ICS 691D: MACHINE LEARNING REVIEW

Peter Y. Washington, PhD

Assistant Professor, ICS

August 24, 2022

THOUGHTS ON OR QUESTIONS STEMMING FROM TODAY'S READING?

EXAMPLES OF REFLECTION PARAGRAPH TOPICS

- What you agreed/disagreed with and why
- Critique of any section, such as (1) the claims in the Introduction and Discussion, (2) the methods used, (3) the evaluation methods used in the Results, etc.
- How you might repurpose something you learned for an area that is of interest to you (including your research proposal assignment)
- How the paper connects to other readings in the class
- The ethical implications of the described technology (would be good to connect to one of the formal ethical frameworks we discussed)
- How the described technology could be designed for real world use and factors to consider when doing so
- …

GASHIPS / POSTDOCS AVAILABLE

http://www.expmed.org/

A. **Using synthetic organic life models to visualize large data sets in a domain and expertise agnostic manner**. Dr. Stokes' research is developing elegant biologically-inspired data processing, simulation and visualization methods to support the S&T, healthcare, and finance sectors. The goal is to demystify complex data and democratize access to the power that these data hold using simulated organic life models. After securing an initial US patent on the core technology, Dr. Stokes is researching applications to biomedical and financial data.

B. **Application of unbiased GEO analytics to cardiovascular disease (CVD) therapeutic development**. Dr. Stokes' GEO Analytics project provides for a new type of 'trans-indication' analytics. It provides for identification of associations between genes, proteins, targets, drugs and mechanisms that are not biased by prior associations. Given the unmet need for new targets in CVD, the project has three goals: (1) Identify and associate new target-gene-mechanism associations for heart failure using advanced analytics and visualization of the GEO database; (2) Re-analyze drugs currently in the cardiovascular pipeline or that have failed developmental stages to identify/explain potential off-target or adverse effects, (3) Provide a broadly applicable online resource for drug discovery that allows unbiased 'trans-indication' analytics of GEO to associate gene-target relationships and mechanisms in ways that overcome prior association biases.

OPPORTUNITY TO USE RESEARCH PROPOSAL FOR RESEARCH WITH ME NEXT SEMESTER AND BEYOND

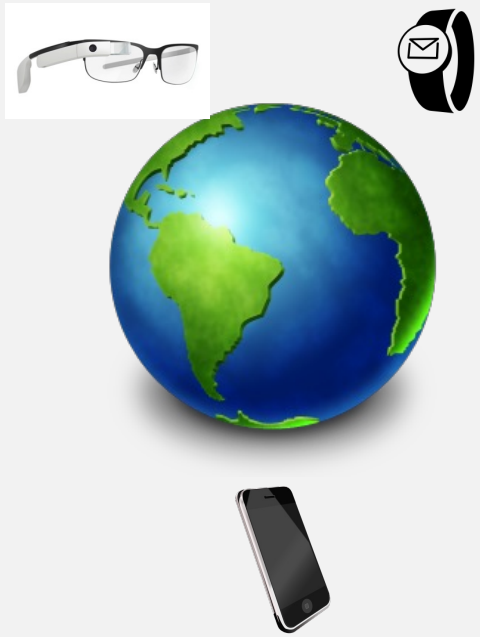
- For BA/BS students
  - ICS 499
  - Undergraduate Research Opportunities Program
- For MS students
  - ICS 699
  - ICS 700
- For PhD students
  - Supervised thesis research

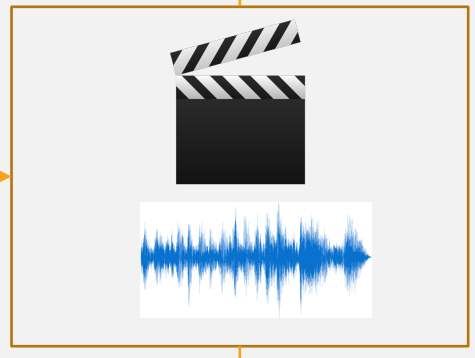
MY RESEARCH AREA

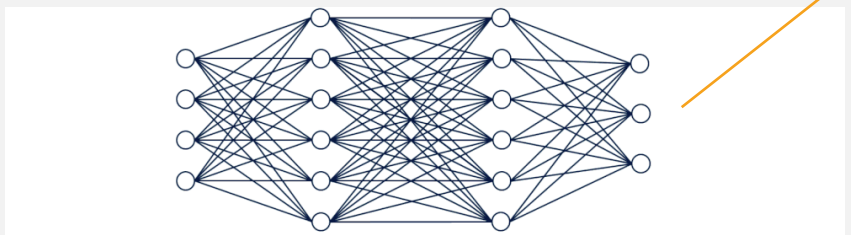
Data capture via distributed digital devices

Feature extraction via automated algorithms and crowdsourcing

Digital diagnostics and adaptive therapies





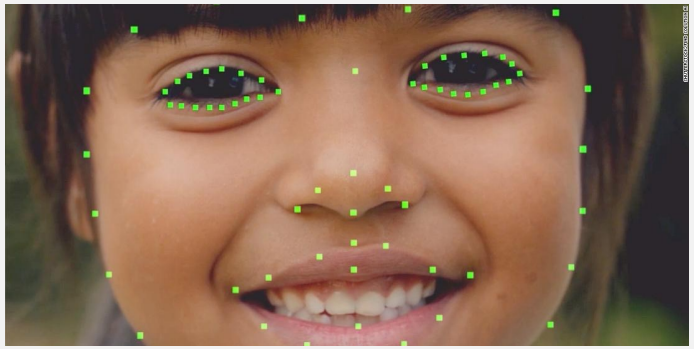


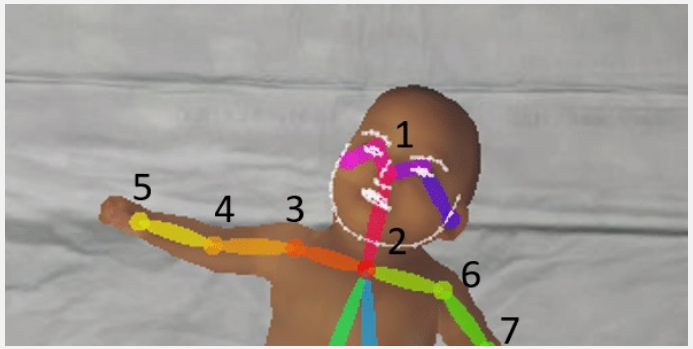
2 2 0 0 1 4

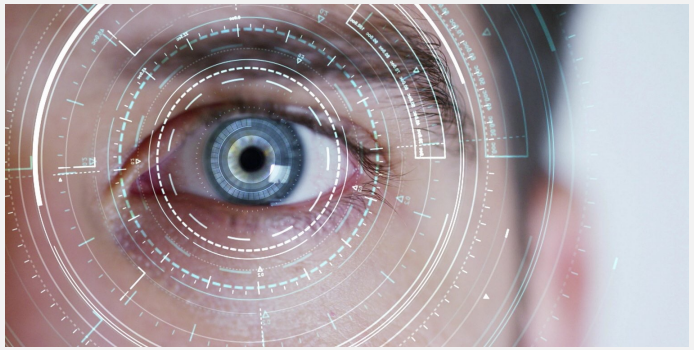
…

-1 1 3 6 0 0

CREATE MACHINE LEARNING MODELS FOR CLASSIFYING HUMAN BEHAVIOR

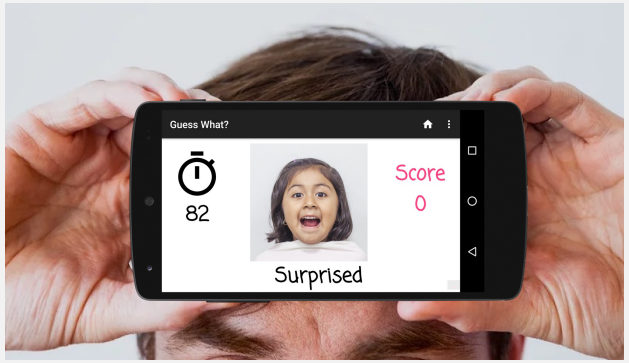






CREATE FUN WEB OR MOBILE GAMES WHICH COLLECT STRUCTURED USER DATA

Parent Guesses → Child Acts → Data Logged

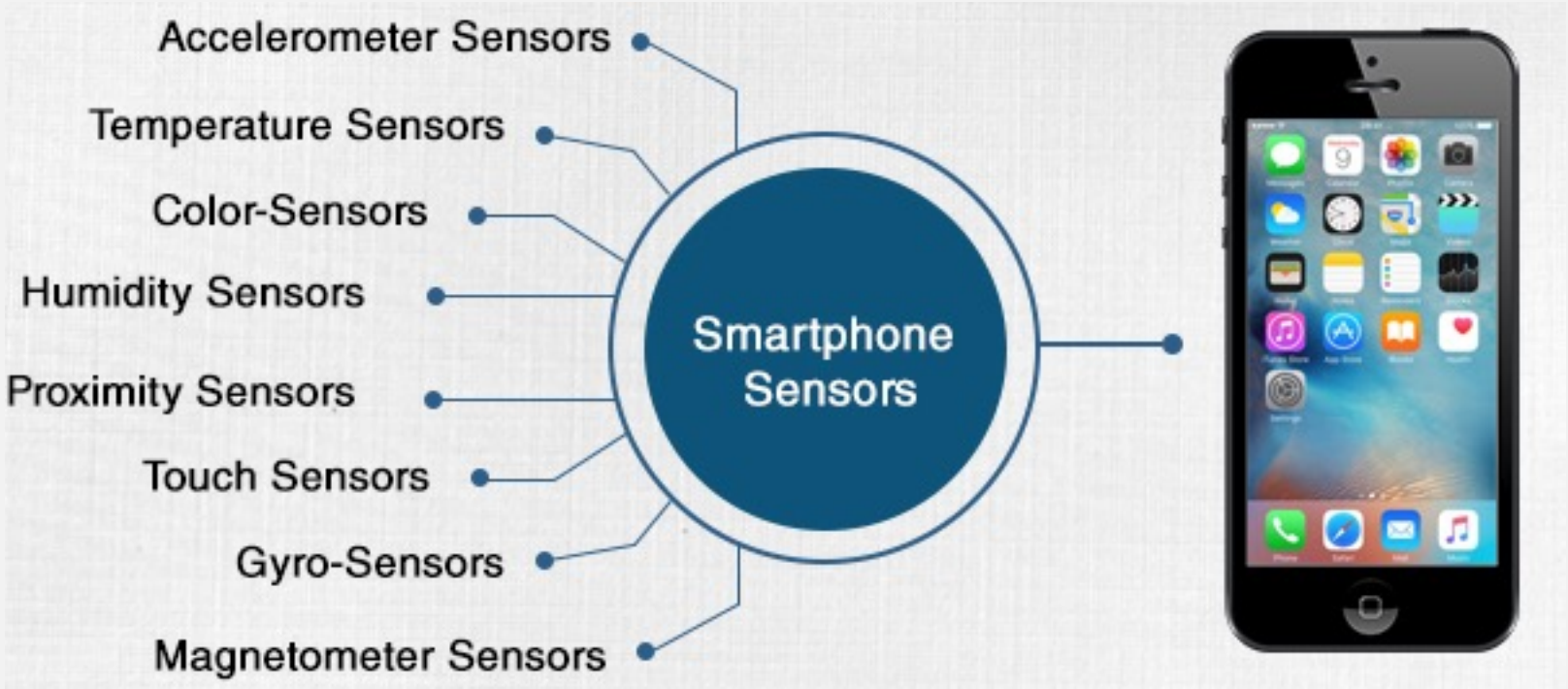


COMPUTER VISION LIBRARIES

GAME PLAY DATA

CREATE NEW DATASETS THROUGH PYTHON DATA MINING

DEVELOP BACKEND SMARTPHONE APPLICATIONS FOR PSYCHIATRIC HEALTHCARE



https://myphonefactor.in/2012/04/sensors-used-in-a-smartphone/

OUTLINE

- Literature Review Topics
- Classic Supervised Learning
  - Regression
  - Classification
- Classic Unsupervised Learning
- Implementing ML in Python
- Discussion

UPCOMING ASSIGNMENTS DUE

- Monday August 29: Literature review topic AND discussion section topic
- Wednesday August 31: Proposal topic
- Wednesday September 7: Haber reflection

OUTLINE

- Literature Review Topics
- Classic Supervised Learning
  - Regression
  - Classification
- Classic Unsupervised Learning
- Implementing ML in Python
- Discussion

LITERATURE REVIEW TOPIC SUBMISSION (DUE MON AUG 29)

- Choose a topic from the course calendar that interests you based on the overview from last class
- Provide an additional sentence summarizing the lens through which you will analyze the topic
- Submit on Laulima
- Example submissions:
  1. *Topic: Crowdsourcing. Lens: I will review the use (or lack thereof) of fair payment practices for data labeling on crowdsourcing platforms such as Amazon Mechanical Turk.*
  2. *Topic: Interpretable ML. Lens: Lens: I will review the state-of-the-art in methods for interpretability of multimodal prediction models.*

LITERATURE REVIEW TOPIC SUBMISSION (DUE MON AUG 29)

- Choose a topic from the course calendar that interests you based on the overview from last class
- Provide an additional sentence summarizing the lens through which you will analyze the topic
- Submit on Laulima
- Example submissions:
  1. *Topic: Crowdsourcing. Lens: I will review the use (or lack thereof) of fair payment practices for data labeling on crowdsourcing platforms such as Amazon Mechanical Turk.*
  2. *Topic: Interpretable ML. Lens: Lens: I will review the state-of-the-art in methods for interpretability of multimodal prediction models.*

Questions about your planned topic? Class time to discuss.

DISCUSSION TOPIC SUBMISSION (DUE MON AUG 29)

- Provide your ranking of the top-10 topics you'd like to lead
- Alternatively, you may say "I am fine with leading any discussion topic"
- Submit on Laulima

PROPOSAL TOPIC SUBMISSION (DUE WED AUG 31)

- Choose a topic from the course calendar that interests you based on the overview from last class
- Provide a paragraph-long summary describing your project proposal
- Submit on Laulima
- Requested format:
  - *Project Category: [choose from course calendar]*
  - *Societal and/or Technical Motivation: [text]*
  - *Dataset: [include size, data description, and whether it is a publicly available dataset or how the data will be collected]*
  - *Methods: [a few sentences about the proposed methodology]*

PROPOSAL TOPIC SUBMISSION (DUE WED AUG 31)

- Choose a topic from the course calendar that interests you based on the overview from last class
- Provide a paragraph-long summary describing your project proposal
- Submit on Laulima
- Requested format:
  - *Project Category: [choose from course calendar]*
  - *Societal and/or Technical Motivation: [text]*
  - *Dataset: [include size, data description, and whether it is a publicly available dataset or how the data will be collected]*
  - *Methods: [a few sentences about the proposed methodology]*

Questions about your planned topic? Class time to discuss.

