarily located on the left side of the image, while the yellow circles are primarily on the right side. There is a small region of overlap in the center. A dark blue rectangular box is positioned in the upper-middle part of the plot, containing white text. Below the box, the linear equation $z = w_0 + w_1x_1 + w_2x_2 + \epsilon$ is written in a black serif font.

There's probably a simple linear equation that can be used to separate the two class

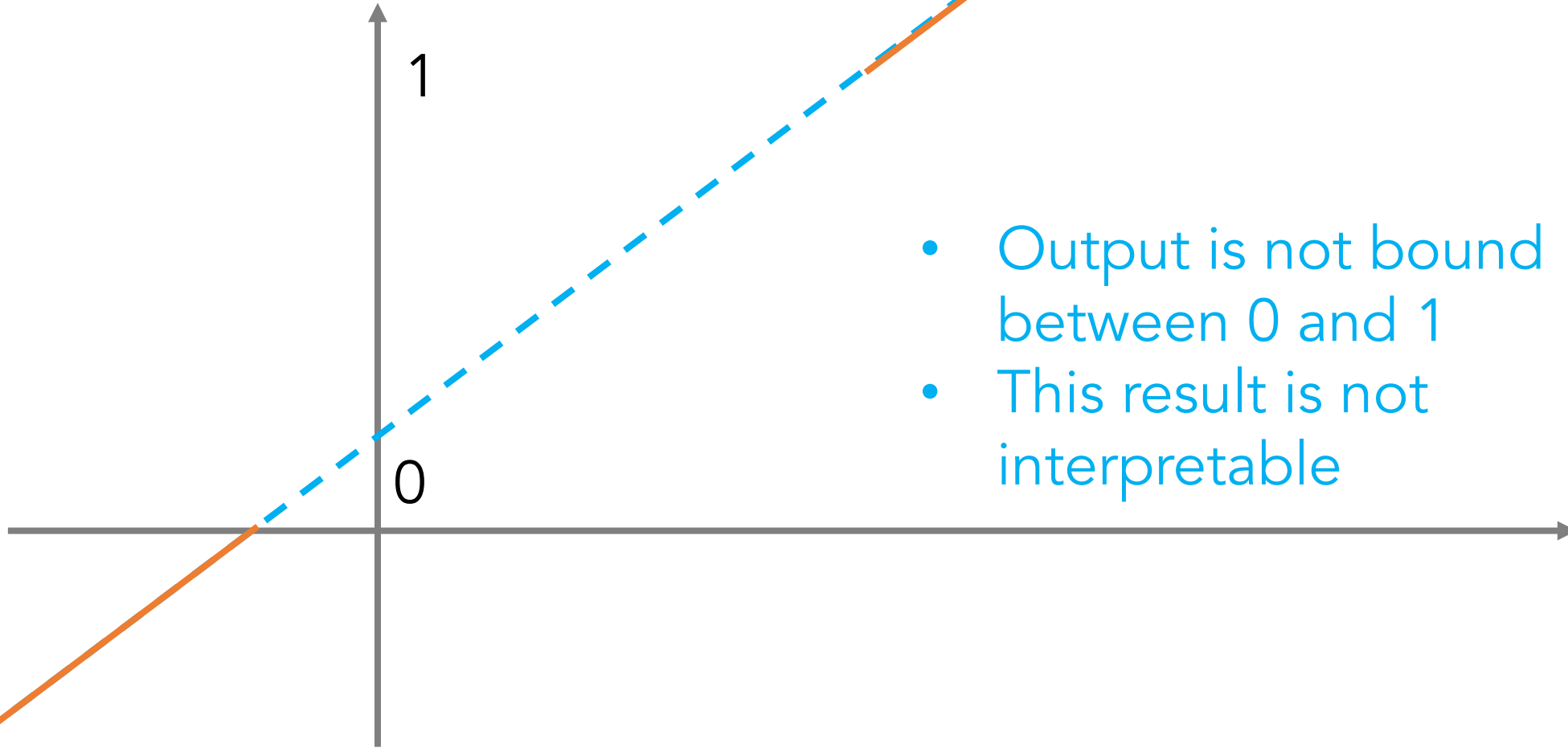
$$z = w_0 + w_1x_1 + w_2x_2 + \epsilon$$

$$z = w_0 + w_1x_1 + w_2x_2 + \epsilon$$



But wait, there's a problem.
What is the problem?

$$z = w_0 + w_1x_1 + w_2x_2 + \epsilon$$

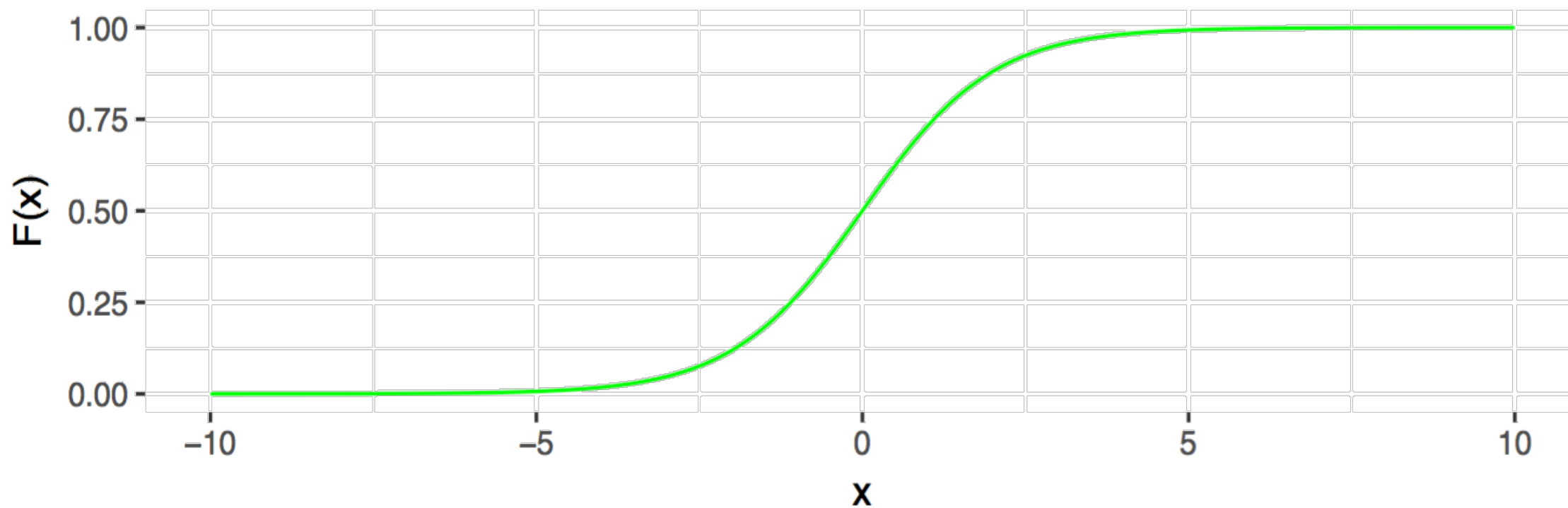


- Output is not bound between 0 and 1
- This result is not interpretable

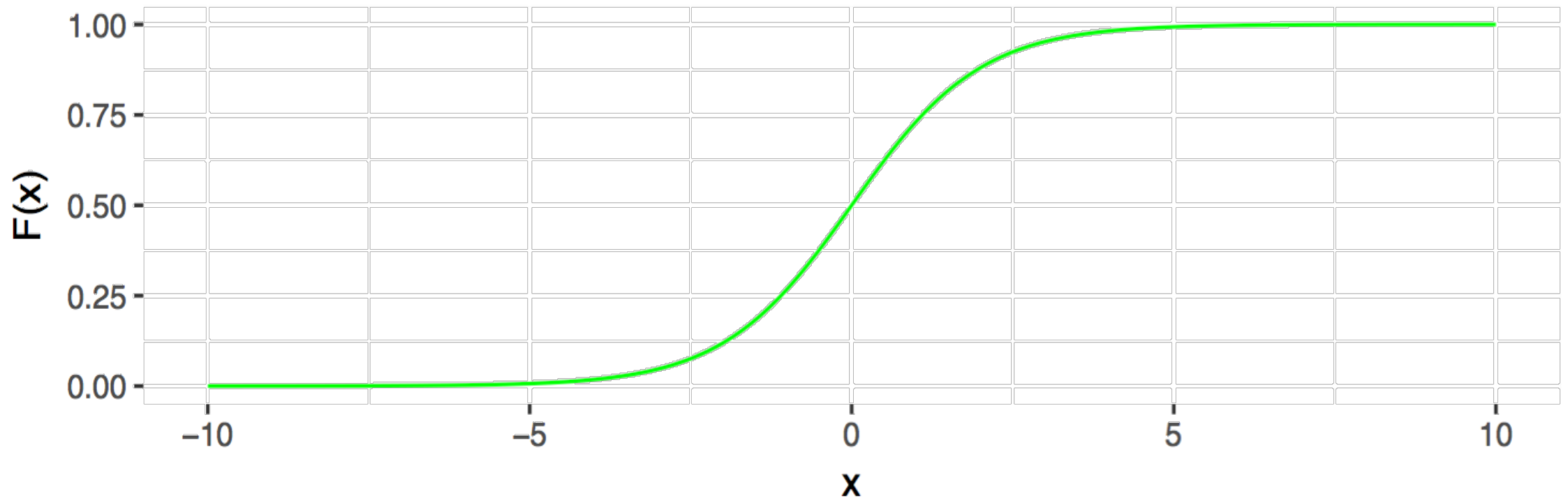
Solution:

Plug $z = w_0 + w_1x_1 + w_2x_2 + \epsilon$ into $F(z) = \frac{1}{1 + e^{-z}}$ to get

$$Pr(Y = 1|X) = F(z) = \frac{1}{1 + e^{-(w_0 + w_1x_1 + w_2x_2)}}$$



What does this sigmoid function do?