TODAY'S CLASS

- Today will probably be the most math-heavy lecture and discussion
- Do not worry about understanding every detail
  - You will not be tested or expect to recall these details
  - Rather, this is to make sure we are all on the same page for the technical aspects of the course
- If you are new to ML, this lecture should still be accessible
  - In this case, focus on understanding the big picture and some of the details

OUTLINE

- Literature Review Topics
- Classic Supervised Learning
  - Regression
  - Classification
- Classic Unsupervised Learning
- Implementing ML in Python
- Discussion

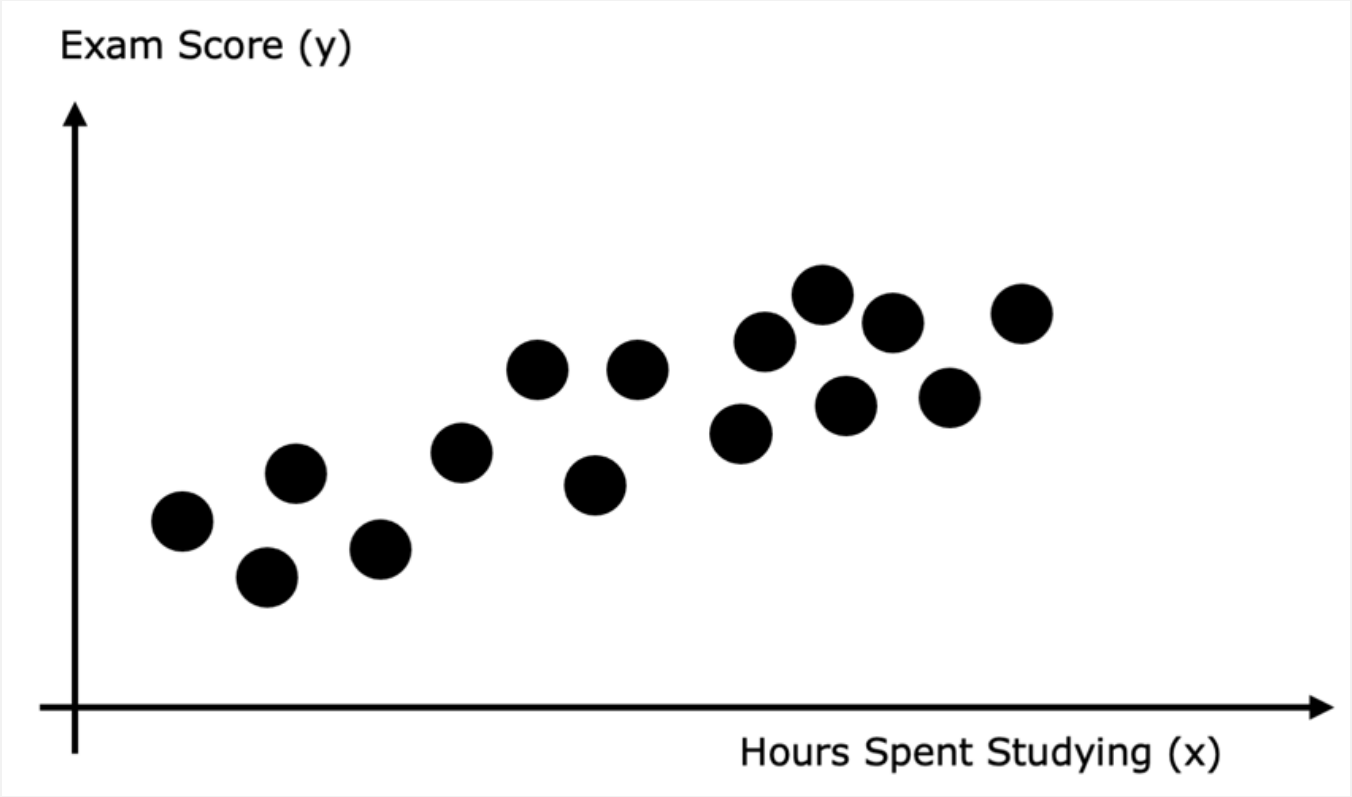
REGRESSION

- Output: continuous number
- Examples?

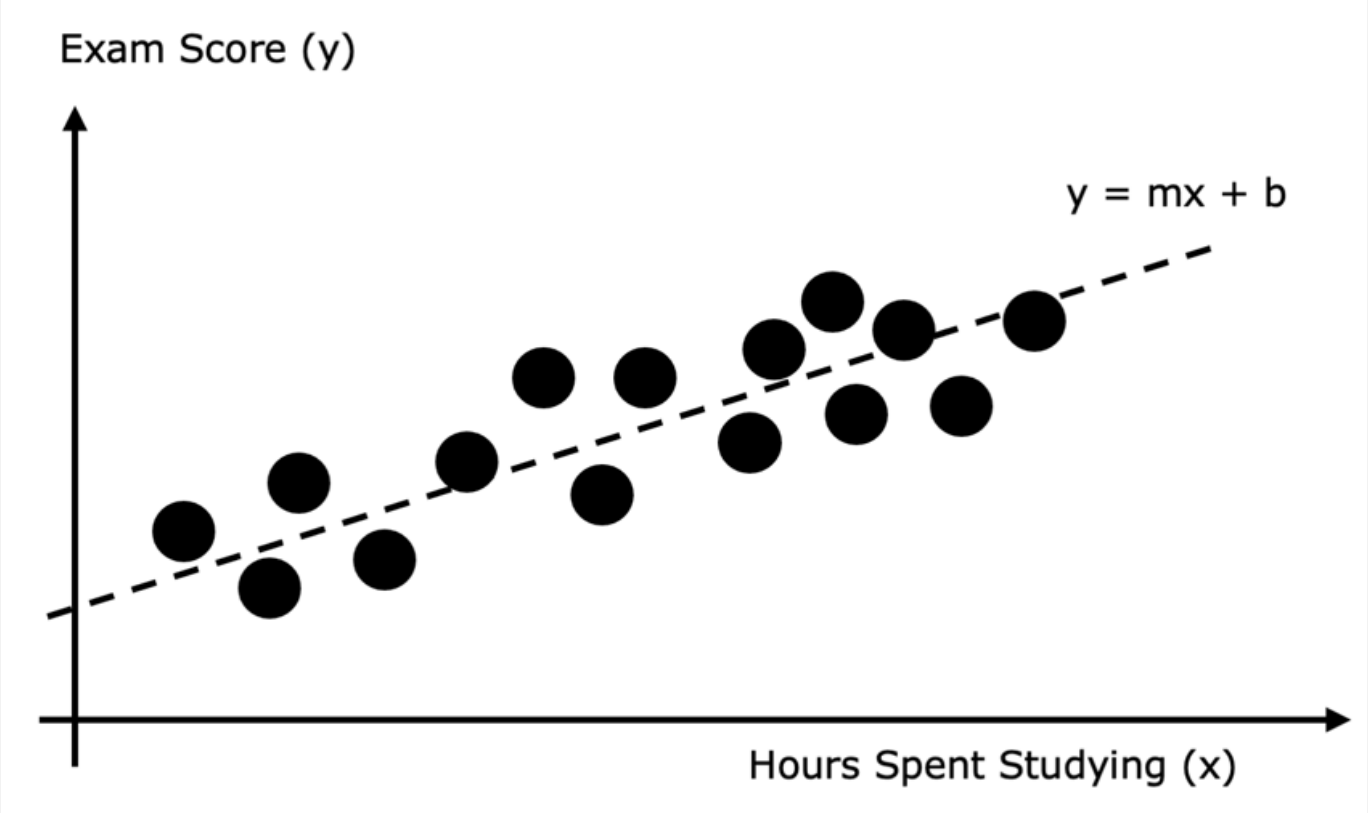
REGRESSION

- Output: continuous number
- Examples?
  - Predict age given selfie
  - Predict house price given properties about the house
  - Predict blood glucose level given DNA and diet log
  - …

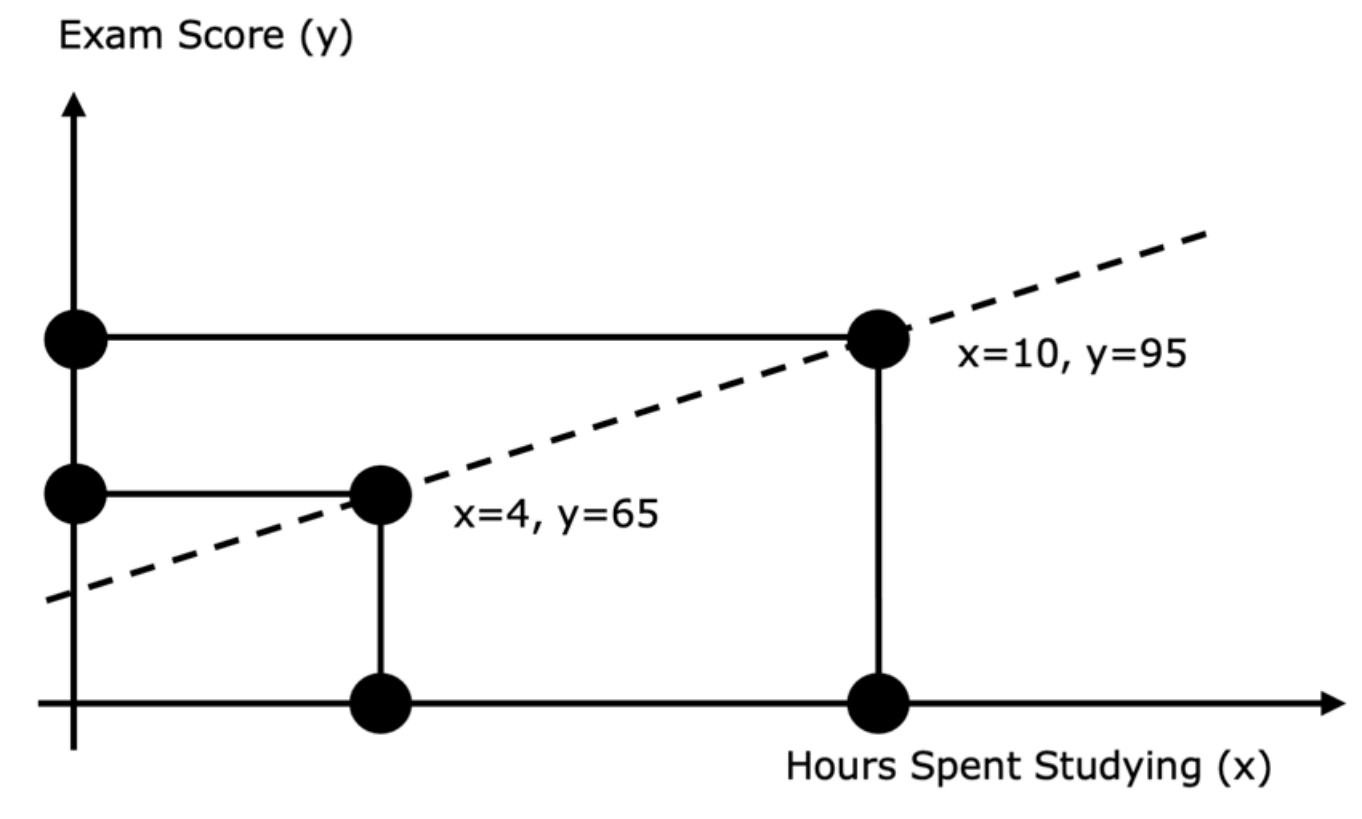
REGRESSION



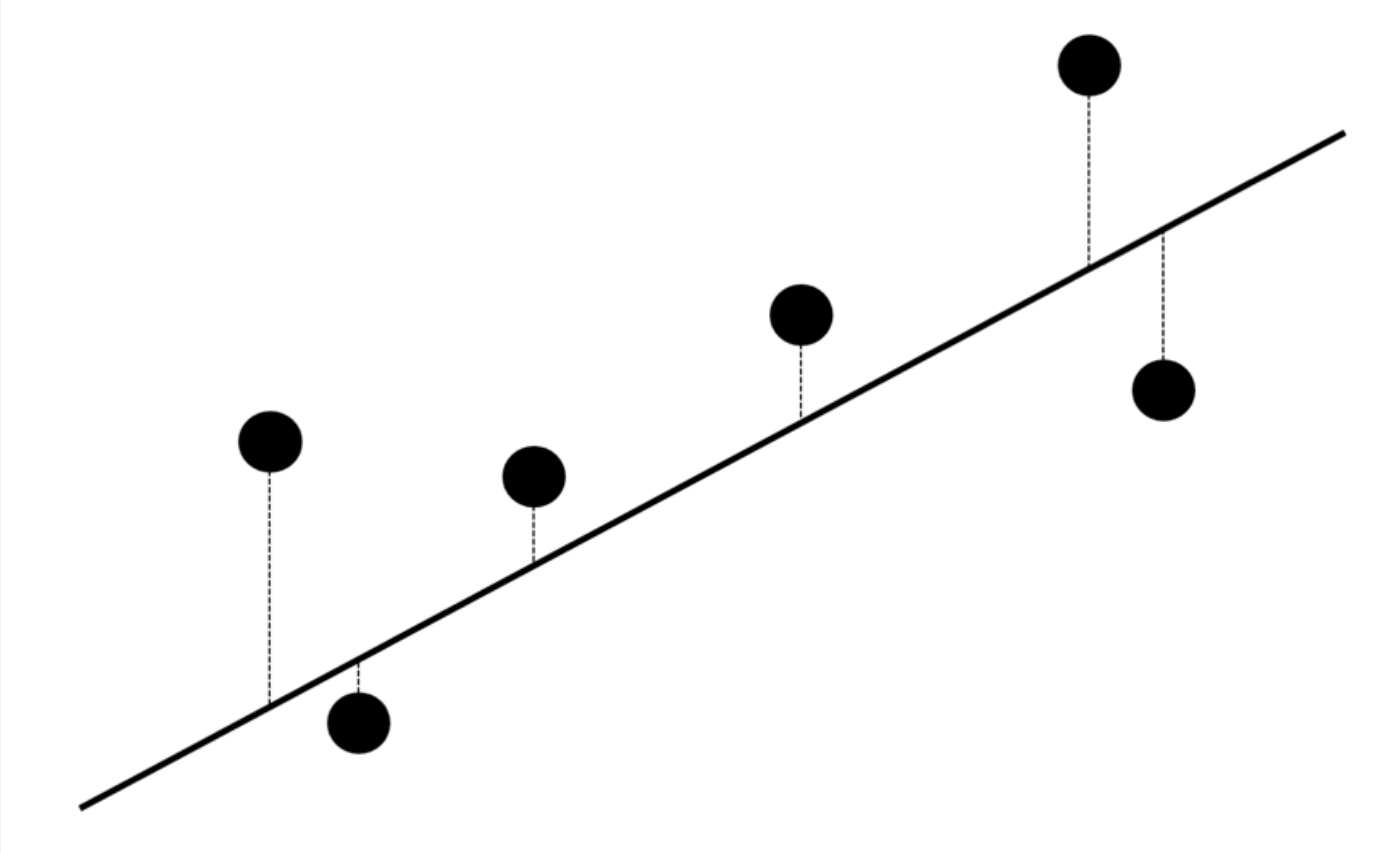
REGRESSION



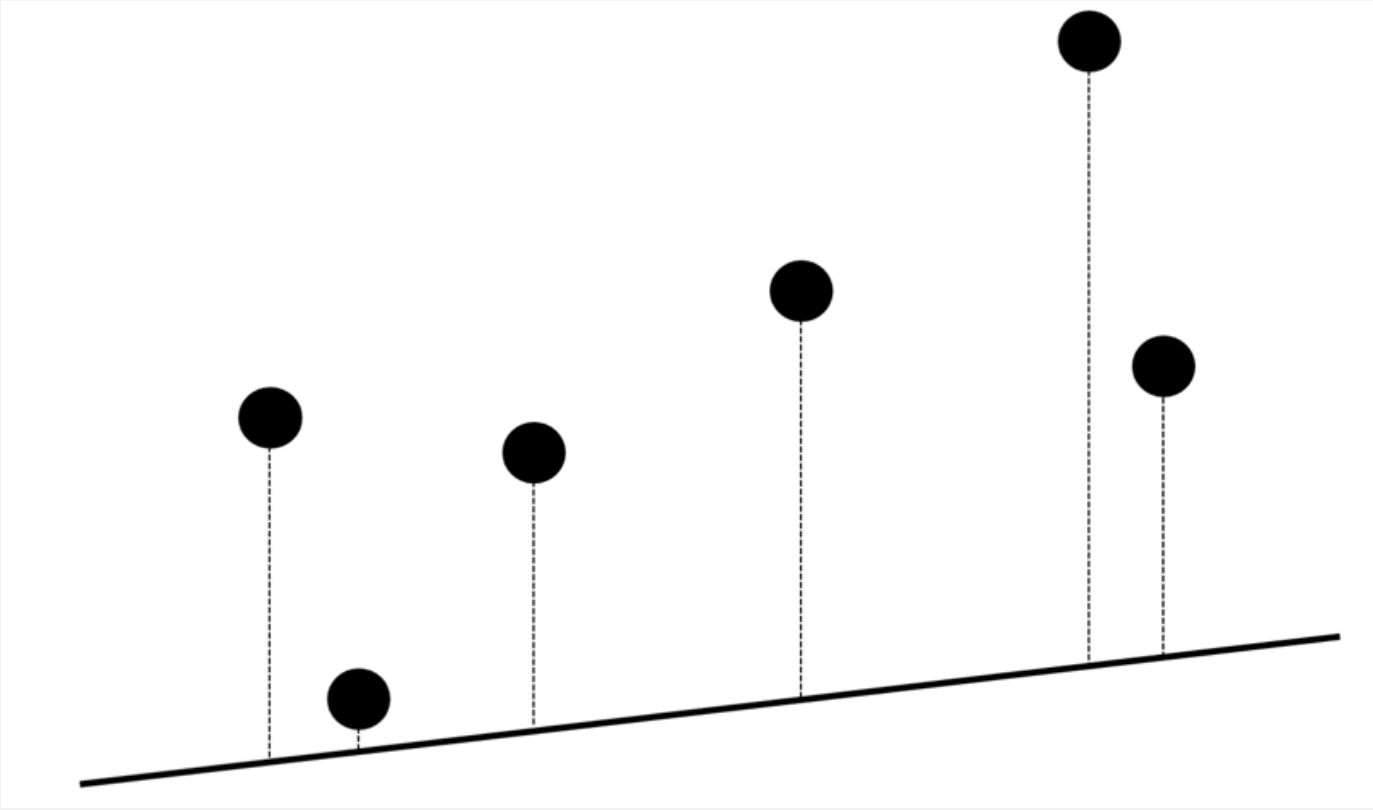
REGRESSION



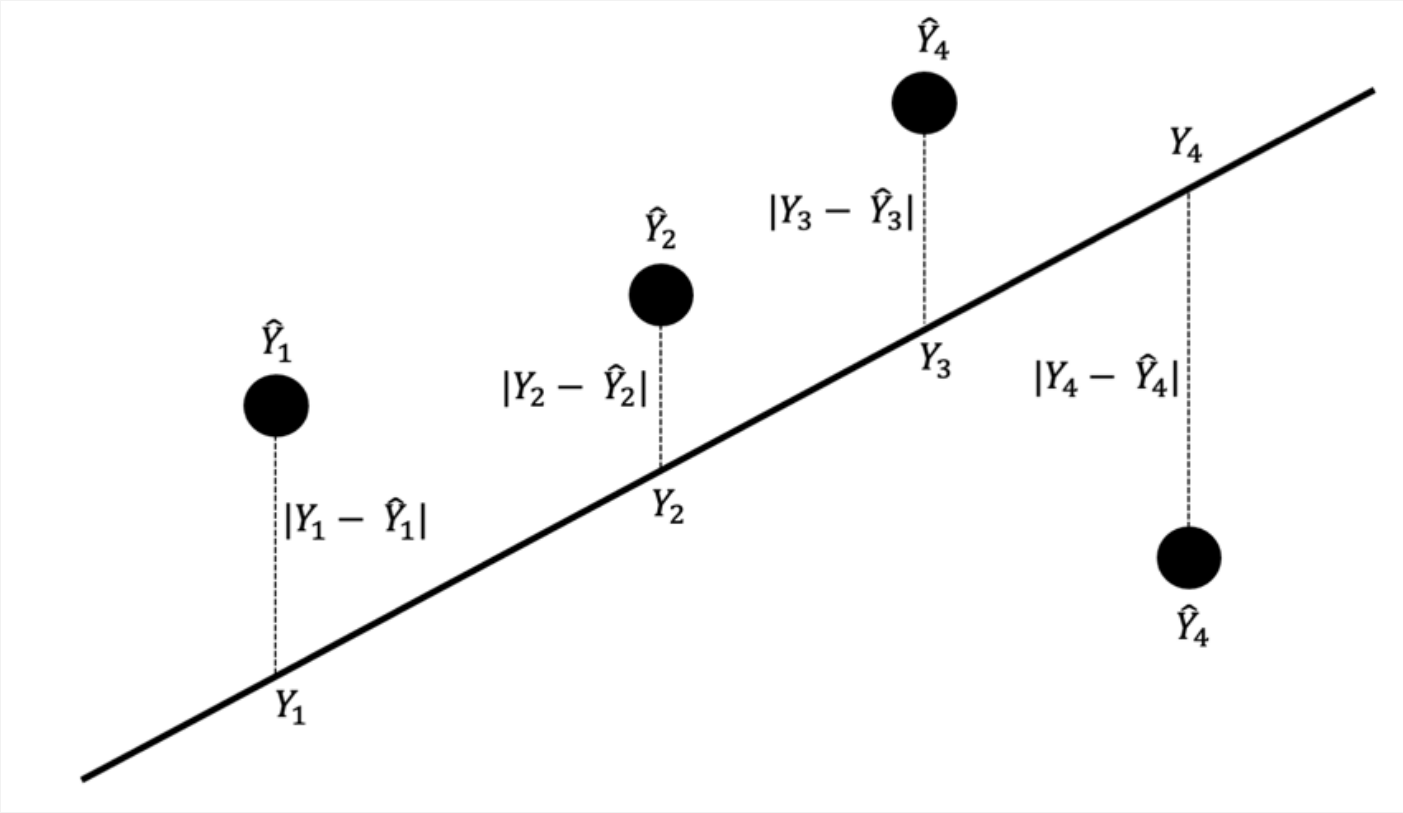
WHAT IS A LINE OF BEST FIT?



GRADIENT DESCENT



GRADIENT DESCENT



LOSS FUNCTION FOR REGRESSION

$$\frac{\sum_{i=1}^{N} |Y_i - \hat{Y}_i|}{N}$$

GOAL OF LINEAR REGRESSION

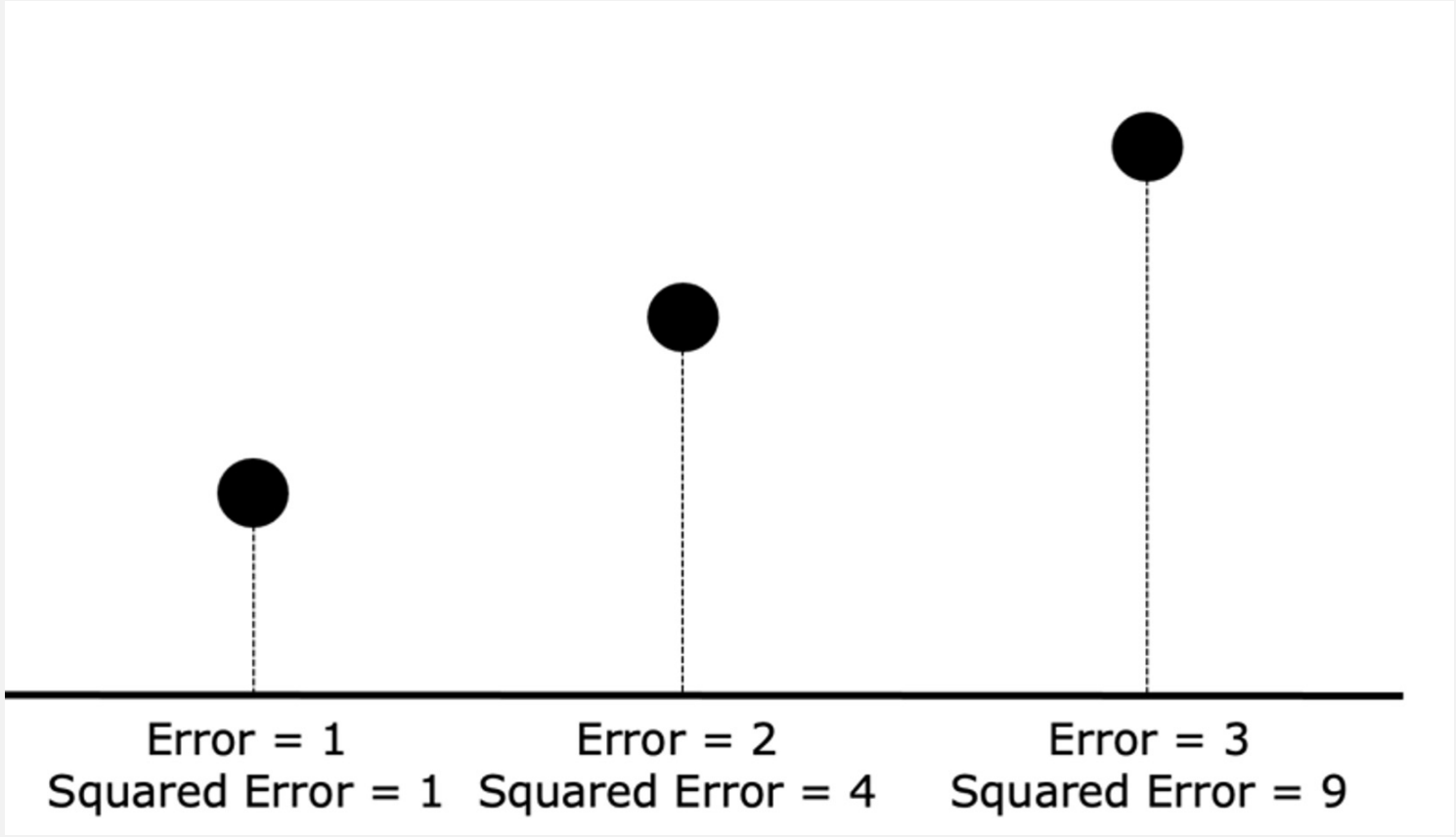
$$min_{m,b} \frac{\sum_{i=1}^{N} |Y_i - (mx_i + b)|}{N}$$

In other words, the goal is to minimize the loss

This is the goal in all supervised machine learning

LOSS FUNCTION FOR REGRESSION

$$\frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{N}$$

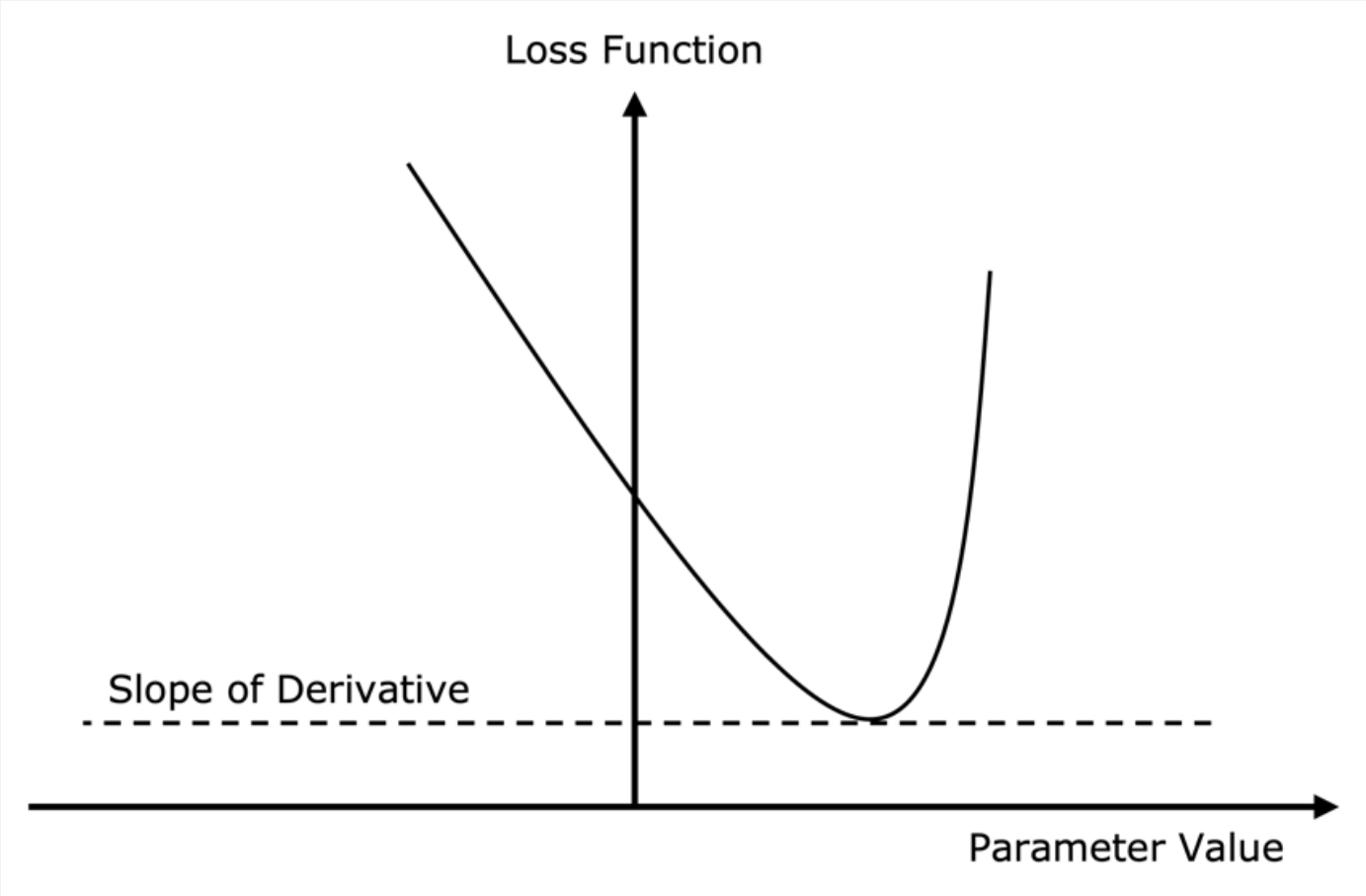


GRADIENT DESCENT MATHEMATICALLY

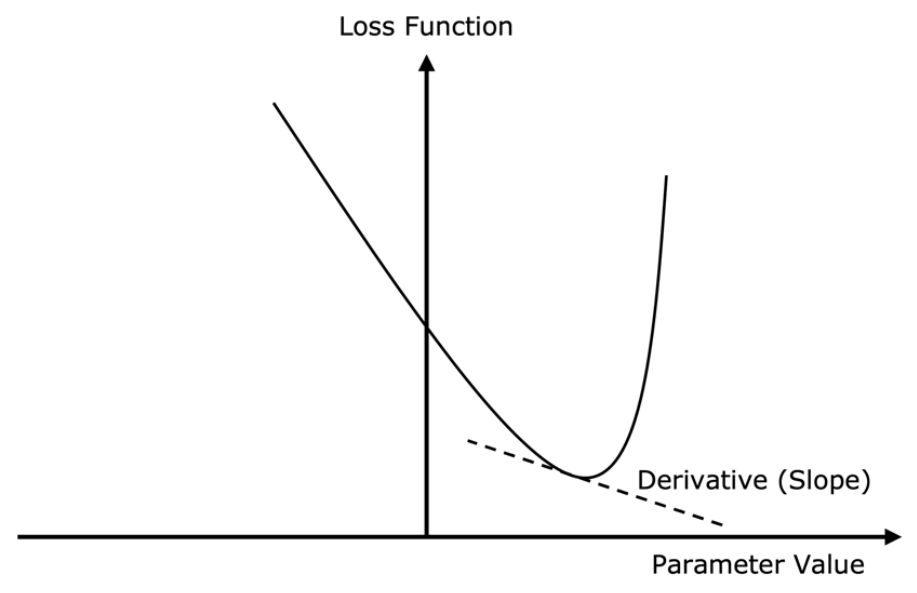
$$m = m - (learning\ rate) \times \frac{\partial}{\partial m}(Loss\ Function)$$