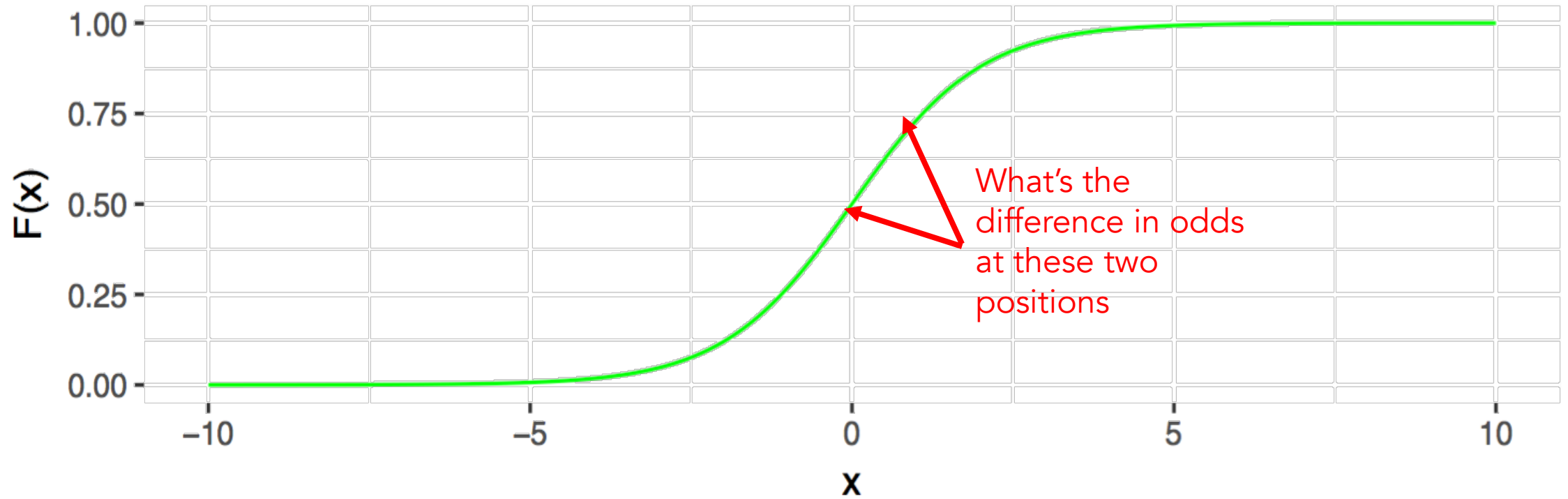
As this is a linear function,

$$Pr(Y = 1|X) = F(z) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + w_2 x_2)}}$$

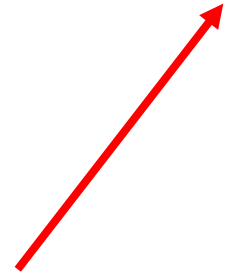
we should be able to interpret the coefficients.
To do so we first define odds (success p divided by failure):

$$odds = \frac{p}{1 - p} = \frac{F(z)}{1 - F(z)} = e^z$$

The odds ratio compares two sets of odds where one set examines the case where $X = 1$ and $X = 0$.



Which can be expressed in terms of the following:

$$OR = \frac{e^{w_0 + w_1(x_1 + 1) + w_2 x_2}}{e^{w_0 + w_1(x_1 + 0) + w_2 x_2}} = e^{w_1}$$


In other words: Just exponentiate the coefficient.

Interpretation

OR = 0.9 = 10% less likely

OR = 4 = 300% more likely or 4x higher chance

OR = 0.1 = 90% less likely

OR = 1 = No significant difference

Logistic regression is optimized using Maximum Likelihood Estimation (MLE)

Single likelihood

$$p(z = z_i | x) = [F(x)]^{z_i} [1 - F(x)]^{1-z_i}$$

Likelihood
(Product of all obs)

$$L = \prod_{i=1}^N [F(x)]^{z_i} [1 - F(x)]^{1-z_i}$$

Log-Likelihood
(Log transformed as
exponents are hard)

$$\log(L) = z_i \log(F(x)) + (1 - z_i) \log(1 - F(x))$$

Solve: Partial derivative of
L with respect to $w = 0$

$$\frac{\partial L}{\partial w_k} = 0$$

Assumptions

- The usual suspects: Independence of error terms, extraneous variables are omitted, etc.
- Records should not be perfectly separable.
- Sample sizes need to be large for each target class (MLE is weak if targets are small). Minimum of 30 to 40 cases per parameter estimated.
- Tuning is largely a matter of feature selection: it all depends on the variables that are available.

The Good, The Bad and The Ugly

Good

- The technique is well-suited for socializing an empirical problem.
- Method of choice for binary problems in economics, sociology and public policy, and applied medical.

Bad

- Often is outperformed in accuracy by more flexible techniques such as Random Forests.
- Limited to linear assumptions. Weak with non-linear problems
- Use tends to be prone to high bias (bias-variance tradeoff)

Ugly

- None unless the technique is misused.

Common uses

- Medical: Relative risks of medicine use
- Economics: Choice models for transport
- Politics: Chance of demographic leaning in one direction
- Program evaluation: Estimated program effect

<Code Time/>

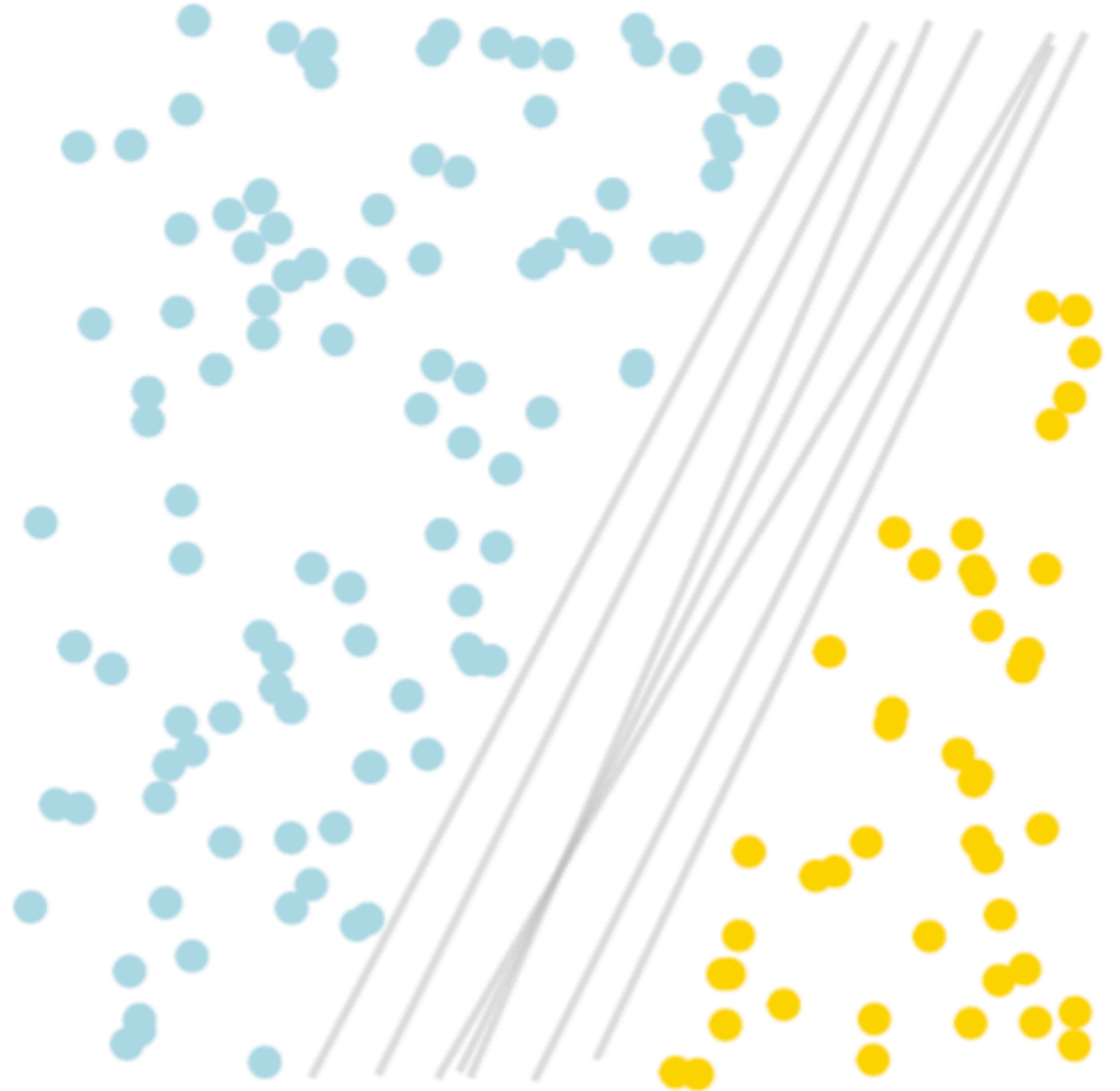
Roadmap

- Homework #2
- Pause to think about applying types of data analytics
- Logistic Regression
- <break>
- Support Vector Machines
- Neural Networks (if we have time)
- Homework assignment

Let's assume a modified case where we have two classes that are separated by a gap.



There should be some line that can separate the classes. But which?



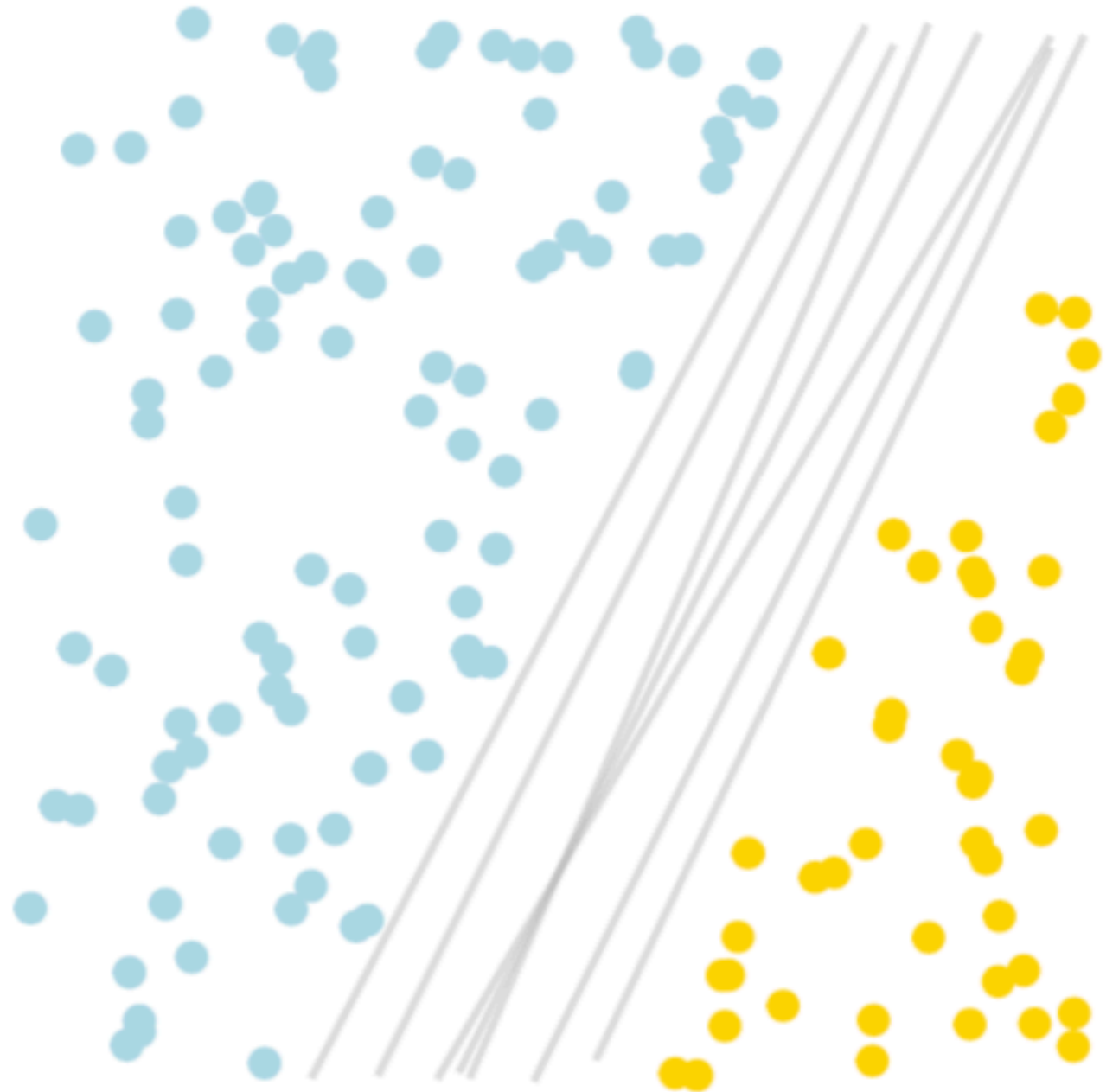
Geometry Refresher:

One line is written as

$$y = mx + b$$

A line in greater in 2D
is a plane:

$$y = w^T x - b$$

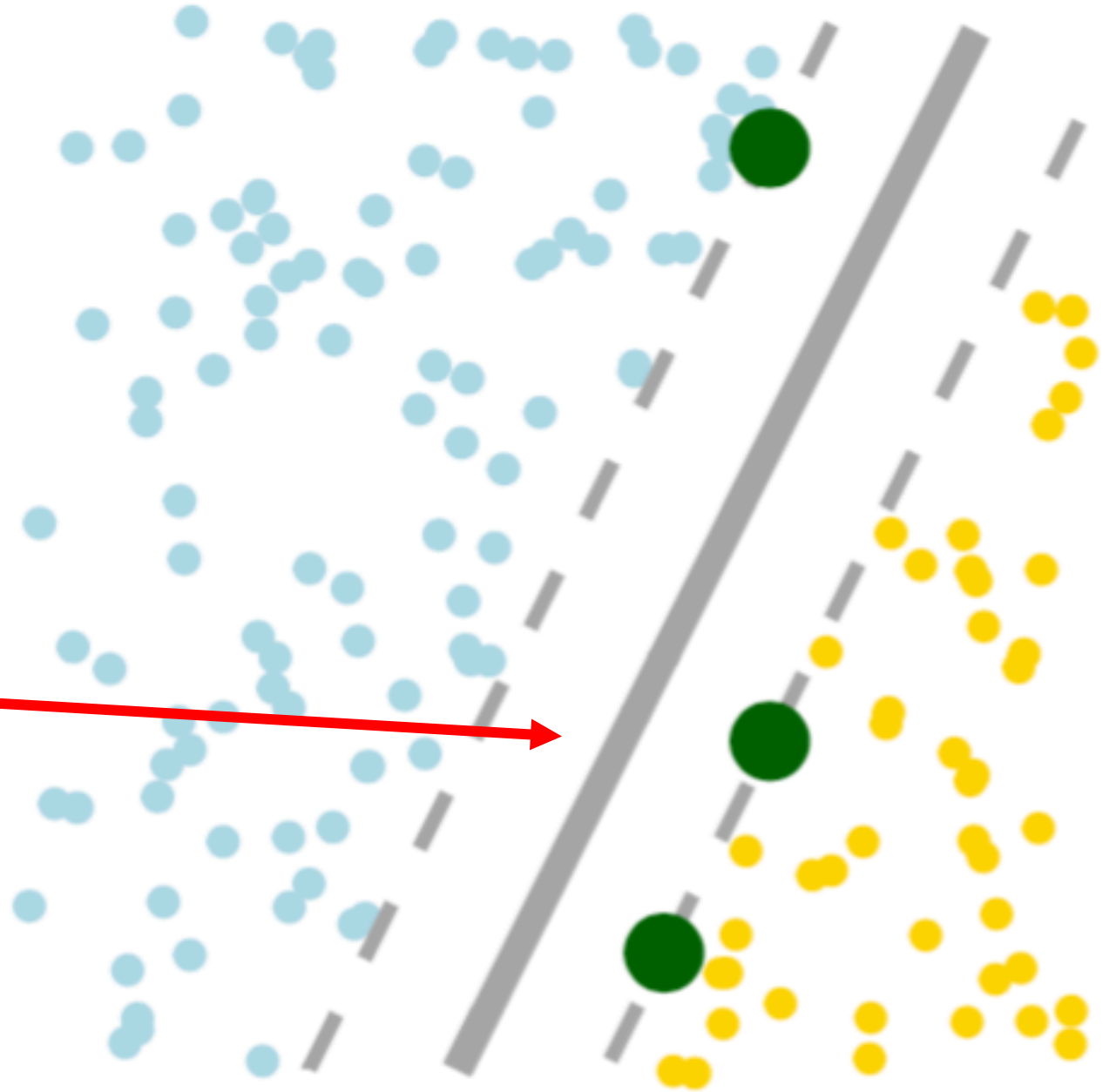


Goal:

Maximize the distance
between each class.

Gap is known as the
margin

$$\text{margin} = x_1 - x_2 = \frac{2}{||w||}$$



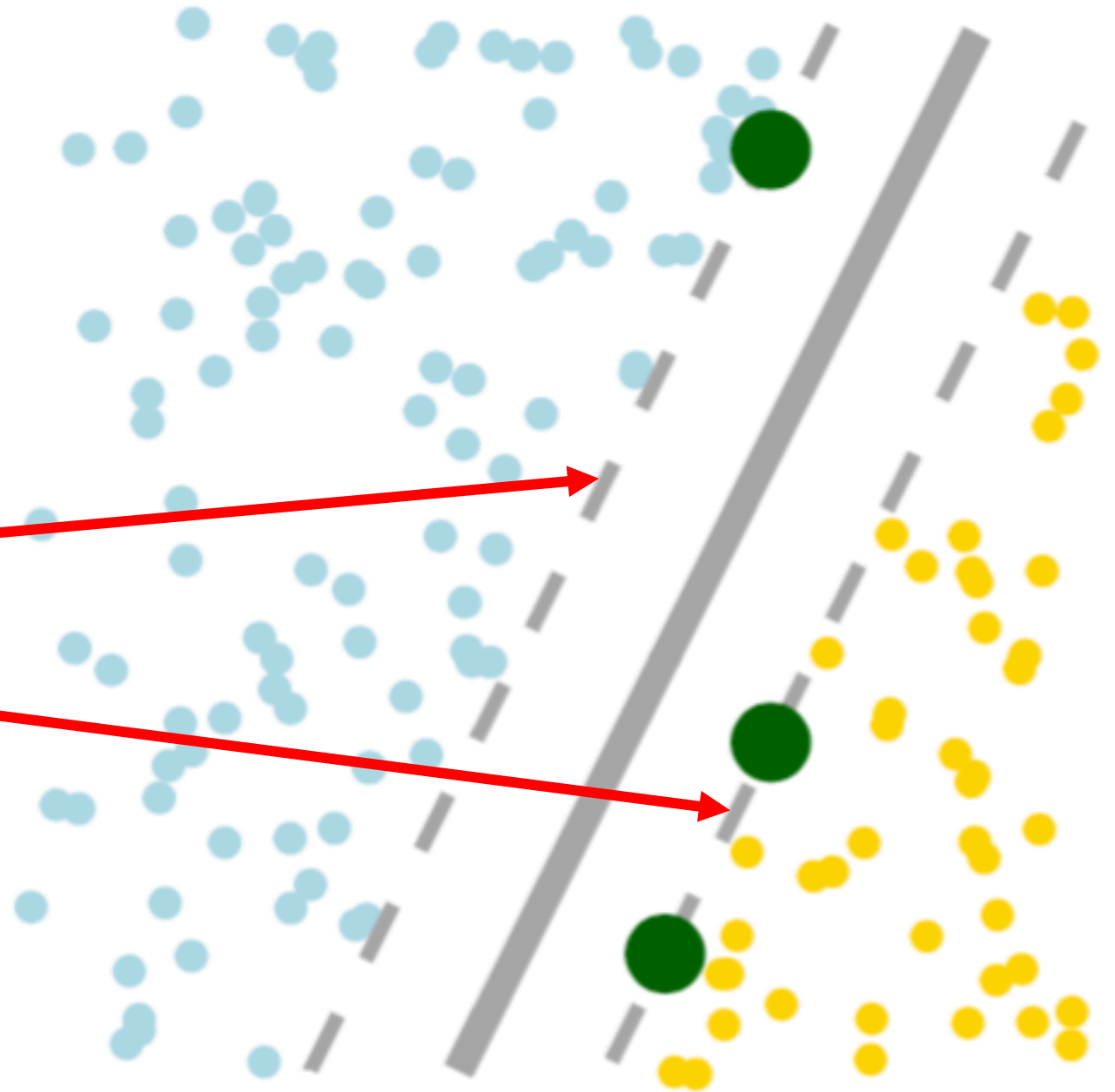
How:

Find hyperplanes that sit on the edge of each group

$$w^T x - b = +1$$

$$w^T x - b = -1$$

Note that points may not sit in the margin.



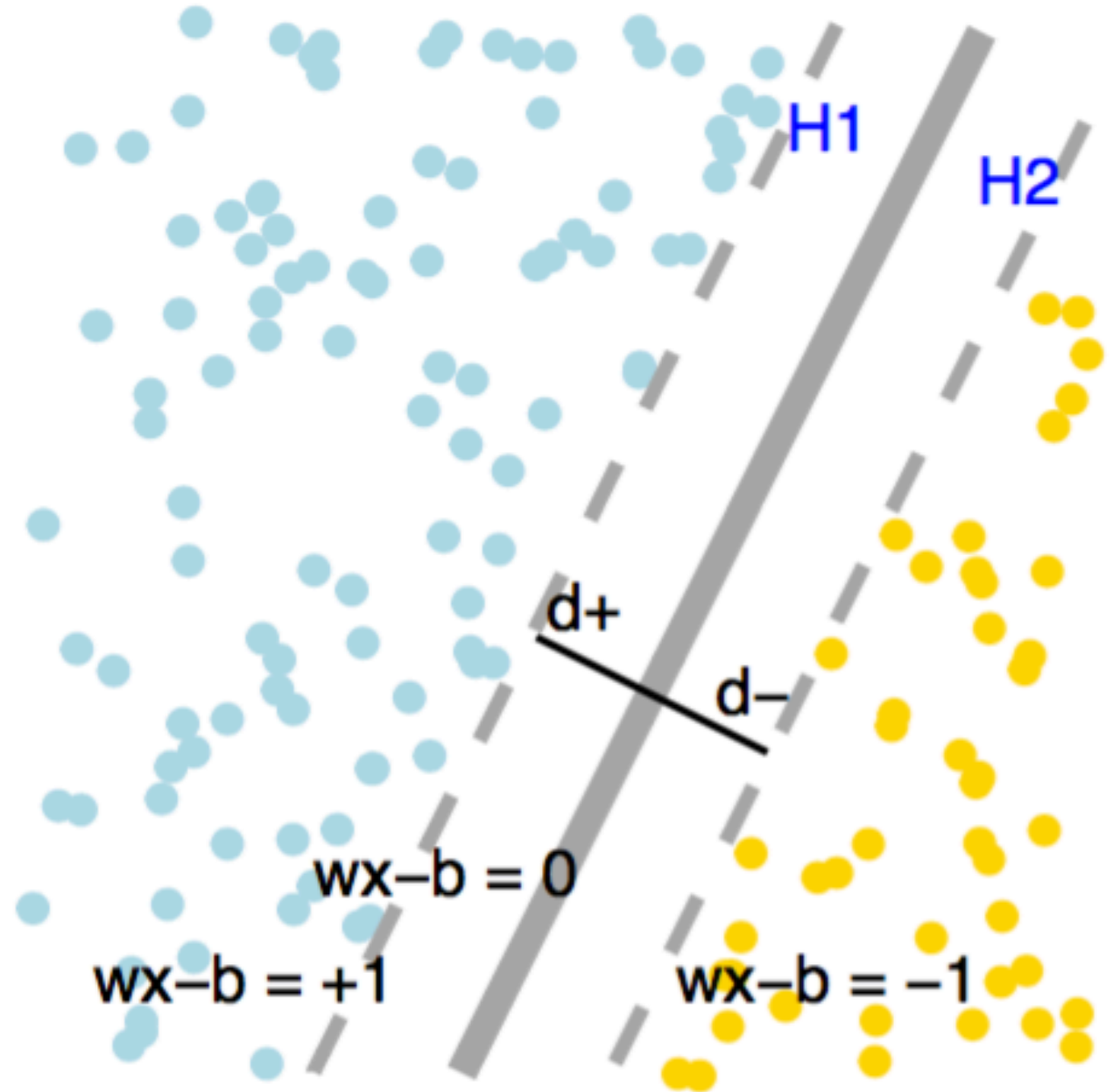
The decision plane is given by the plane that is equidistant from the two decision planes.

$$w^T x - b = 0$$

Where the decision criteria are:

$$+1 = w^T x - b > 0$$

$$-1 = w^T x - b < 0$$



We find the optimal hyperplanes by maximizing the margin

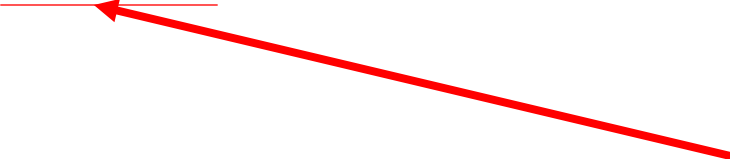
$$\text{margin} = x_1 - x_2 = \frac{2}{\|w\|}$$

But it is easier to minimize this transformed version of distance using quadratic programming.

$$\min \frac{1}{2} \|w\|^2 \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1 :$$

Sometimes the $\frac{1}{2}$ is dropped as it's a constant.

The value of w is learned as a function of α .

$$w(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \alpha_1 \alpha_0 y_1 y_0 \underline{x_1^T x_0}$$


Subject to

$$\sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

Alpha is determined through convex optimization

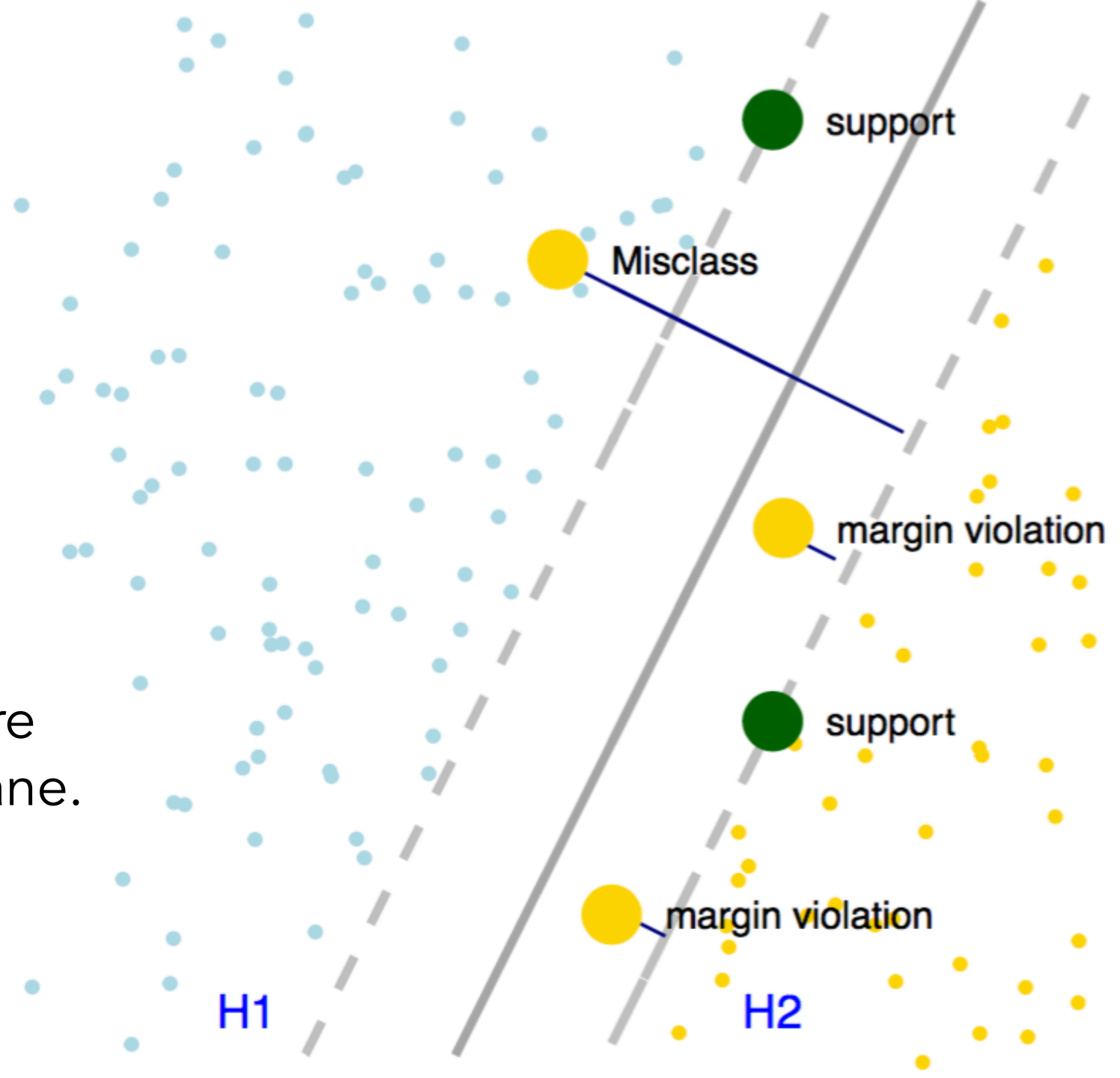
*“Linear Discriminant” Dot Product
Or Sum of the Products*

This is key for modifying the shape of hyperplane.