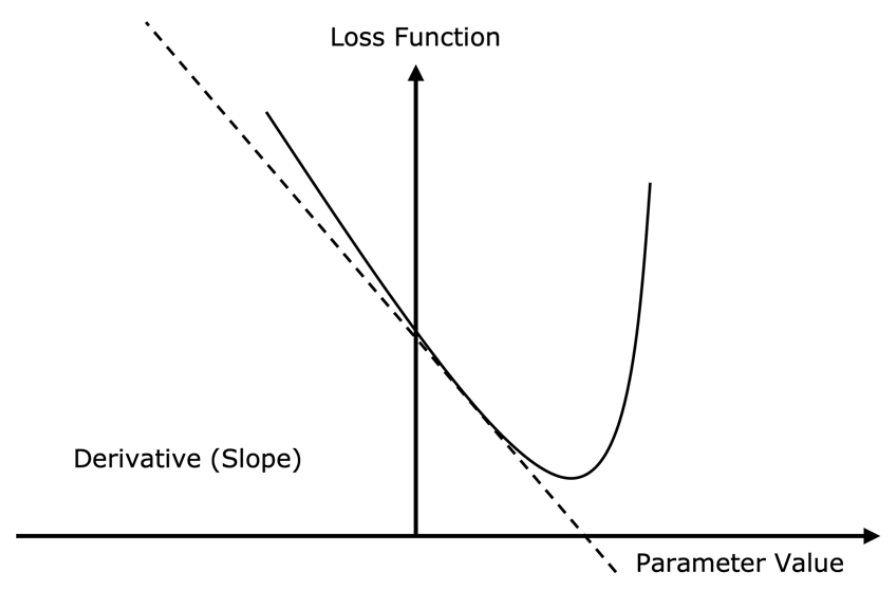
$$b = b - (learning\ rate) \times \frac{\partial}{\partial b}(Loss\ Function)$$

INTUITION BEHIND TAKING THE DERIVATIVE OF THE LOSS FUNCTION



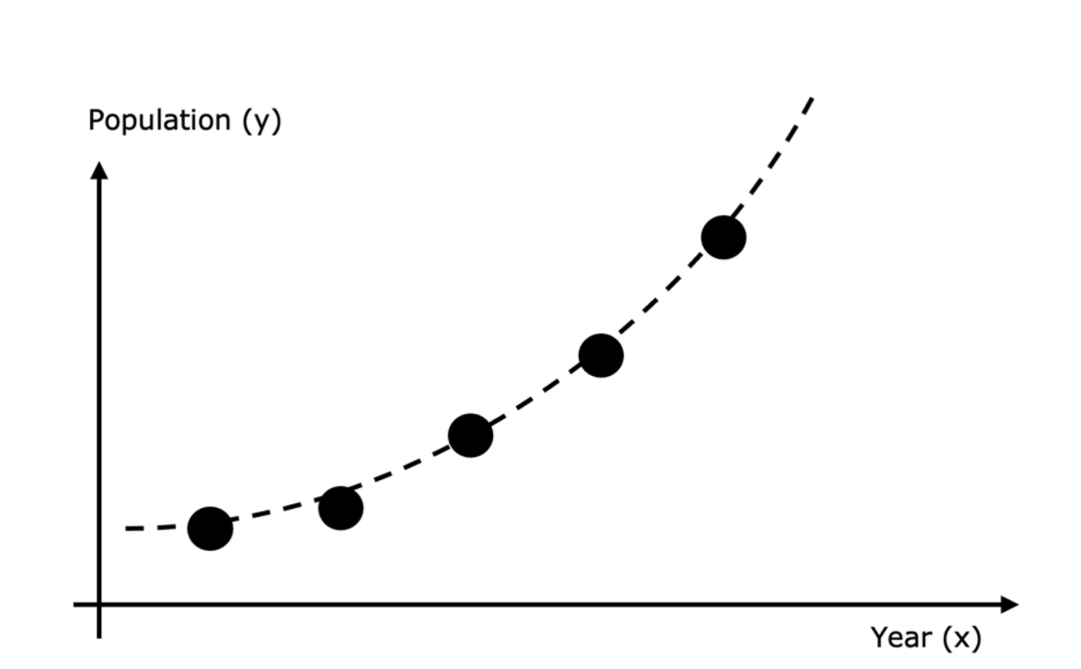
INTUITION BEHIND TAKING THE DERIVATIVE OF THE LOSS FUNCTION





NON-LINEAR REGRESSION

$$y = a\,x^{b} + c$$



NON-LINEAR REGRESSION

$$a = a - (learning\ rate) \times \frac{\partial}{\partial a}(Loss\ Function)$$

$$b = b - (learning\ rate) \times \frac{\partial}{\partial b}(Loss\ Function)$$

$$c = c - (learning\ rate) \times \frac{\partial}{\partial c}(Loss\ Function)$$

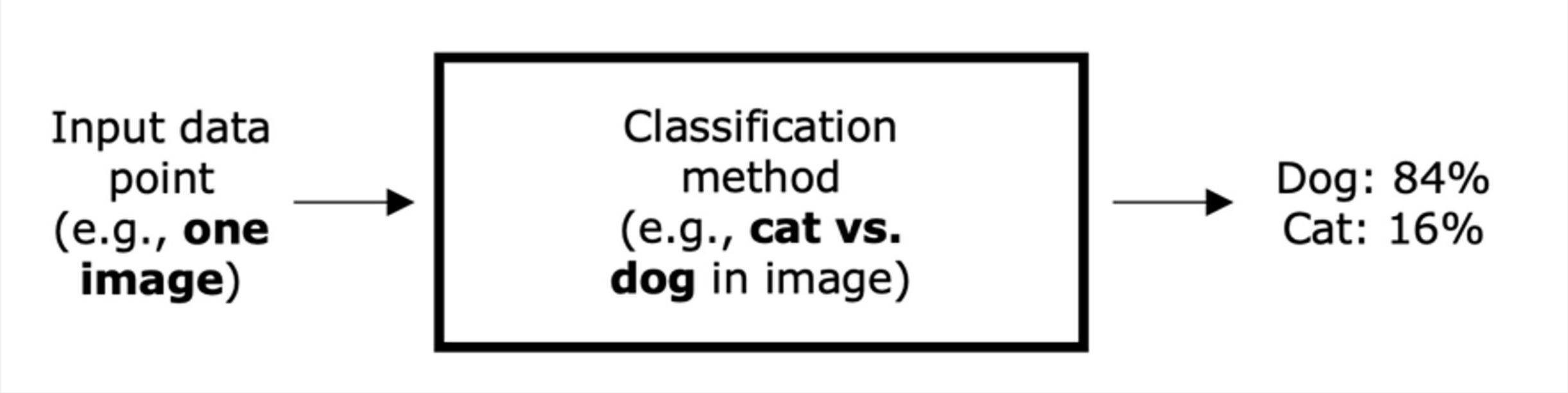
MULTIPLE REGRESSION

$$y = m_1 x_1 + m_2 x_2 + \cdots + m_N x_N + b$$

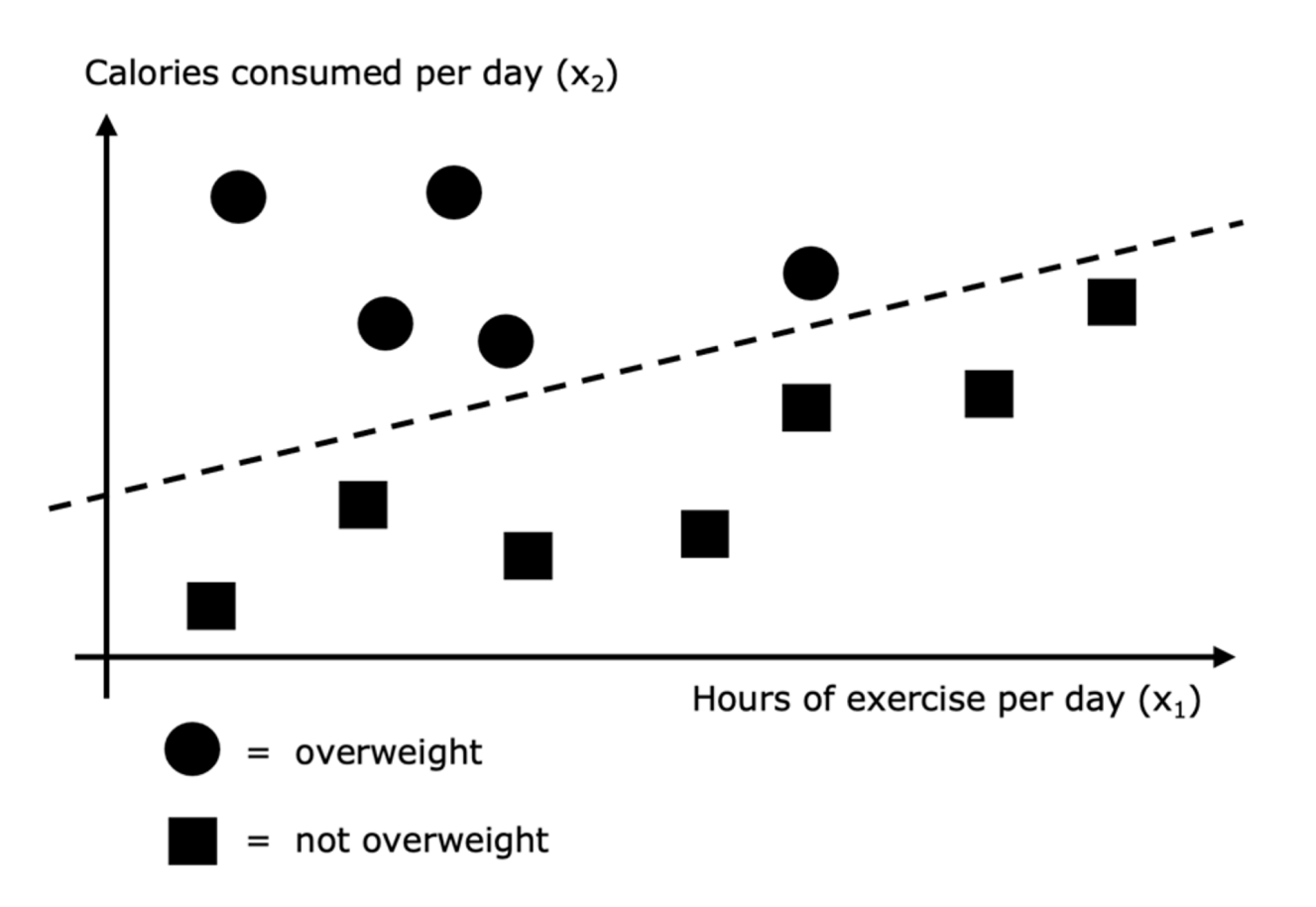
OUTLINE

- Literature Review Topics
- Classic Supervised Learning
  - Regression
  - Classification
- Classic Unsupervised Learning
- Implementing ML in Python
- Discussion

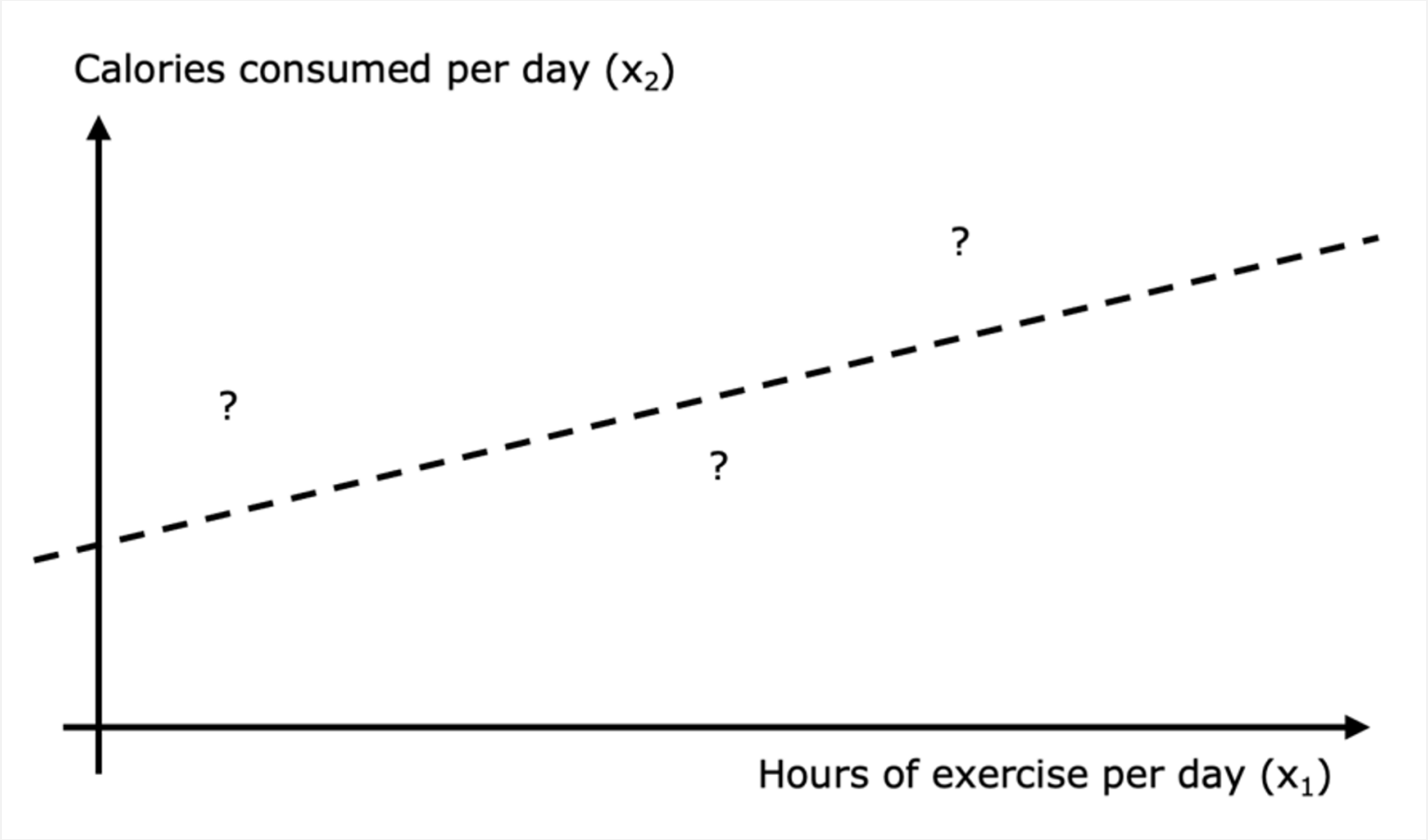
CLASSIFICATION



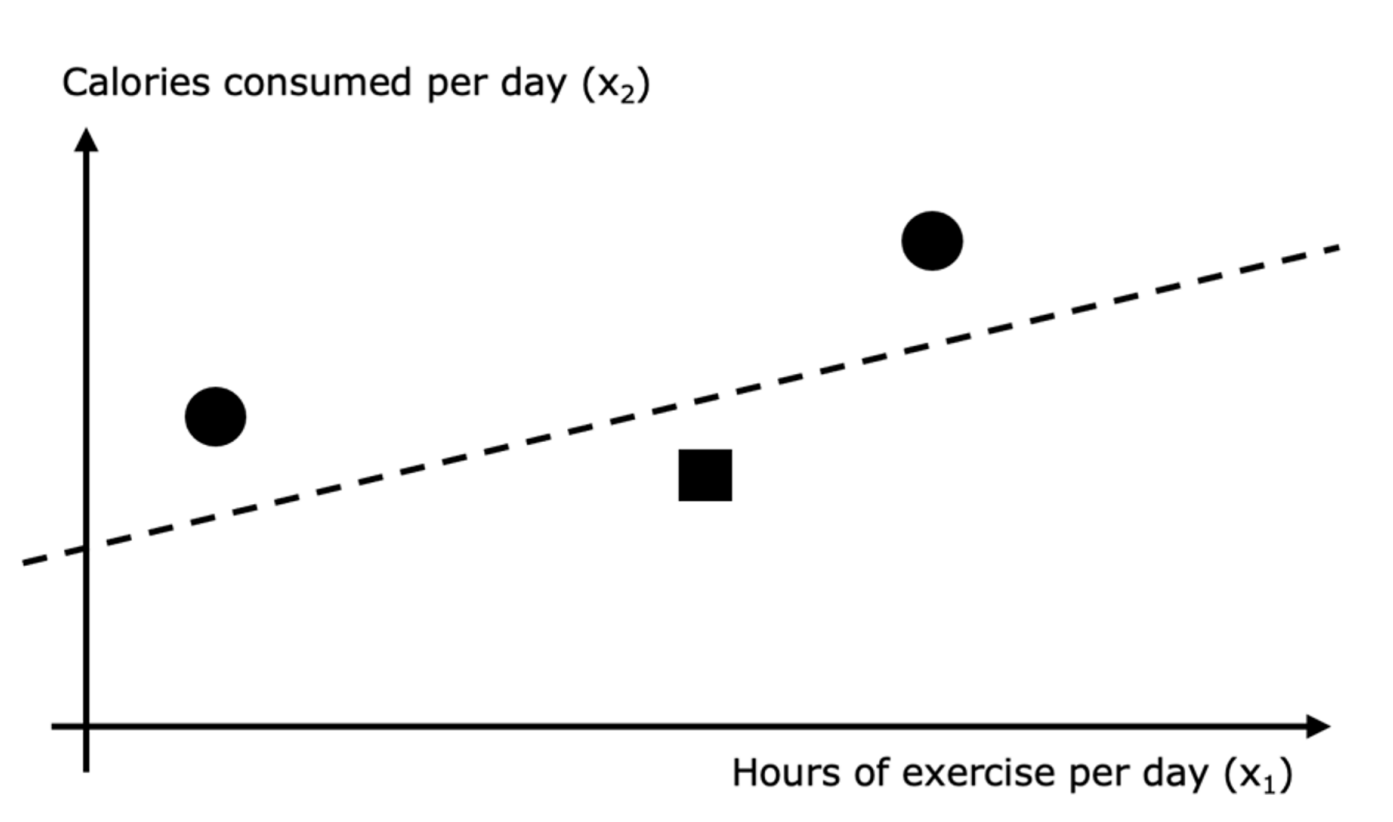
CLASSIFICATION



CLASSIFICATION

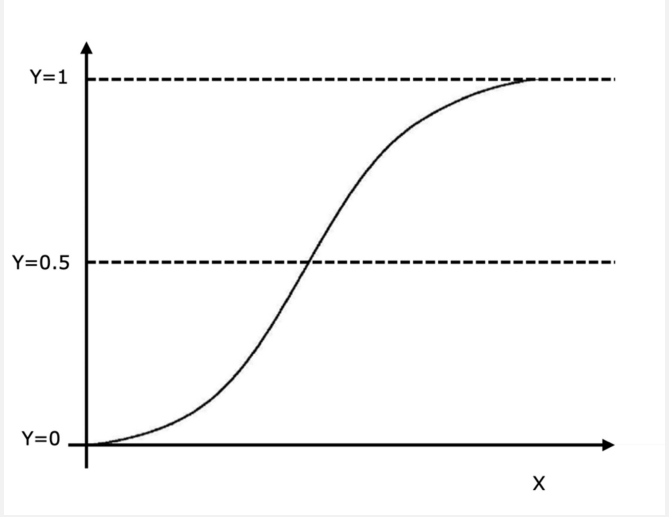


CLASSIFICATION



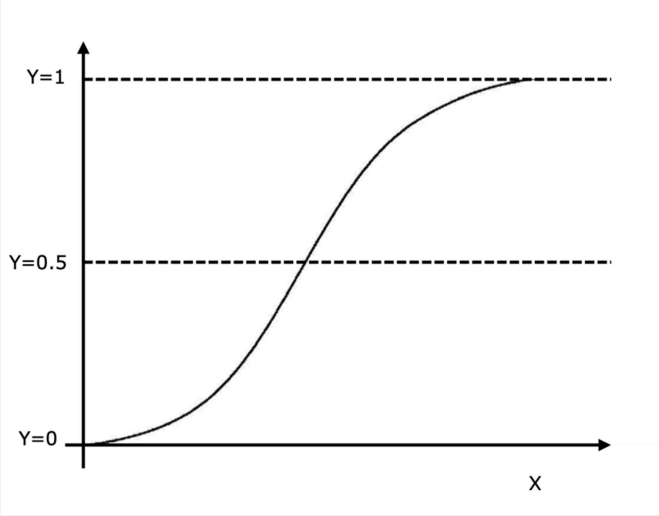
LOGISTIC REGRESSION

$$f(x) = \frac{1}{1 + e^{-x}}$$



LOGISTIC REGRESSION

$$Sigmoid\ Activation(mx + b) = f(mx + b) = \frac{1}{1 + e^{-mx+b}}$$



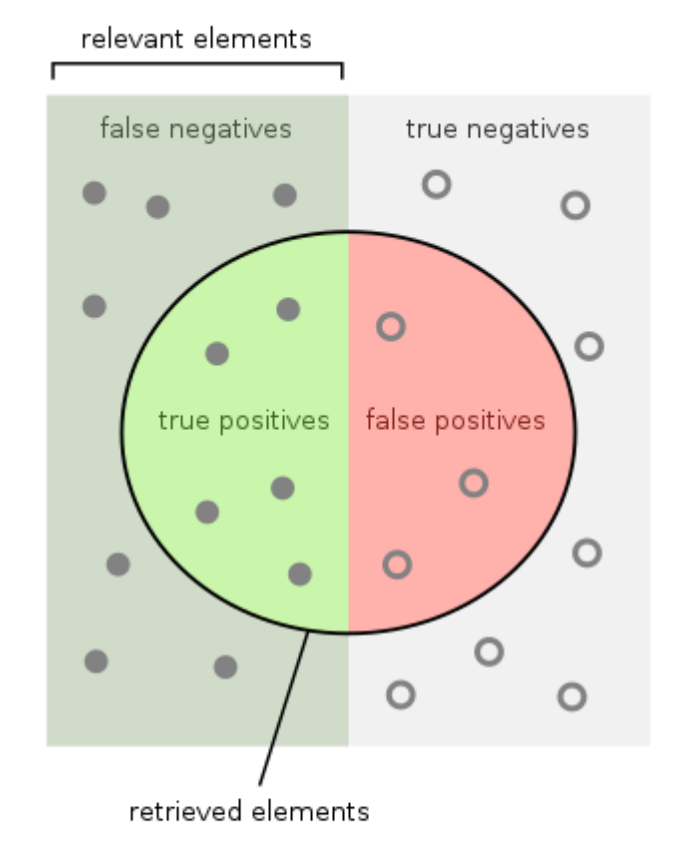
LOSS FUNCTION FOR CLASSIFICATION

$$\begin{cases} -\log(\hat{y}) & y_{true} = 1 \\ -\log(1 - \hat{y}) & y_{true} = 0 \end{cases}$$

AKA

$$\sum_{i=1}^{N} y_{true,i} \log(\hat{y}_i) + (1 - y_{true,i}) \log(1 - \hat{y}_i)$$

EVALUATING CLASSIFICATION



How many retrieved items are relevant?

Precision =

How many relevant items are retrieved?

Recall =

How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.

Sensitivity=

How many negative selected elements are truly negative? e.g. How many healthy people are identified as not having the condition.

Specificity =

Wikipedia

SUMMARY OF STEPS TO THE SUPERVISED LEARNING PROCESS*

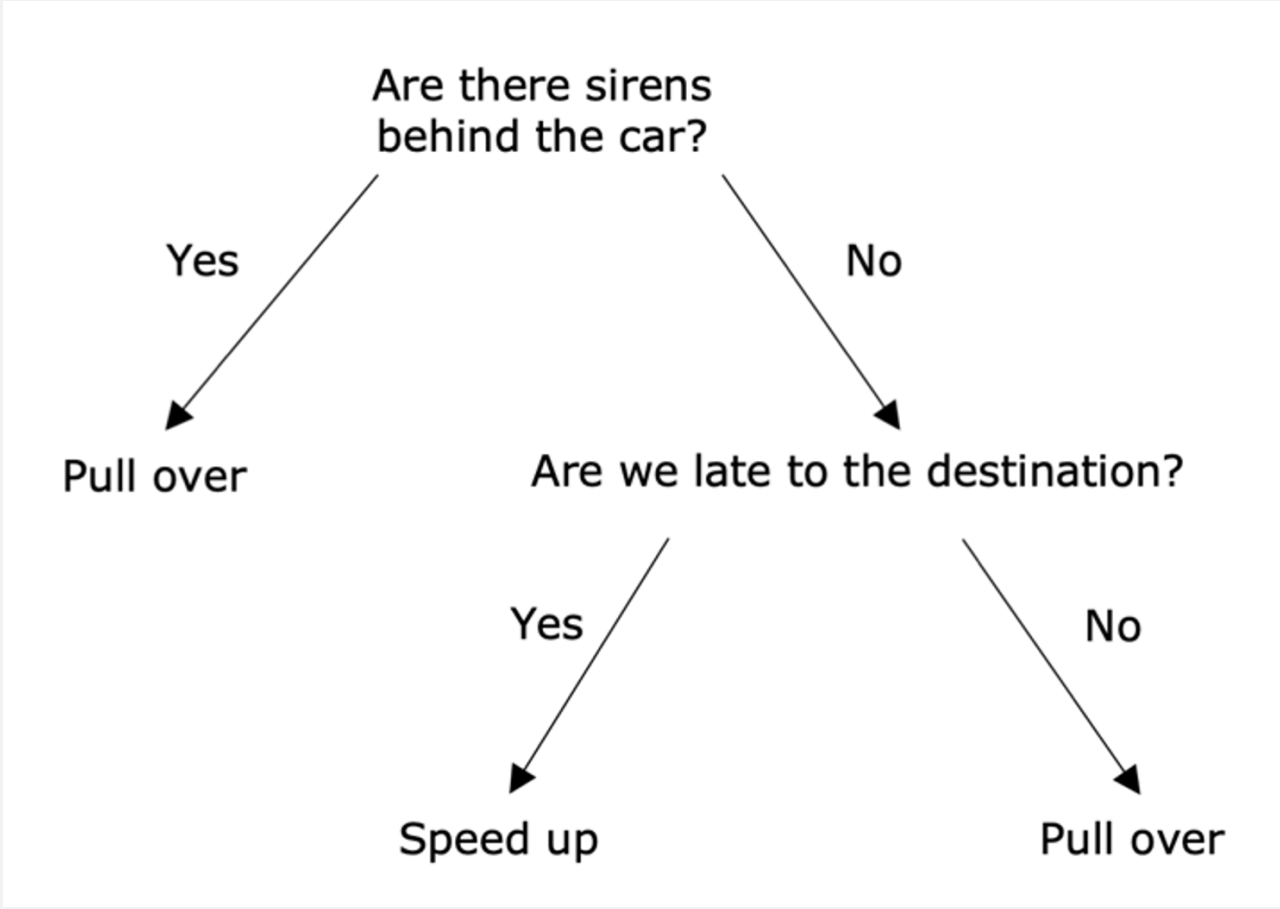
- Start with random parameter values (i.e., the coefficients)
- Define a loss function
- Iteratively:
  - Measure the loss
  - Update each of the model parameters based on the loss
  - Stop when the loss barely changes between iterations
- Evaluate the performance of the model using pre-defined metrics

*There are clever variations to all these steps that are used in more advanced techniques, but this is the basic process

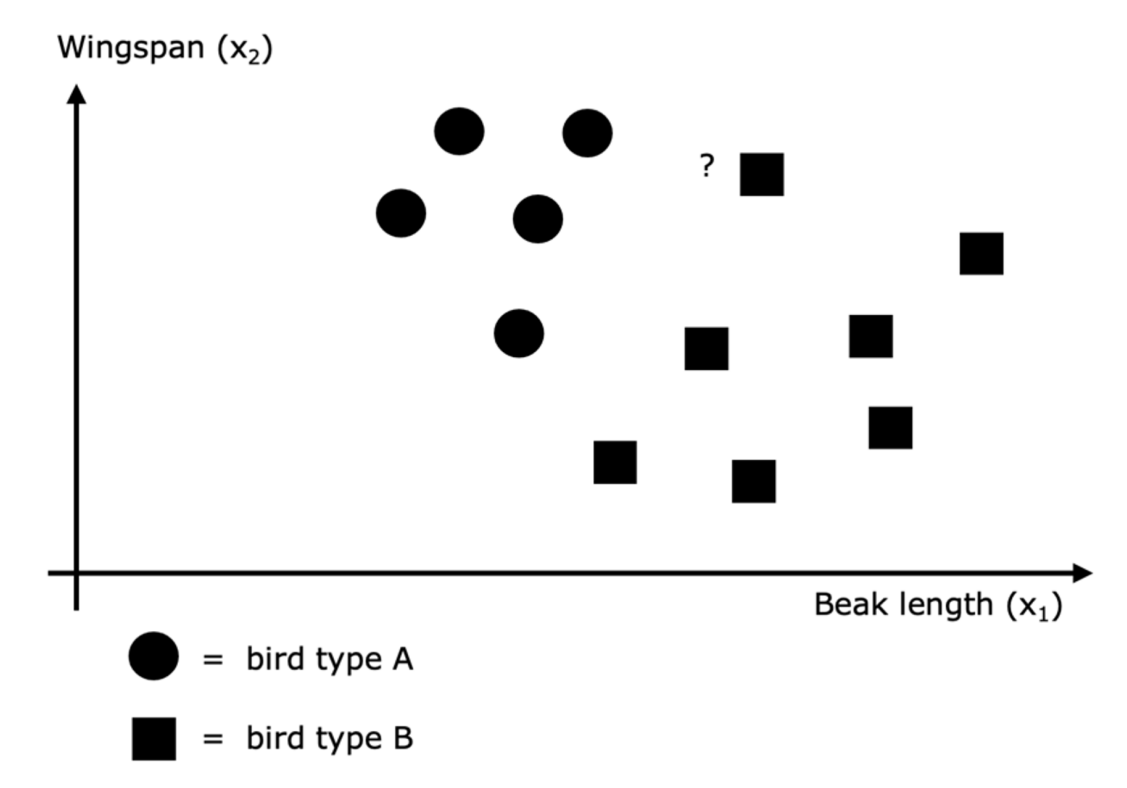
NON-LINEAR PREDICTIONS

- Linear and logistic regression are inherently linear (i.e., coefficients are of the form mx + b),
- Many regression and classification tasks are not linear
- We often do not know the relationship in advance (i.e; canoot assume we can use a polynomial regression)