Honest truth: Most problems don't have clear margins (e.g. hard margin). Most data have 'soft margins'

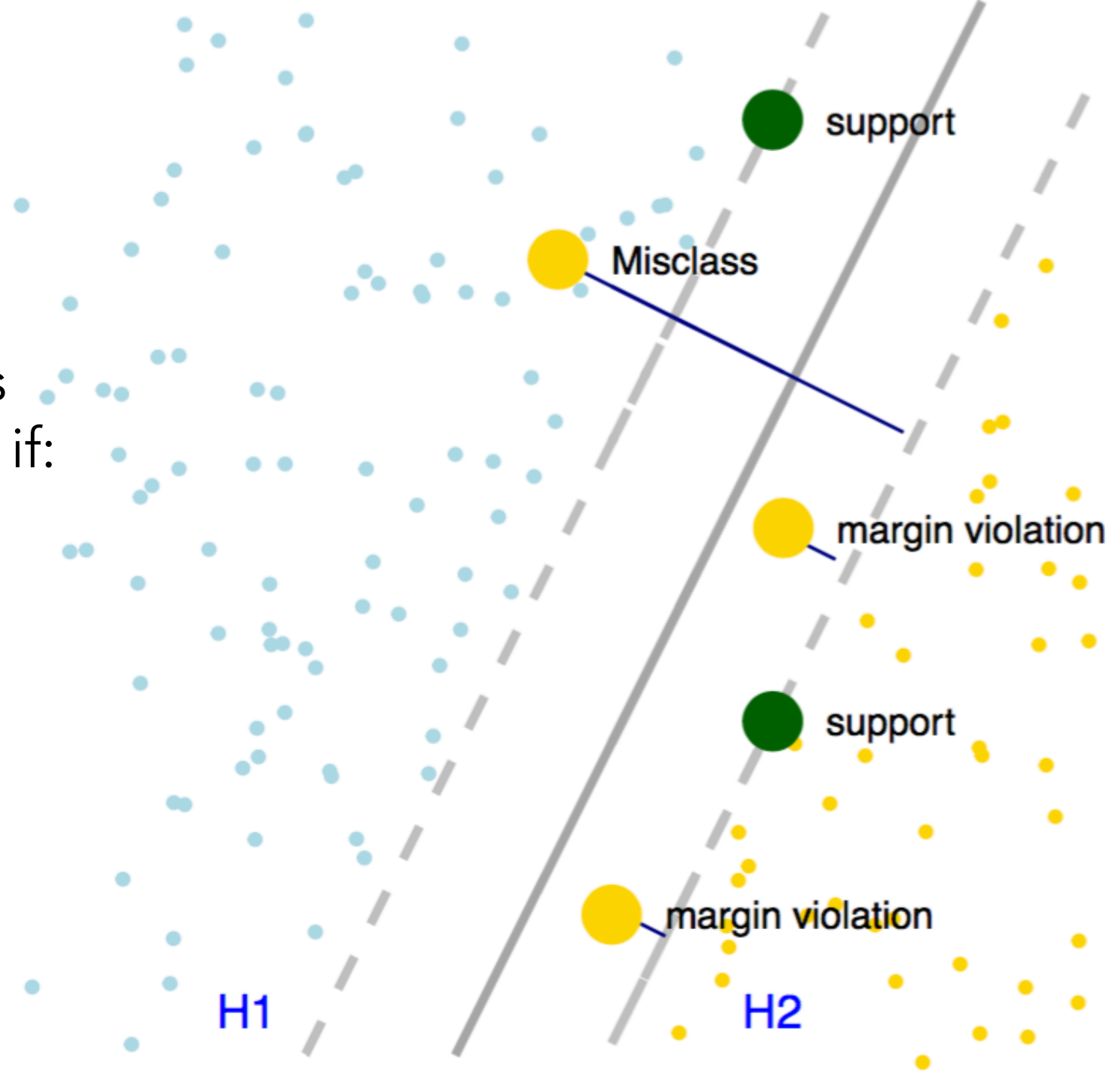
Slack variables ξ_i are a measure of how far misclassified points and margin violation points are from the correct hyperplane.

"How much error can the model take?"



In terms of l_1 , what values do ξ_i slack values take on if:

- margin violation
- misclassification



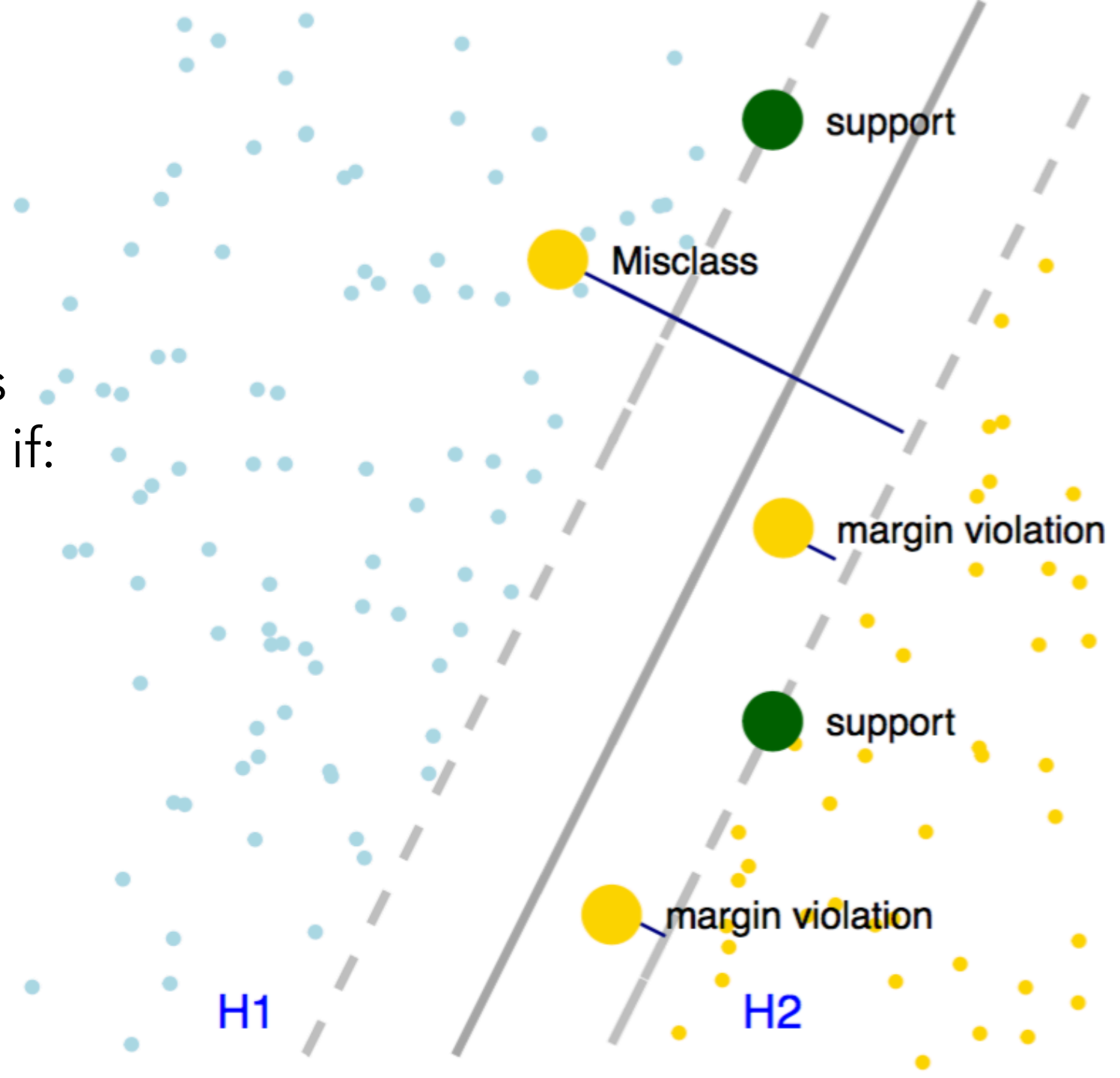
In terms of $\|w\|$, what values do ξ_i slack values take on if:

- margin violation

$$0 \leq \xi \leq \frac{1}{\|w\|}$$

- misclassification

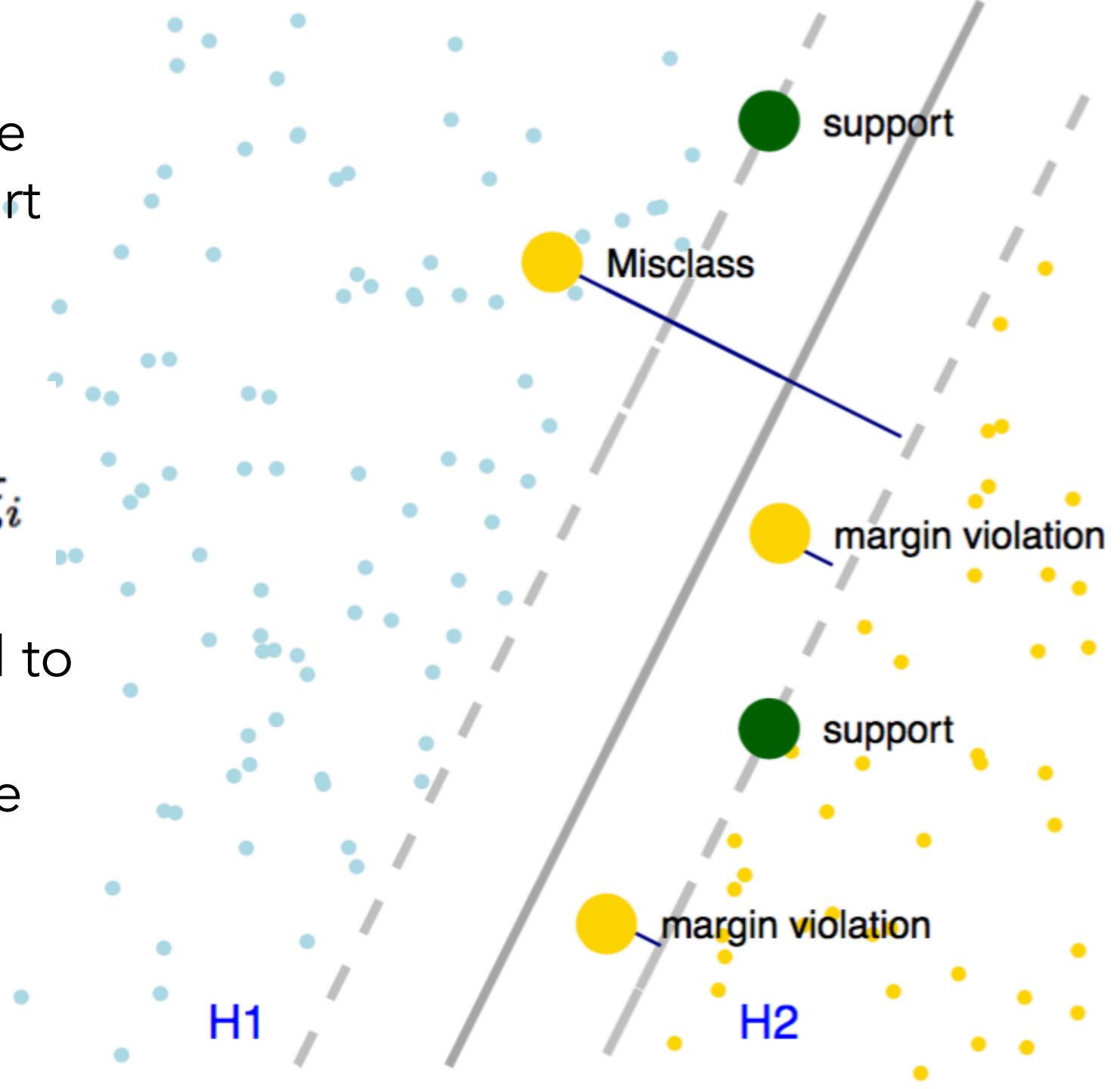
$$\xi > \frac{2}{\|w\|}$$



Slack variables enter into the optimization equation as part of a cost “penalty” – also known as regularization.

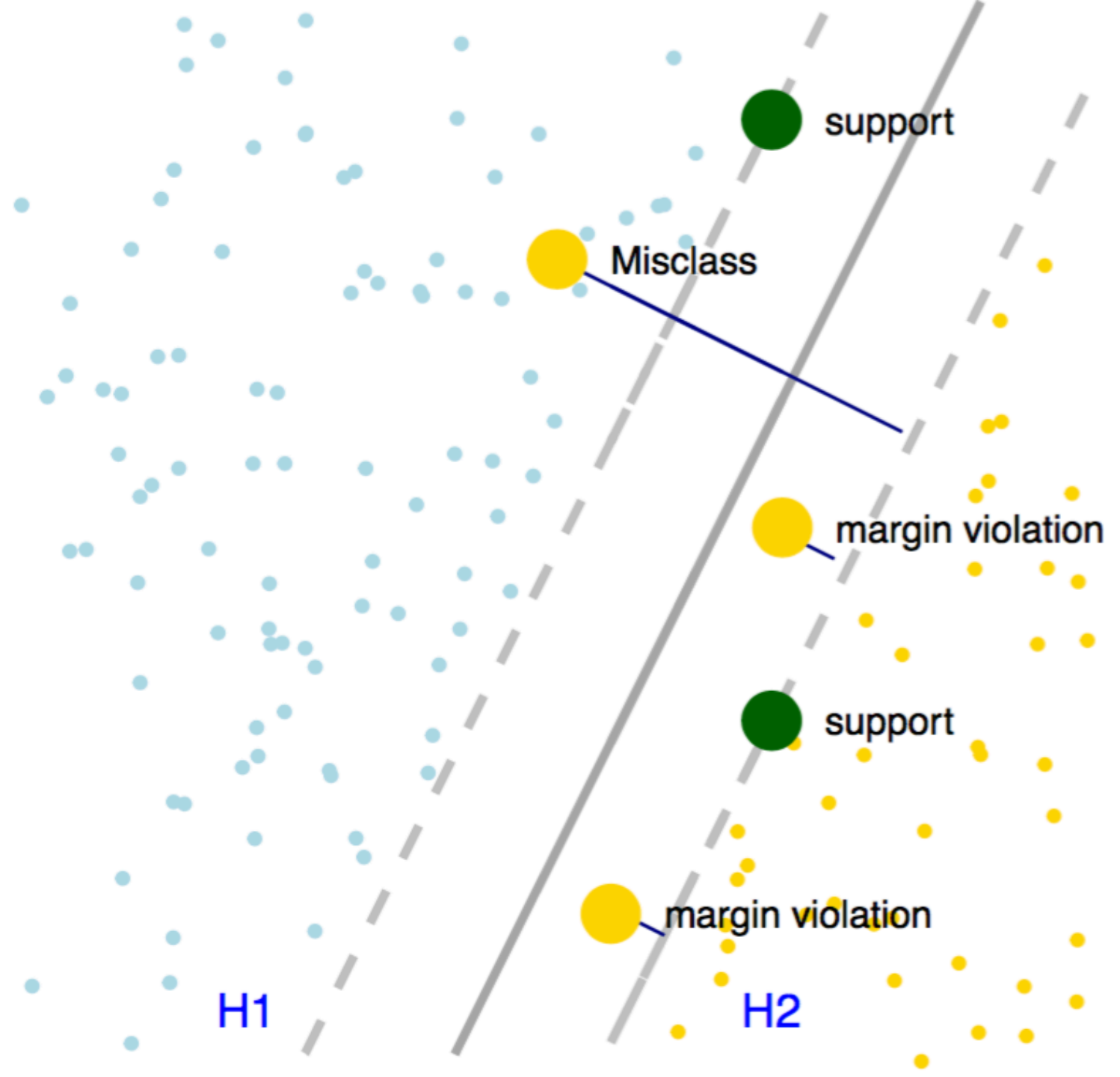
$$\min \frac{1}{2} ||w||^2 + C \sum_i^N \xi_i$$

C is a hyperparameter used to tune how much “slack” is given when determining the margin.



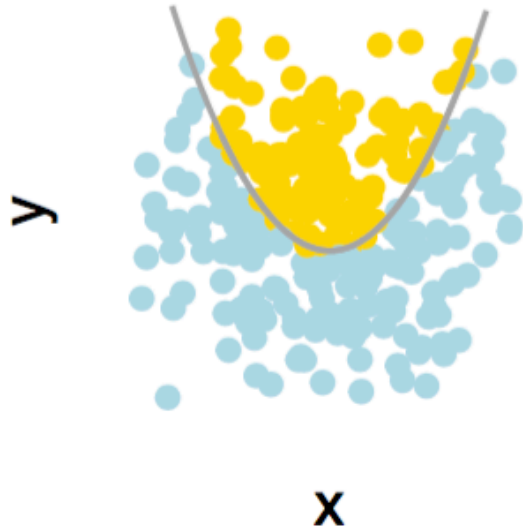
How is C determined?

What does C mean if it's large?

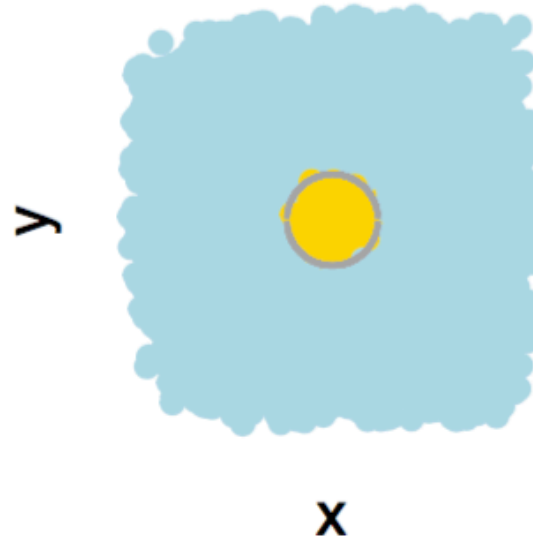


These problems cannot be solved using linear models.