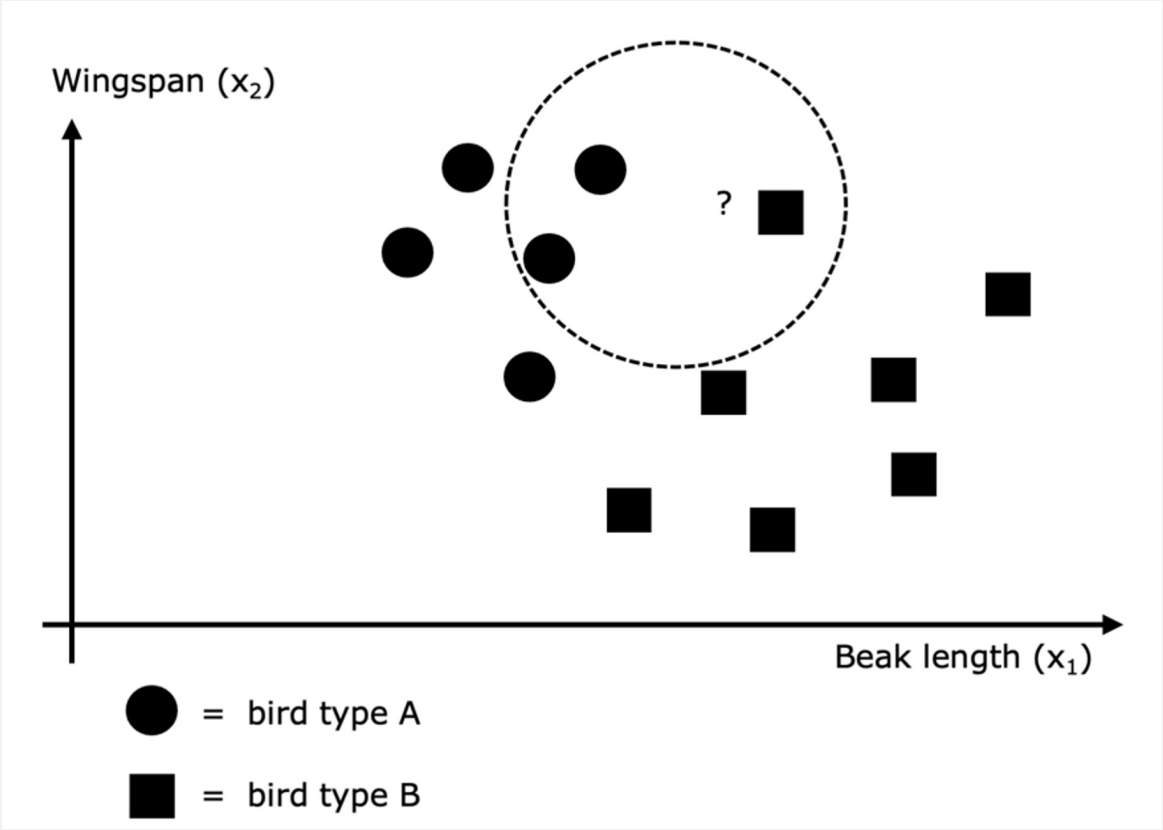
DECISION TREES



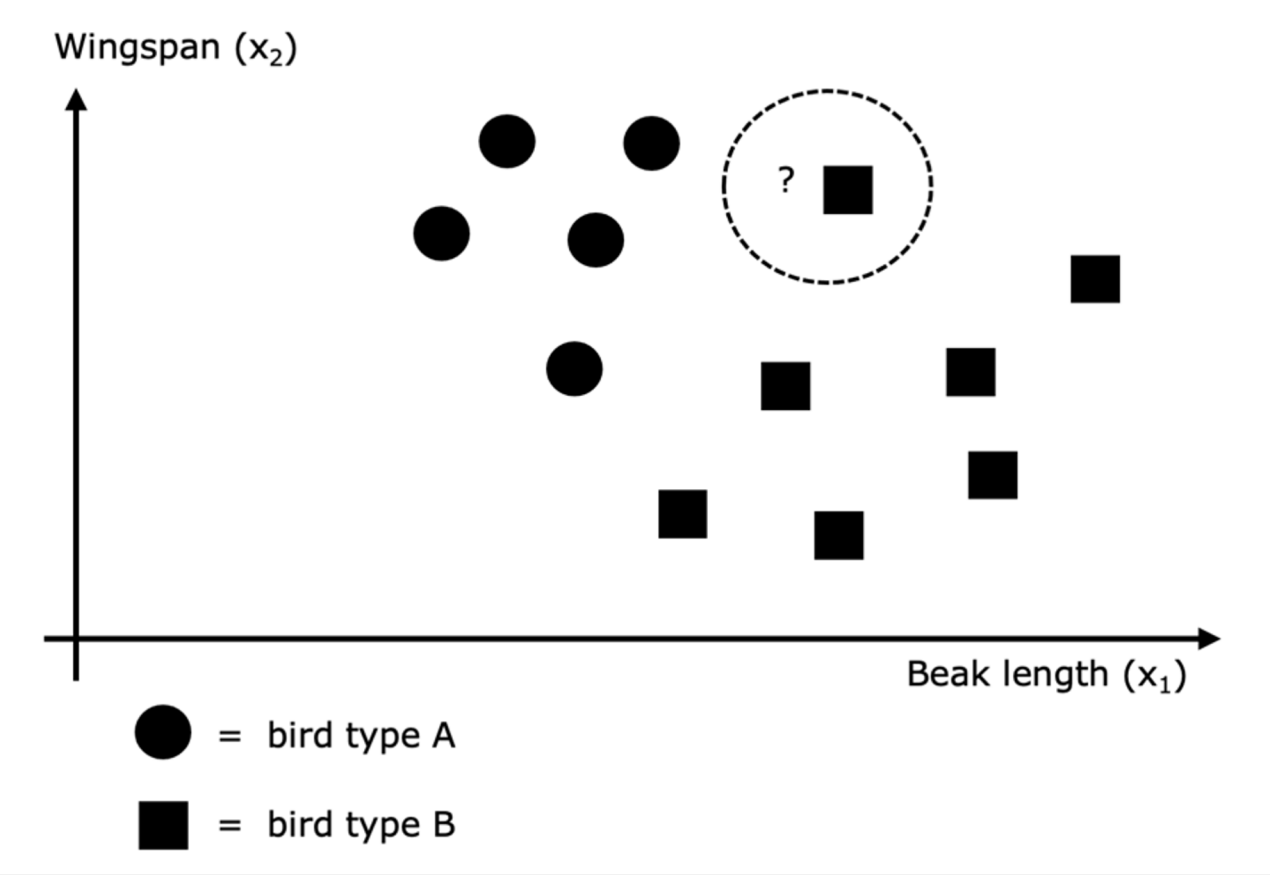
K-NEAREST NEIGHBORS



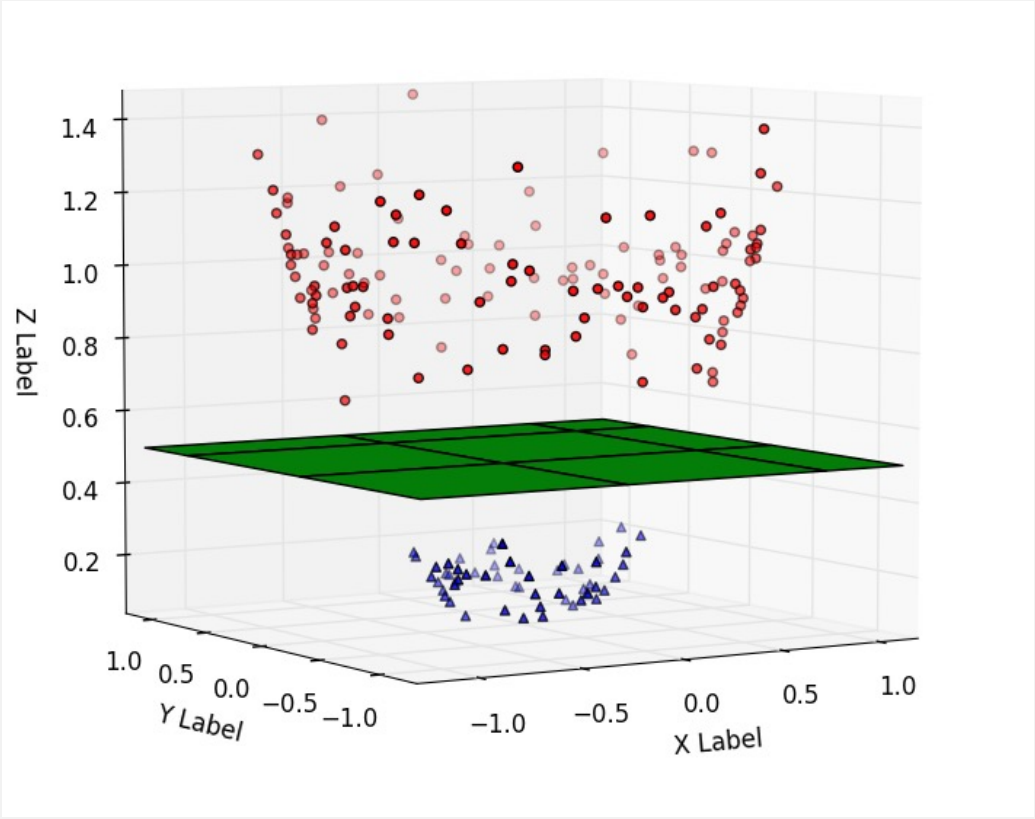
K-NEAREST NEIGHBORS



K-NEAREST NEIGHBORS

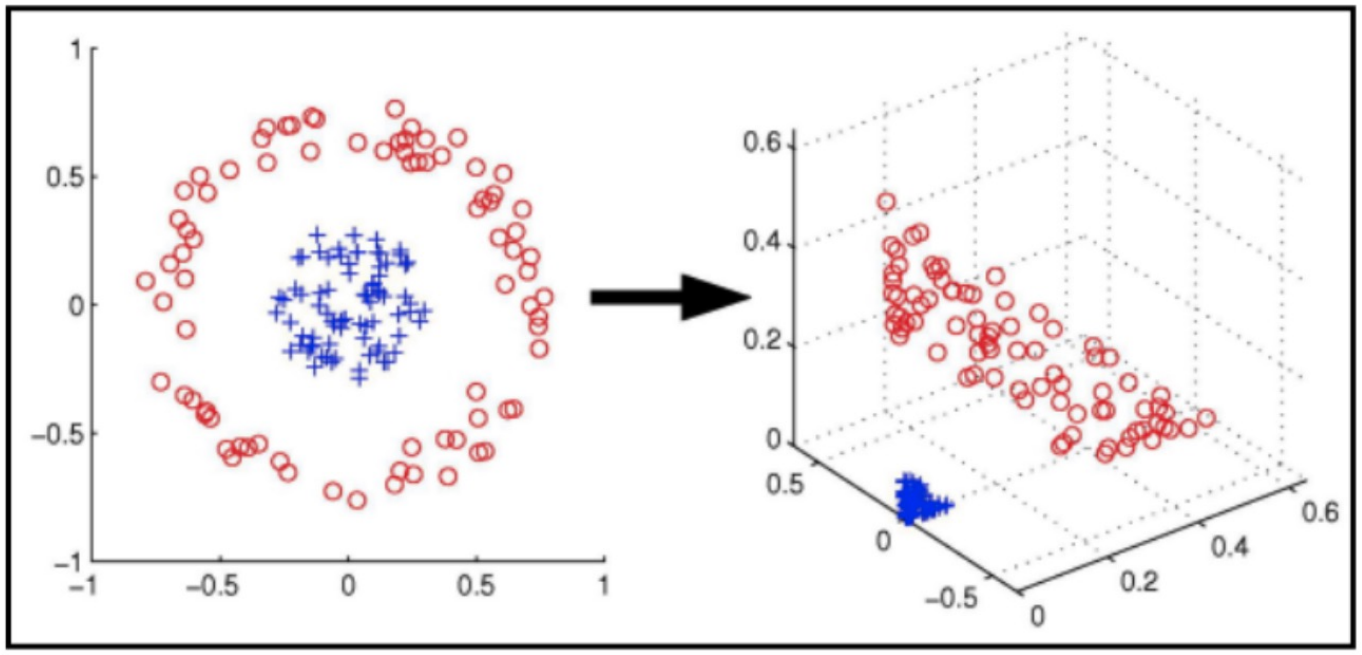


MULTI-DIMENSIONAL MACHINE LEARNING



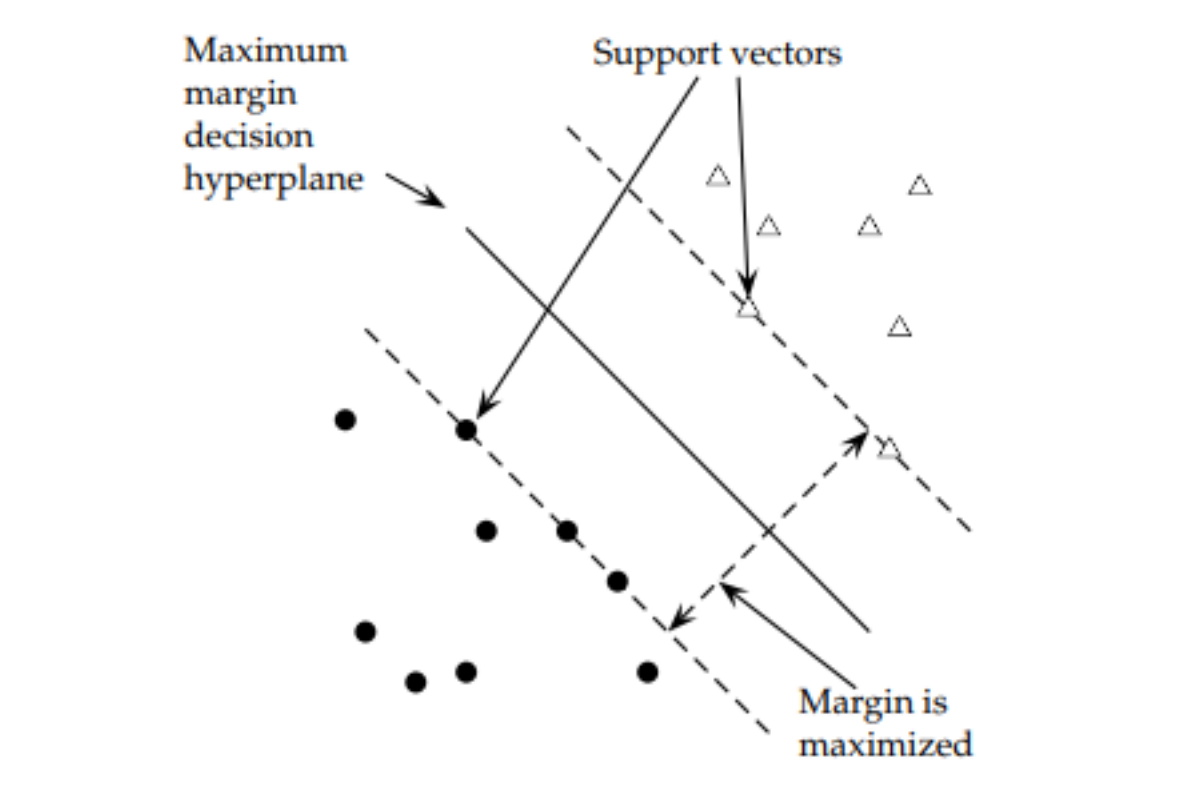
https://www.analyticsvidhya.com/blog/2021/05/support-vector-machines/

SUPPORT VECTOR MACHINES (SVMS)



https://medium.com/@zachary.bedell/support-vector-machines-explained-73f4ec363f13

SUPPORT VECTOR MACHINES (SVMS)

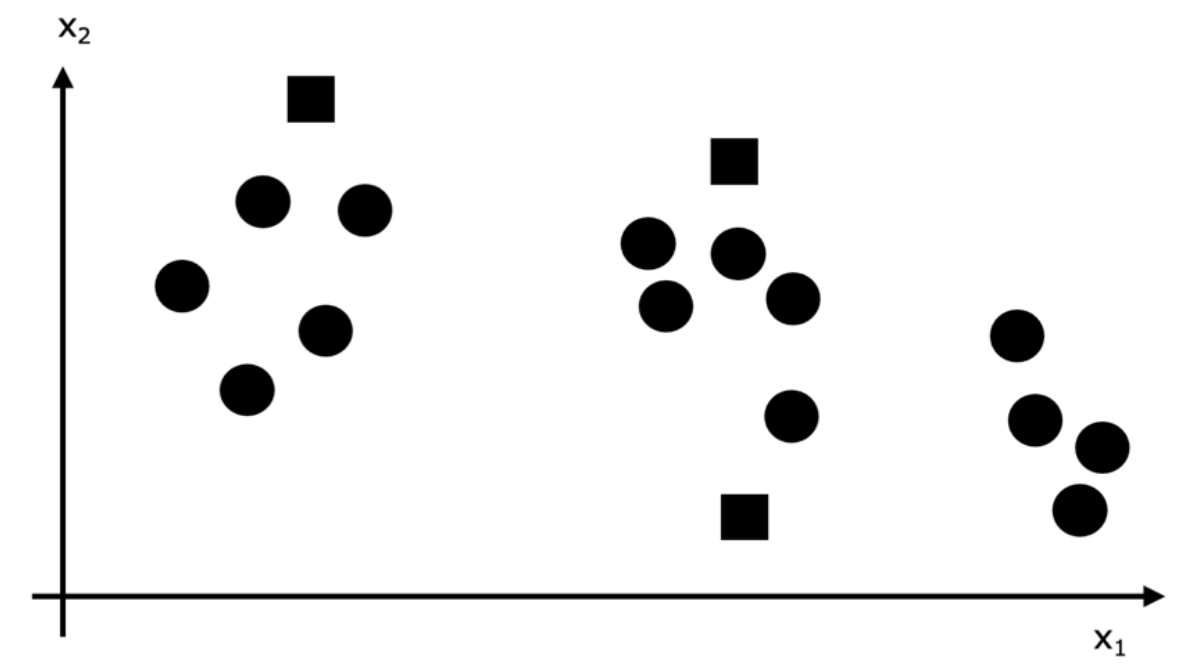


https://bookdown.org/mpfoley1973/data-sci/maximal-margin-classifier.html

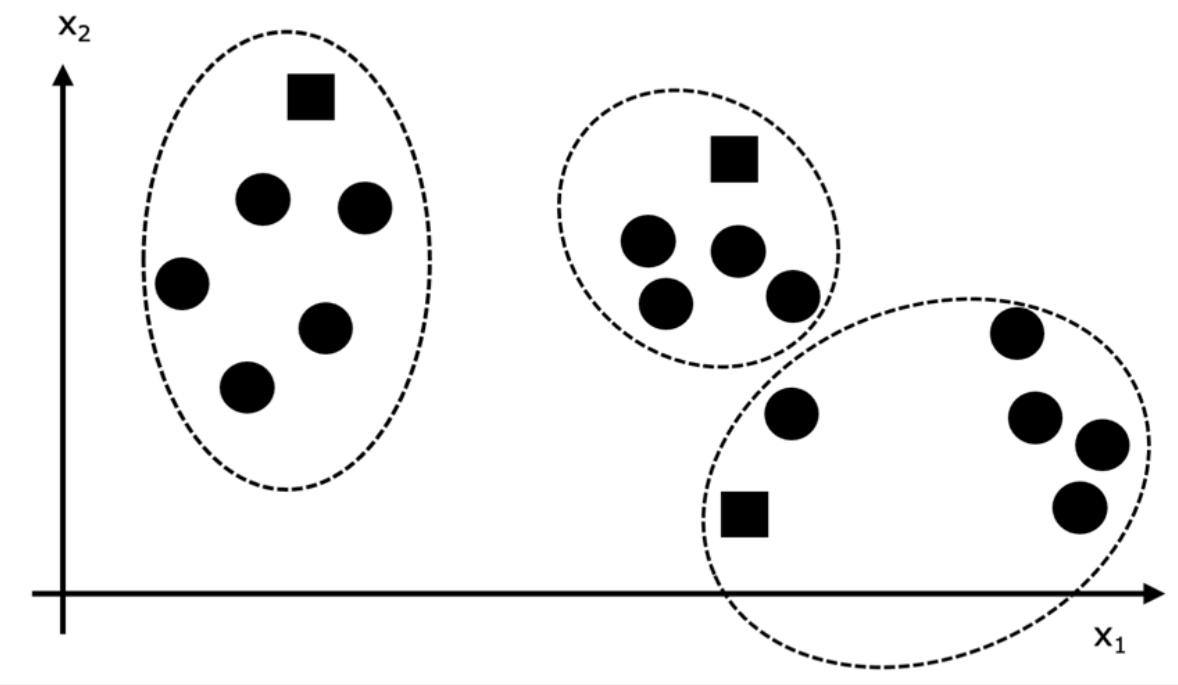
OUTLINE

- Literature Review Topics
- Classic Supervised Learning
  - Regression
  - Classification
- Classic Unsupervised Learning
- Implementing ML in Python
- Discussion

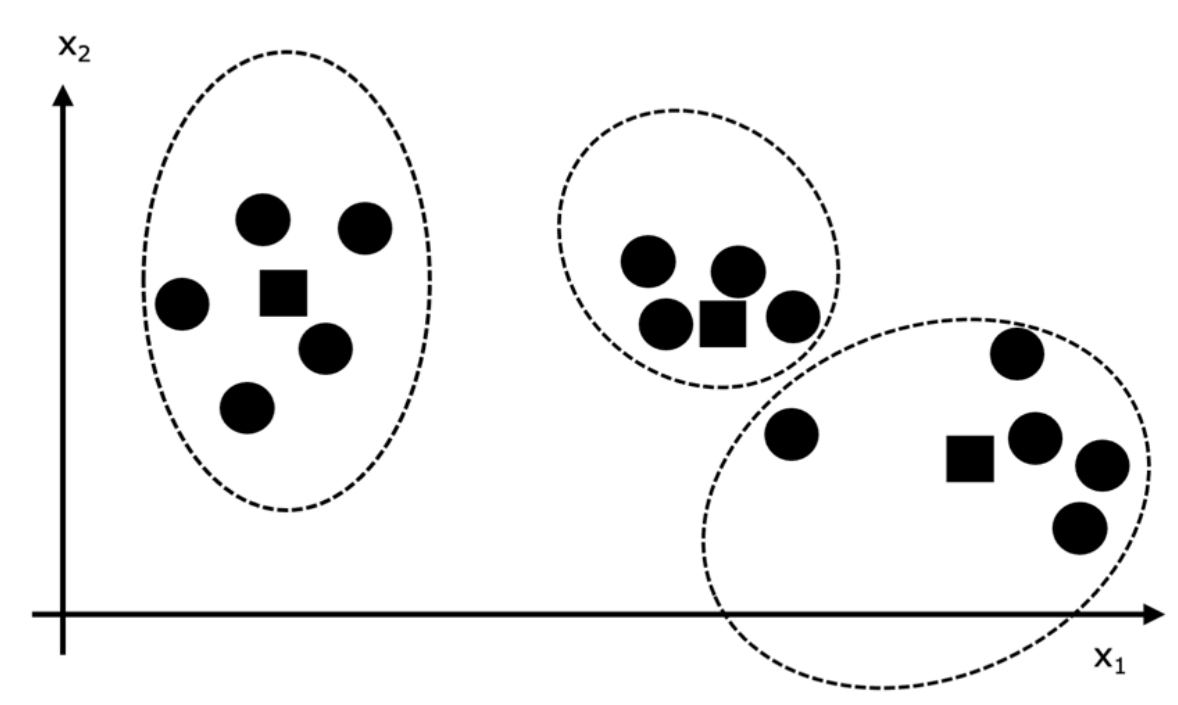
K-MEANS CLUSTERING



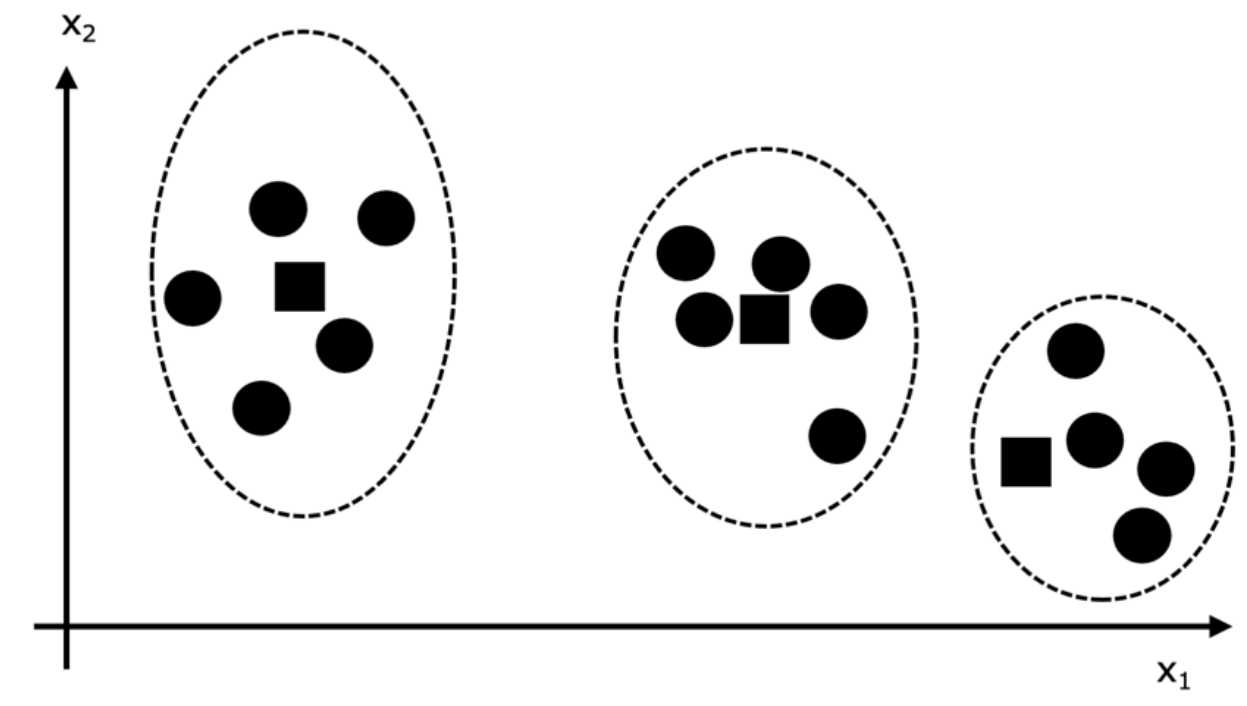
K-MEANS CLUSTERING



K-MEANS CLUSTERING



K-MEANS CLUSTERING



K-MEANS CLUSTERING

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x \in C_i} distance(x, mean(i))$$

SAMPLE APPLICATIONS OF UNSUPERVISED ML

- Detect similar groups of YouTube / Spotify / Amazon users to recommend similar content
- Targeted advertising
- Data exploration
- Anomaly/outlier detection, such as credit card fraud
- …

OUTLINE

- Literature Review Topics
- Classic Supervised Learning
  - Regression
  - Classification
- Classic Unsupervised Learning
- Implementing ML in Python
- Discussion

SKLEARN

```
>>> from sklearn.datasets import load_iris
>>> from sklearn.linear_model import LogisticRegression
>>> X, y = load_iris(return_X_y=True)
>>> clf = LogisticRegression(random_state=0).fit(X, y)
>>> clf.predict(X[:2, :]) array([0, 0])
>>> clf.predict_proba(X[:2, :])
array([[9.8...e-01, 1.8...e-02, 1.4...e-08], [9.7...e-01, 2.8...e-02, ...e-
08]])
>>> clf.score(X, y)
0.97...
```

SKLEARN

Implementing established models is easy using the sklearn module in Python

```
>>> from sklearn import tree
>>> X = [[0, 0], [1, 1]]
>>> Y = [0, 1]
>>> clf = tree.DecisionTreeClassifier()
>>> clf = clf.fit(X, Y)
>>> clf.predict([[2., 2.]])
array([1])
```

SKLEARN

- Implementing established models is easy using the sklearn module in Python
- Knowledge of how these algorithms work is not required to code them
- However, understanding how they work is ***crucial*** for selecting which algorithm and hyperparameters to choose for a given dataset, as well as for extending an algorithm for a new problem domain

CODING DEMO

https://colab.research.google.com/drive/1o7qfDOxrXMjr7auR4SzU6PPQqMqn9bwn?usp=sharing

USEFUL ML RESOURCES FOR BEGINNERS

High level easy-to-understand resources on the math / theory:

- https://www.youtube.com/c/joshstarmer/videos
- "The StatQuest Illustrated Guide to Machine Learning" by Josh Starmer

High level easy-to-understand resources on programming:

- https://scikit-learn.org/stable/tutorial/index.html
- https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb

OUTLINE

- Literature Review Topics
- Classic Supervised Learning
  - Regression
  - Classification
- Classic Unsupervised Learning
- Implementing ML in Python
- Discussion