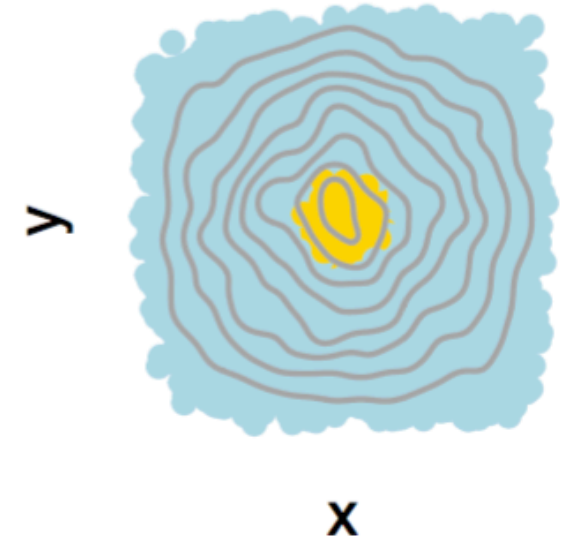
(1) Parabola



(2) Circle



(3) Circle – 3–dimension



The linear dot product can be swapped out for other non-linear relationships to take advantage of curvature and transformations in multiple dimensions.

$$w(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \alpha_1 \alpha_0 y_1 y_0 \underbrace{x_1^T x_0}_{\begin{array}{l} \text{Polynomial} = (1 + x_1^T x_0)^d \\ \text{---} \text{---} \text{---} \text{---} \\ \text{RBF} = \exp(-\gamma \|x_1 - x_0\|^2) \end{array}}$$

Video!

<https://www.youtube.com/watch?v=3liCbRZPrZA>

The Good, The Bad and The Ugly

Good

- Very strong at handling non-linear problems
- Great for problems without a simple explanation (e.g. images classification, text classification, genetics, health research)
- Not sensitive to multicollinearity issues

Bad

- Takes a long time to train in high dimensional space
- Simple explanation is not possible as problem is formulated as a geometric one
- Deriving probabilities may at times require fancy footwork

Ugly

- None unless the technique is misused.

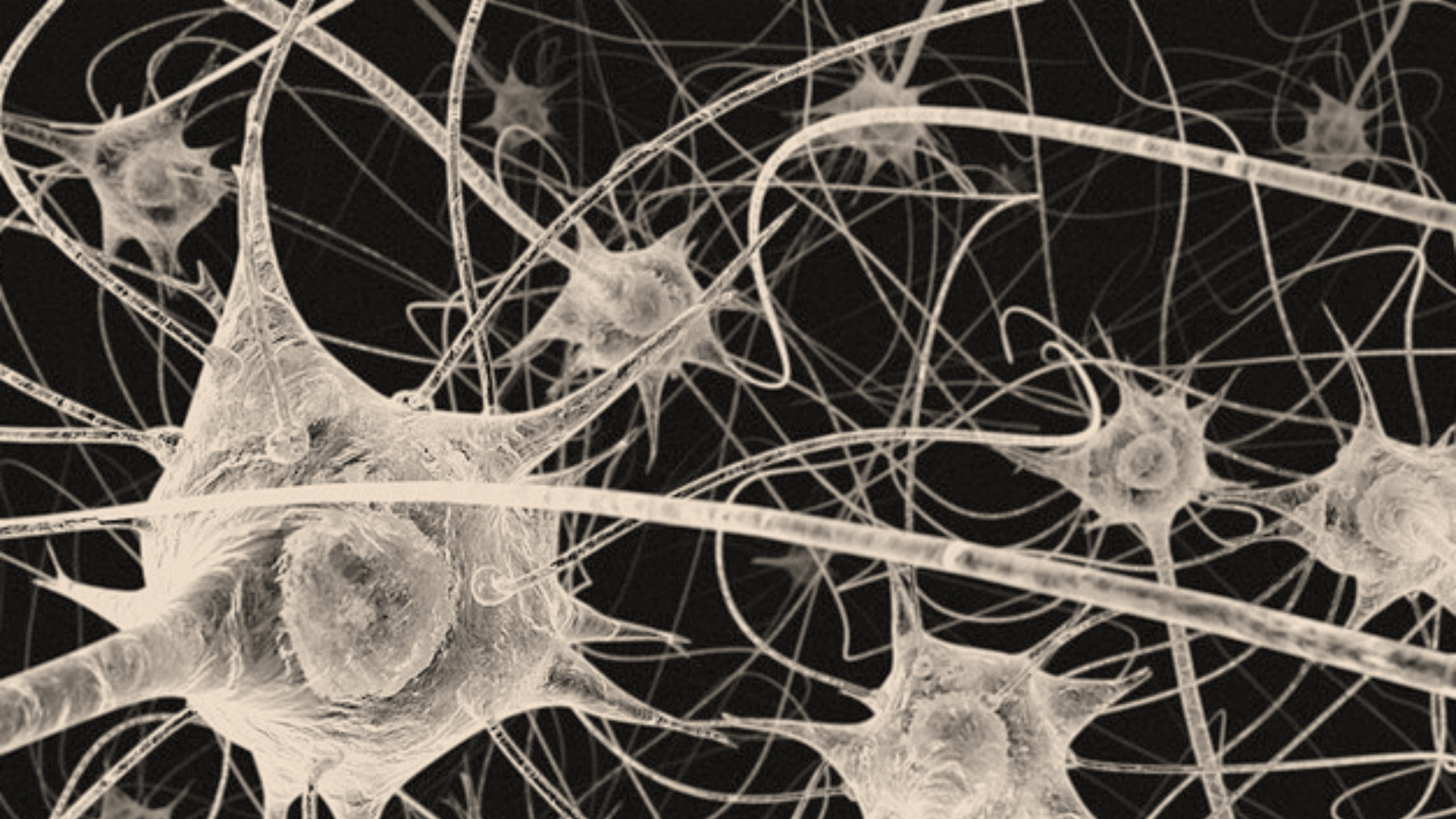
Common uses

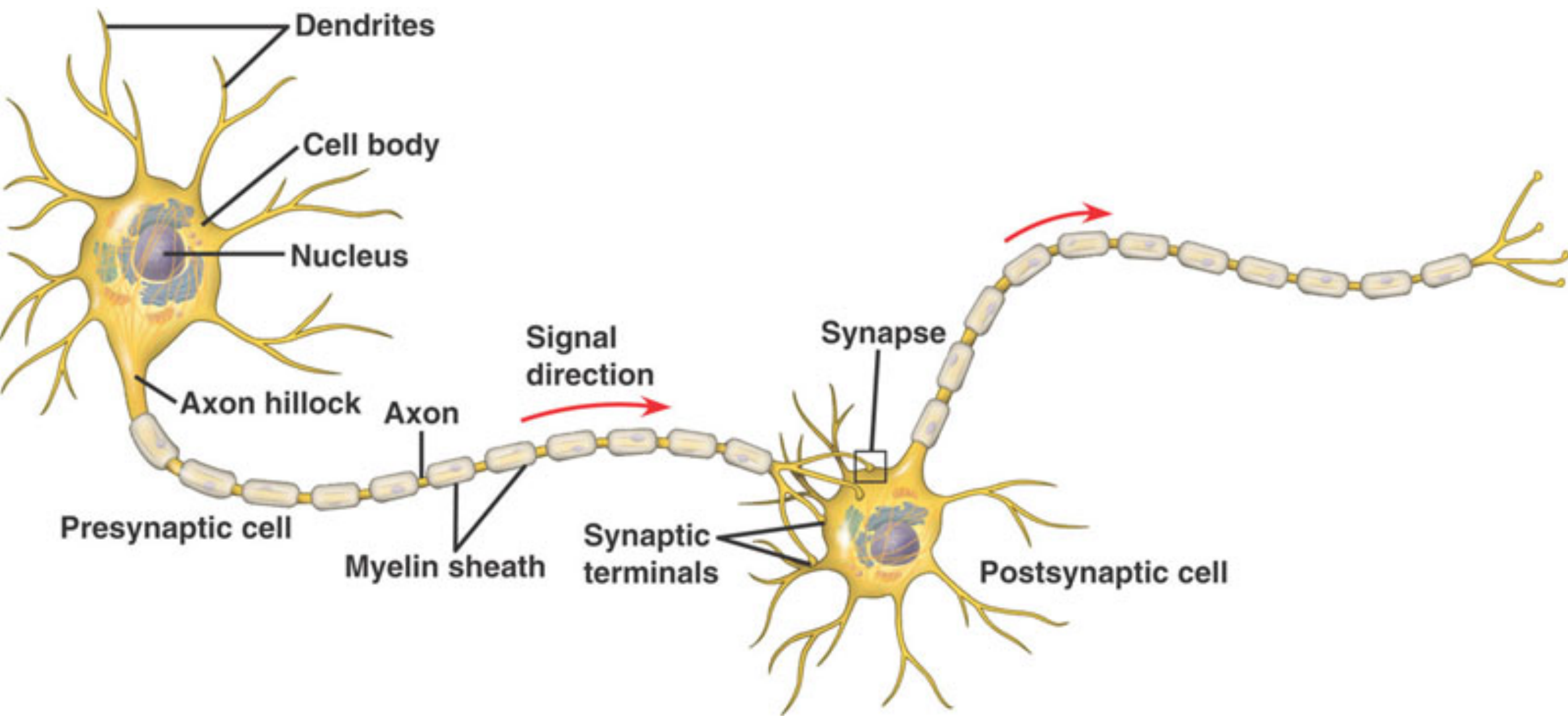
- Pattern recognition in images:
 - Image recognition and retrieval
 - Facial expression recognition
- Biology:
 - Gene expression in high dimensional spaces
 - Detecting protein folds
- Text:
 - Classification of topics
 - Personalized content for e-learning and social media

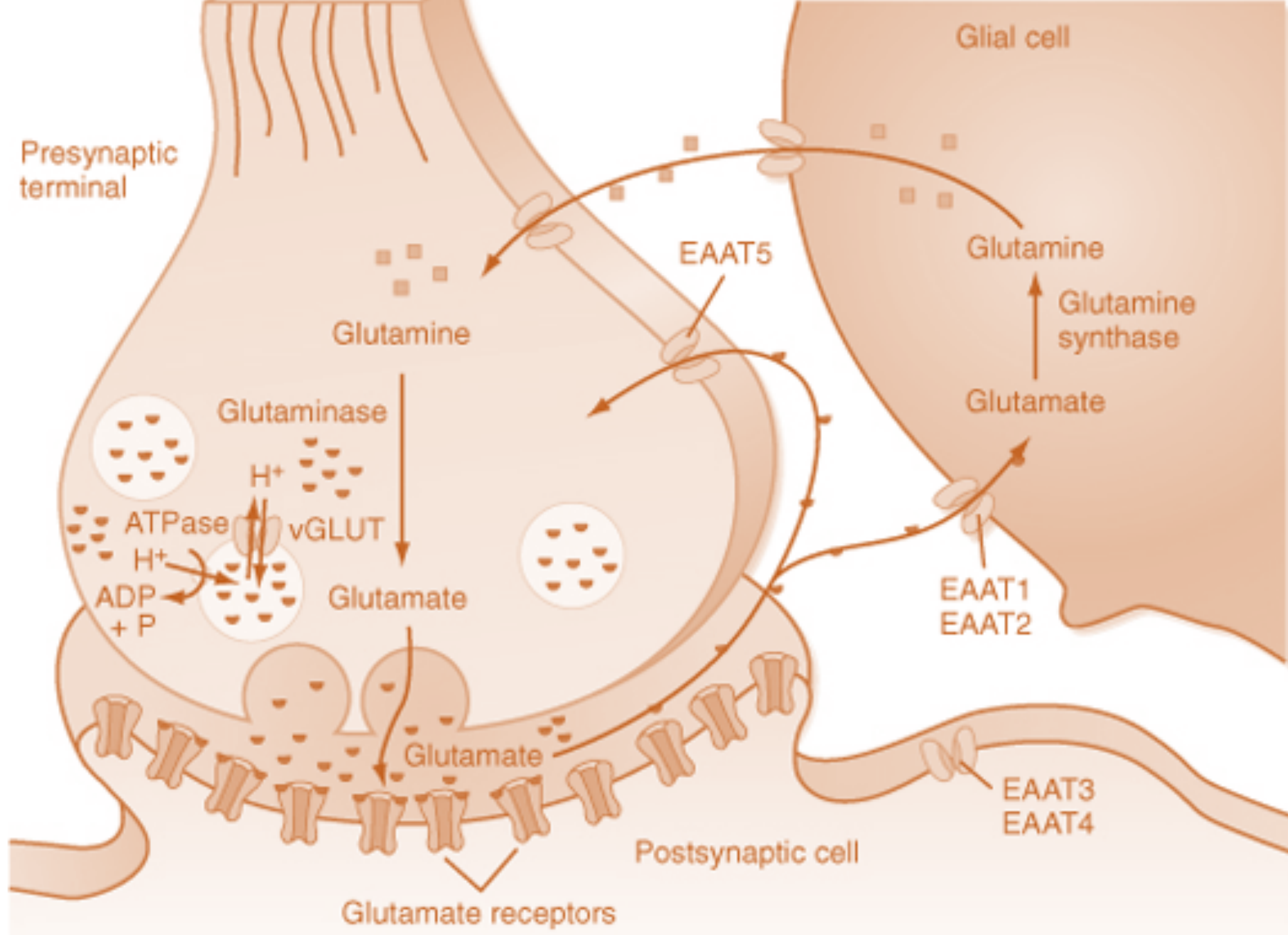
<Code Time/>

Roadmap

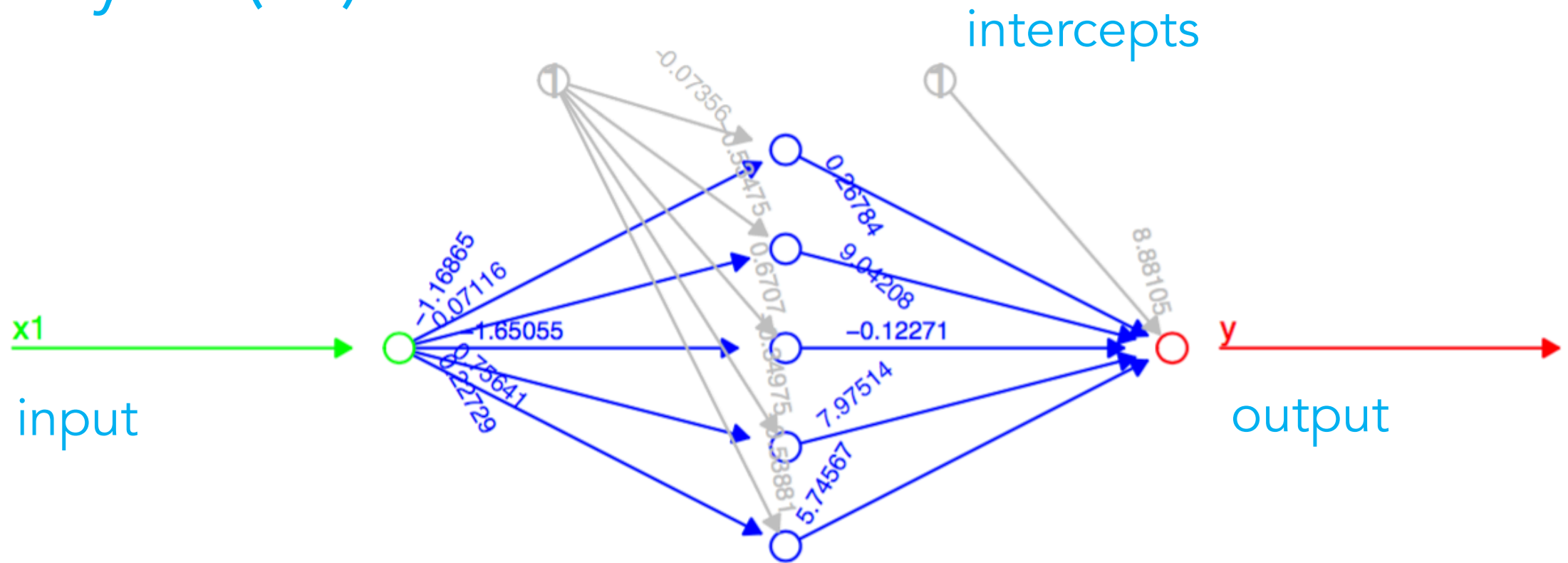
- Homework #2
- Pause to think about applying types of data analytics
- Logistic Regression
- <break>
- Support Vector Machines
- Neural Networks (if we have time)
- Homework assignment





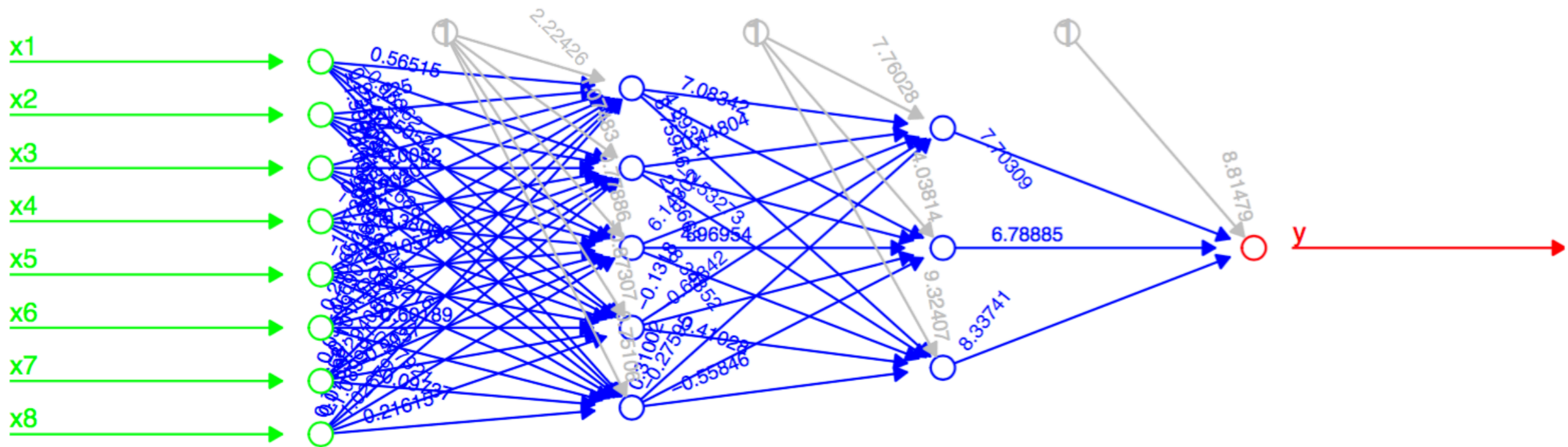


$$y = f(x_1)$$



hidden layer = controls non-linearity
each node

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$$



Artificial Neural Networks are the basis of “Deep Learning”.

The Good, The Bad and The Ugly

Good

- Handles large number of features
- Captures non-linear relationships
- Strong performer with missing values
- Used to learn complex patterns

Bad

- Enormous data requirements
- Data requirements grow following a quadratic with each additional node
- Computationally expensive and time consuming

Ugly

- Requires enormous amounts of data
- Determining node and layer architecture is very challenging

Roadmap

- Homework #2
- Pause to think about applying types of data analytics
- Logistic Regression
- <break>
- Support Vector Machines
- Neural Networks (if we have time)
- Homework assignment