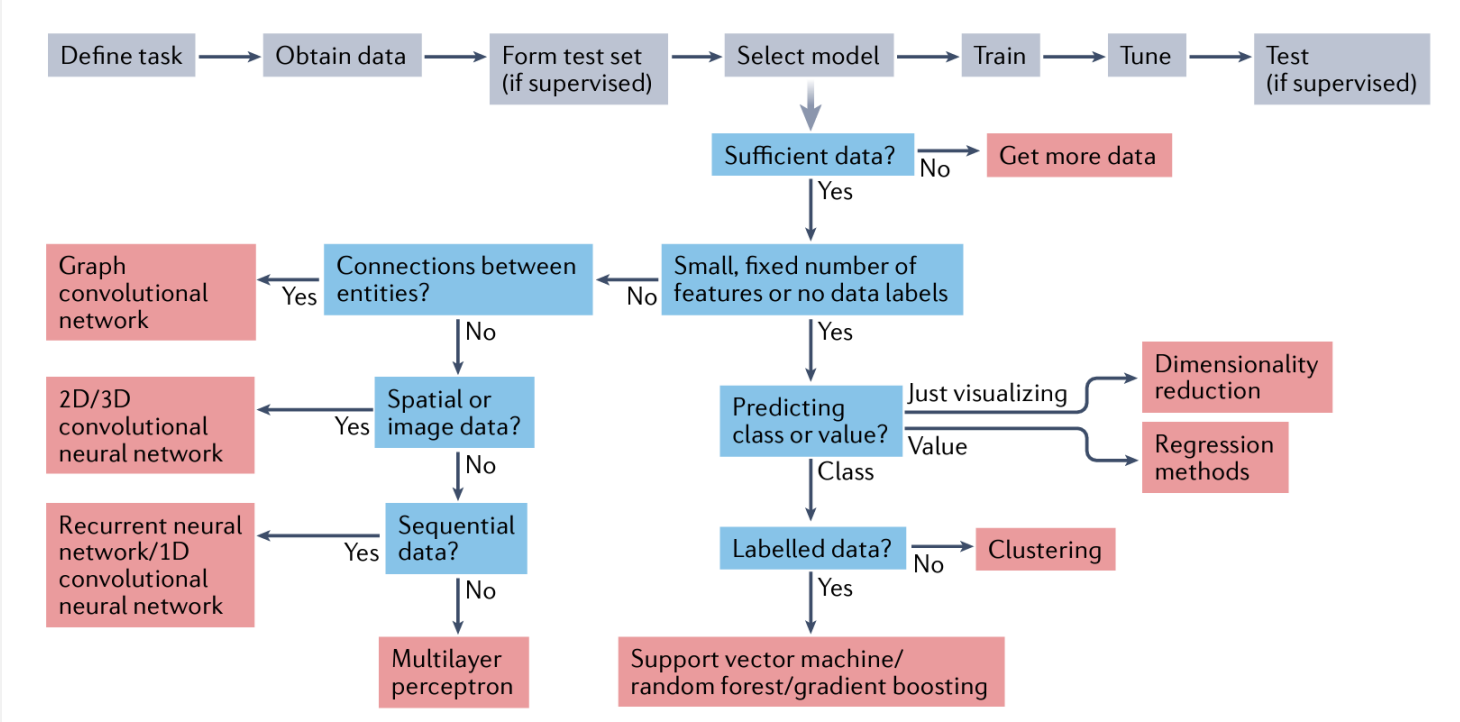
TODAY'S PAPER

A guide to machine learning for biologists

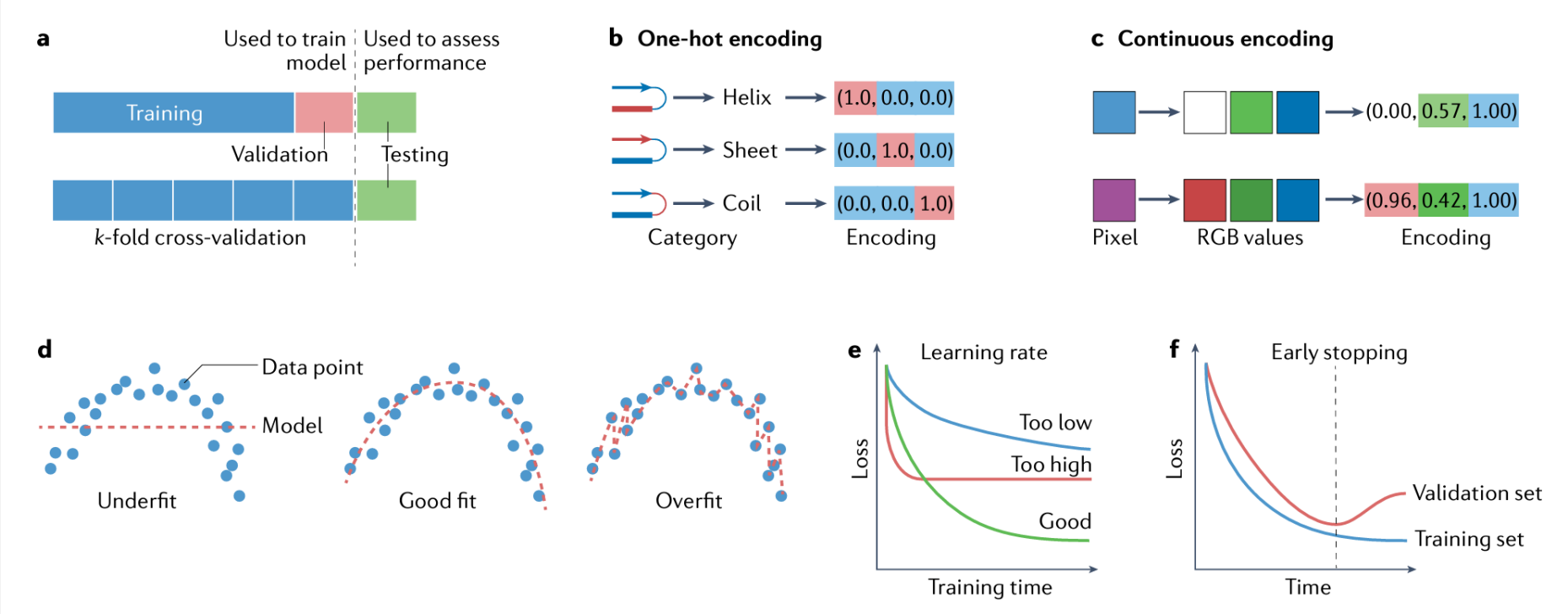
Joe G. Greener[1,2], *Shaun M. Kandathil*[1,2], *Lewis Moffat*[1] *and David T. Jones*[1] ✉

Abstract | The expanding scale and inherent complexity of biological data have encouraged a growing use of machine learning in biology to build informative and predictive models of the underlying biological processes. All machine learning techniques fit models to data; however, the specific methods are quite varied and can at first glance seem bewildering. In this Review, we aim to provide readers with a gentle introduction to a few key machine learning techniques, including the most recently developed and widely used techniques involving deep neural networks. We describe how different techniques may be suited to specific types of biological data, and also discuss some best practices and points to consider when one is embarking on experiments involving machine learning. Some emerging directions in machine learning methodology are also discussed.

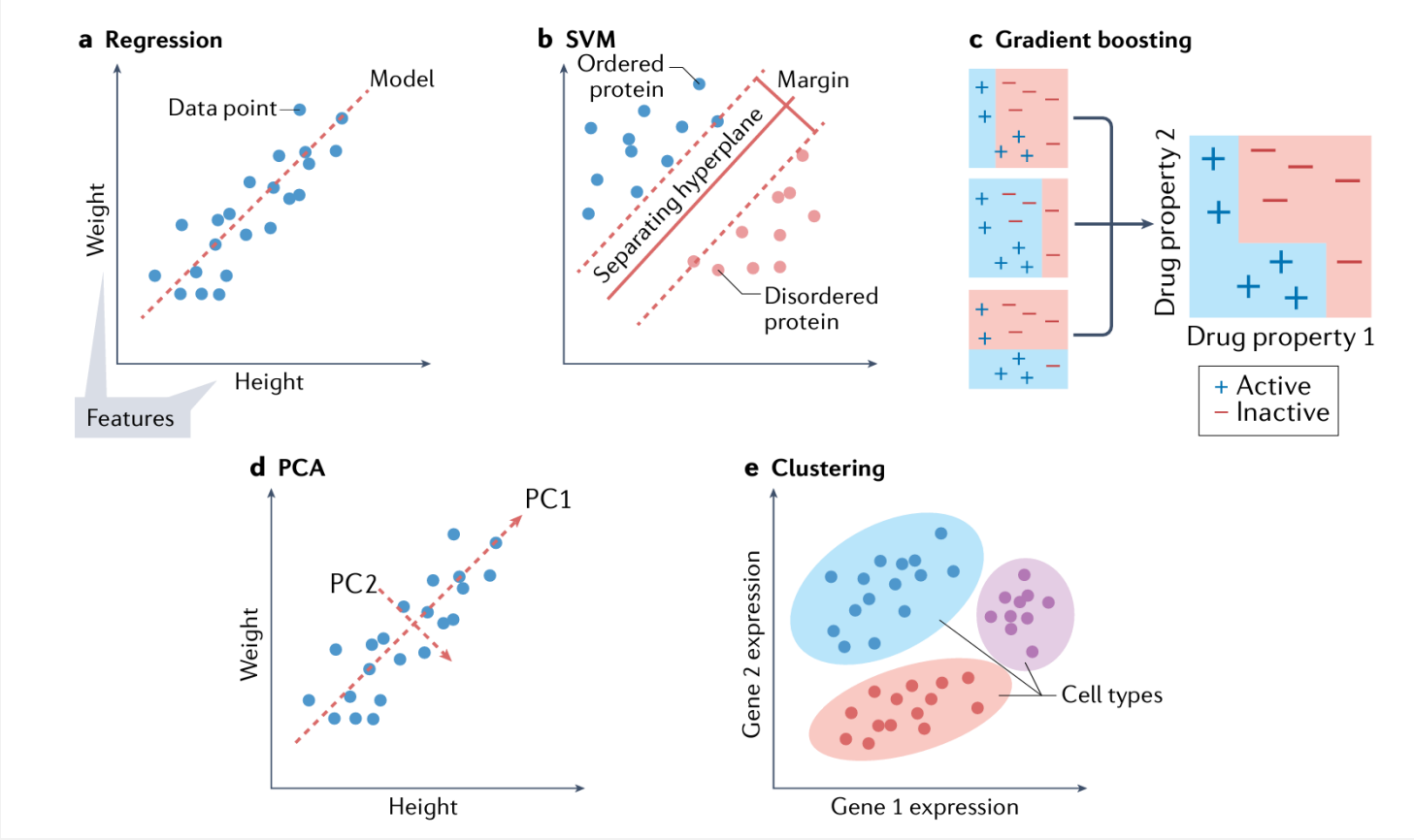
ML PIPELINE / CHOOSING A MODEL



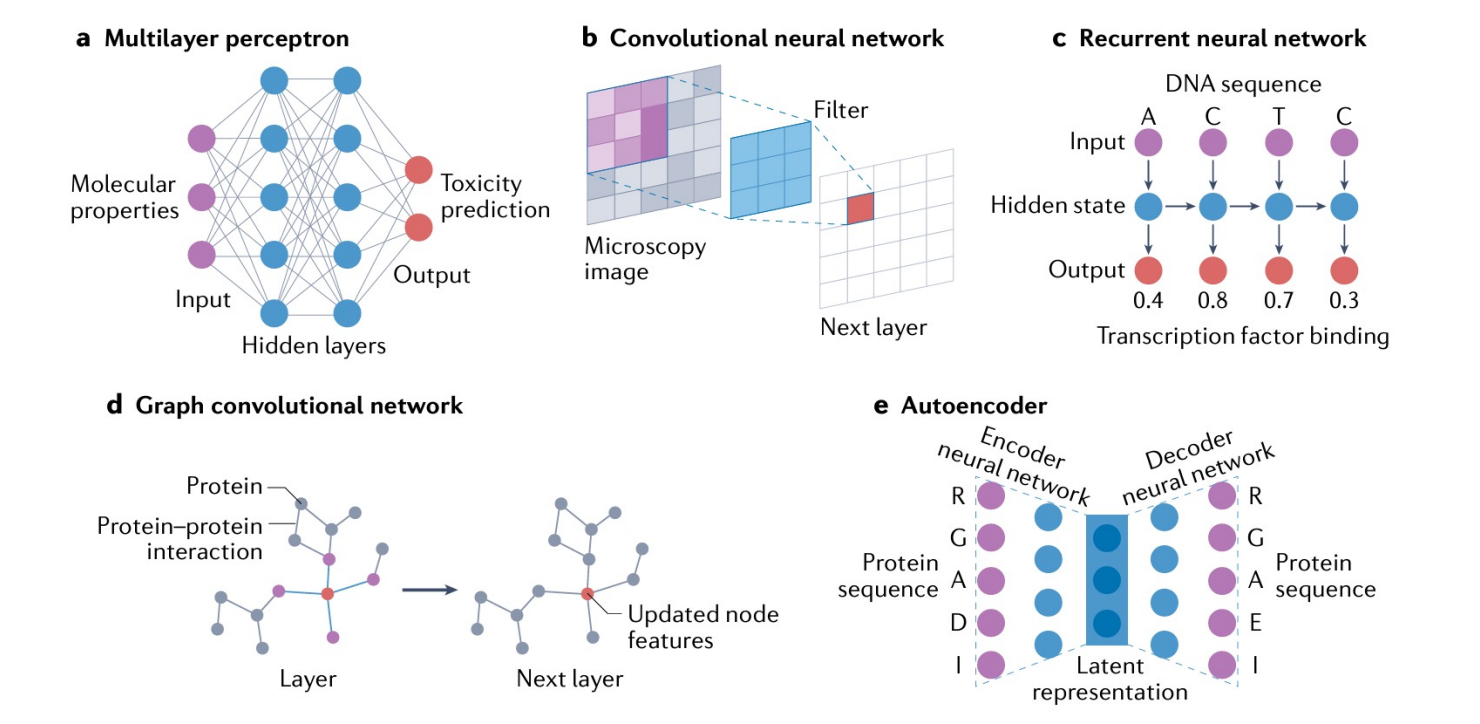
MODEL TRAINING CONSIDERATIONS



TRADITIONAL ML METHODS



NEXT TIME: NEURAL NETWORKS AND DEEP LEARNING



USEFUL TOOLS FOR EVALUATING THE ARTICLES WE READ IN THIS CLASS

Box 2 | **Evaluating articles that use machine learning**

Here are some questions to consider when reading or reviewing articles that use machine learning on biological data. It is useful to bear these considerations in mind, even if the answers are not apparent, and these questions can be used as the basis for a discussion with collaborators with the required expertise. A surprising number of articles do not fulfil these criteria[148].

Is the dataset adequately described?
Complete steps to assemble the dataset should be provided, ideally with the dataset or summary data (for example, biological database IDs) available at a persistent URL. In our experience, a thorough description of the machine learning method but with only a cursory reference to the data is a red flag. If a standard dataset or a dataset from another study is being used, then this should be adequately justified in the article.

Is the test set valid?
Based on the discussion in the section Challenges for biological applications, check that the test set is sufficient to benchmark the property under investigation. There should be no data leakage between the training set and the test set, the test set should be of large enough to give reliable results and the test set should mirror the range of examples a standard user of the tool would be likely to use it on. The composition and size of the training and test sets should be discussed in detail. Authors have a responsibility to ensure that all steps have been taken to avoid data leakage, and these steps should be described in the article, along with the rationale behind them. Journal editors and peer reviewers should also ensure that these tasks have been performed to a good standard, and certainly should never just assume that they have been.

Is the model choice justified?
Reasons should be given for the choice of machine learning method. Neural networks should be used because they are appropriate for the data and question in hand, and not just because everyone else is using them. Discussion of models that were tried and did not work should be encouraged as it may help others; too often a complex model is presented without any discussion of the inevitable trial and error that will have been required to end up with that model.

Has the method been compared with other methods?
A novel method should be compared with existing methods that show good performance and are used in the community. Ideally methods using a variety of model types should be compared, which can aid in interpreting results. It is surprising how many complex models can be matched in performance by simple regression methods.

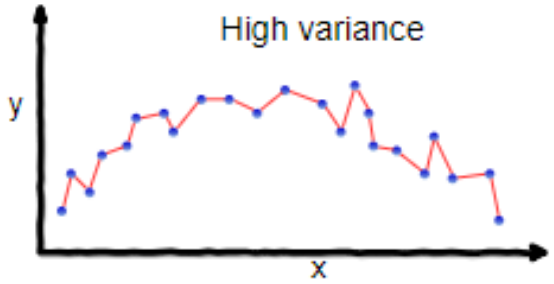
Are the results too good to be true?
Claims of greater than 99% accuracy are not uncommon in machine learning articles in biology. Usually, this is a sign of a problem with the testing rather than an amazing breakthrough. Both authors and reviewers should take note of this point.

Is the method available?
At the very least, someone who wants to use a trained model from an article should be able to run a prediction using a Web service or binary file. Ideally, at least source code and the trained model should be available at a persistent URL and under a common licence[149,150]. Also making the training code available is the ideal scenario, as this further increases the reproducibility of the article and allows other researchers to build on the method without essentially having to start from scratch. Journals should bear some responsibility here to ensure that this becomes the norm.

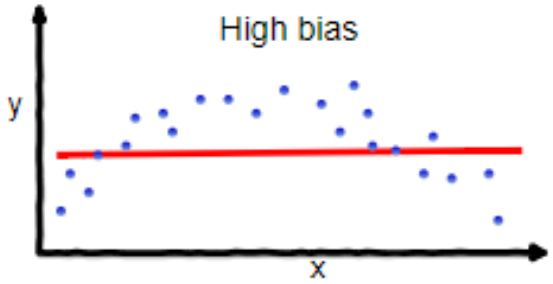
DISCUSSION QUESTIONS

1. What are human-centered implications of the bias-variance tradeoff described in the paper?
2. What are the tradeoffs between sensitivity and specificity for a medical diagnostic classifier? If we had to choose one, which would be prefer to be higher? What about for precision and recall?
3. What are the advantages / disadvantages between non-parametric models (e.g., k-Nearest Neighbors) and parametric models (everything else we talked about)?
4. How might unsupervised learning approaches be used for supervised learning tasks?
5. What are the potential long-term implications of ML algorithm details being abstracted away into APIs?
6. Are fundamental aspects of programming, data analysis, and ML learnable in grade school? If so, should these concepts be taught as core courses?
7. ML seems to be integrated into almost every field. What are some examples of domains where ML would be a bad solution?

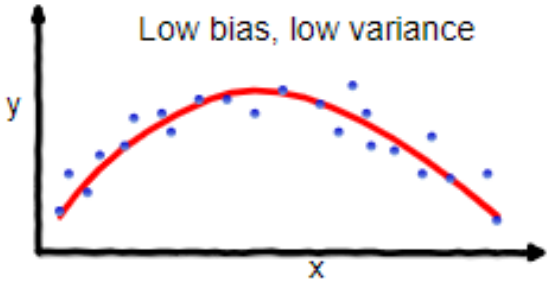
BIAS VARIANCE TRADEOFF



overfitting



underfitting



Good balance

https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229