Intro to Machine Learning

Federal Research Division - Feb 1, 2024

"A breakthrough in machine learning would be worth ten Microsofts"

— Bill Gates, Microsoft Co-Founder

“Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed”

— Arthur L. Samuel, AI pioneer, 1959

(This is likely not an original quote but a paraphrased version of Samuel's sentence ”Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort.”)

Arthur L Samuel. “Some studies in machine learning using the game of checkers”. In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229.

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

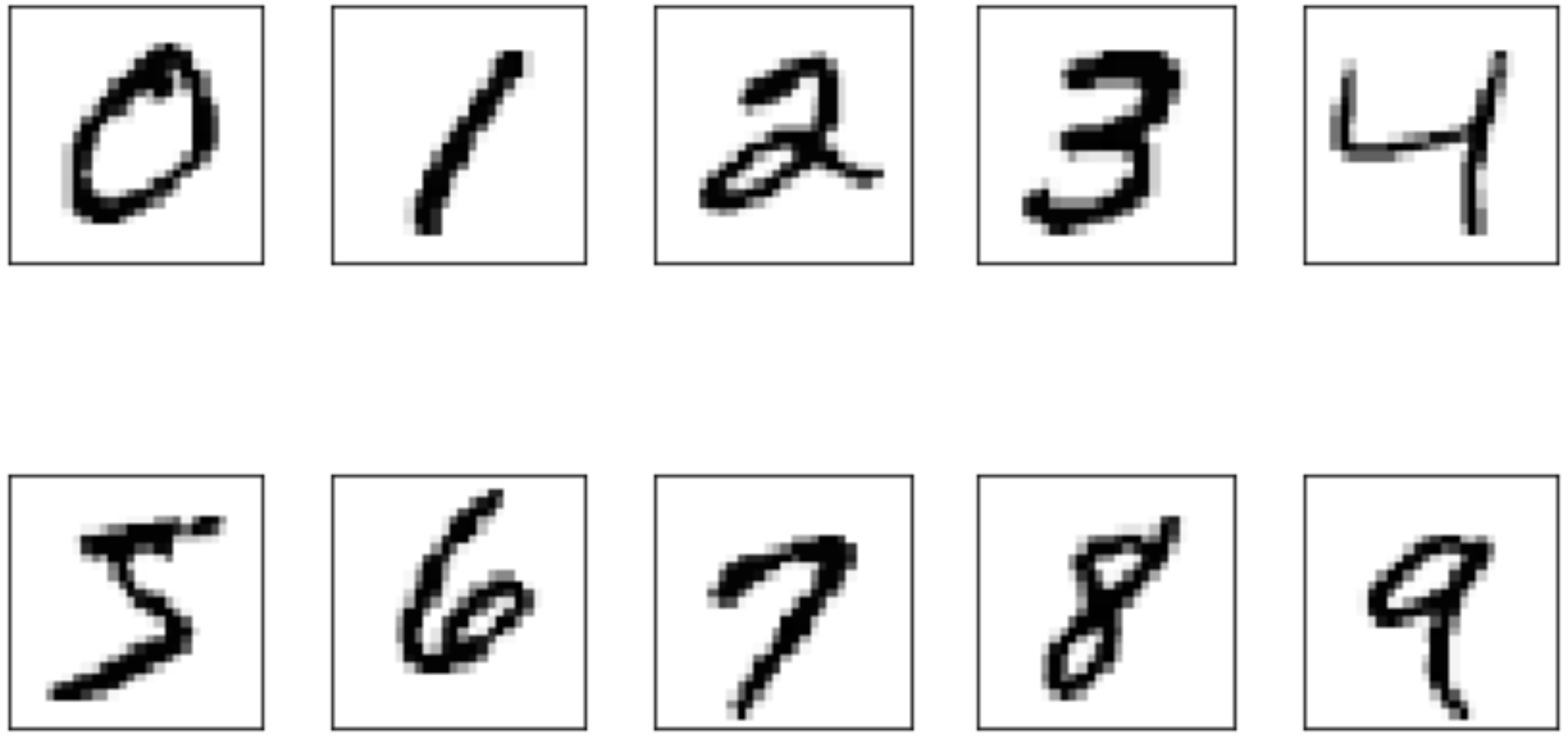
— Tom Mitchell, Professor at Carnegie Mellon University

Tom M Mitchell et al. "Machine learning. 1997". In: *Burr Ridge, IL: McGraw Hill* 45.37 (1997), pp. 870–877.

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

— Tom Mitchell, Professor at Carnegie Mellon University

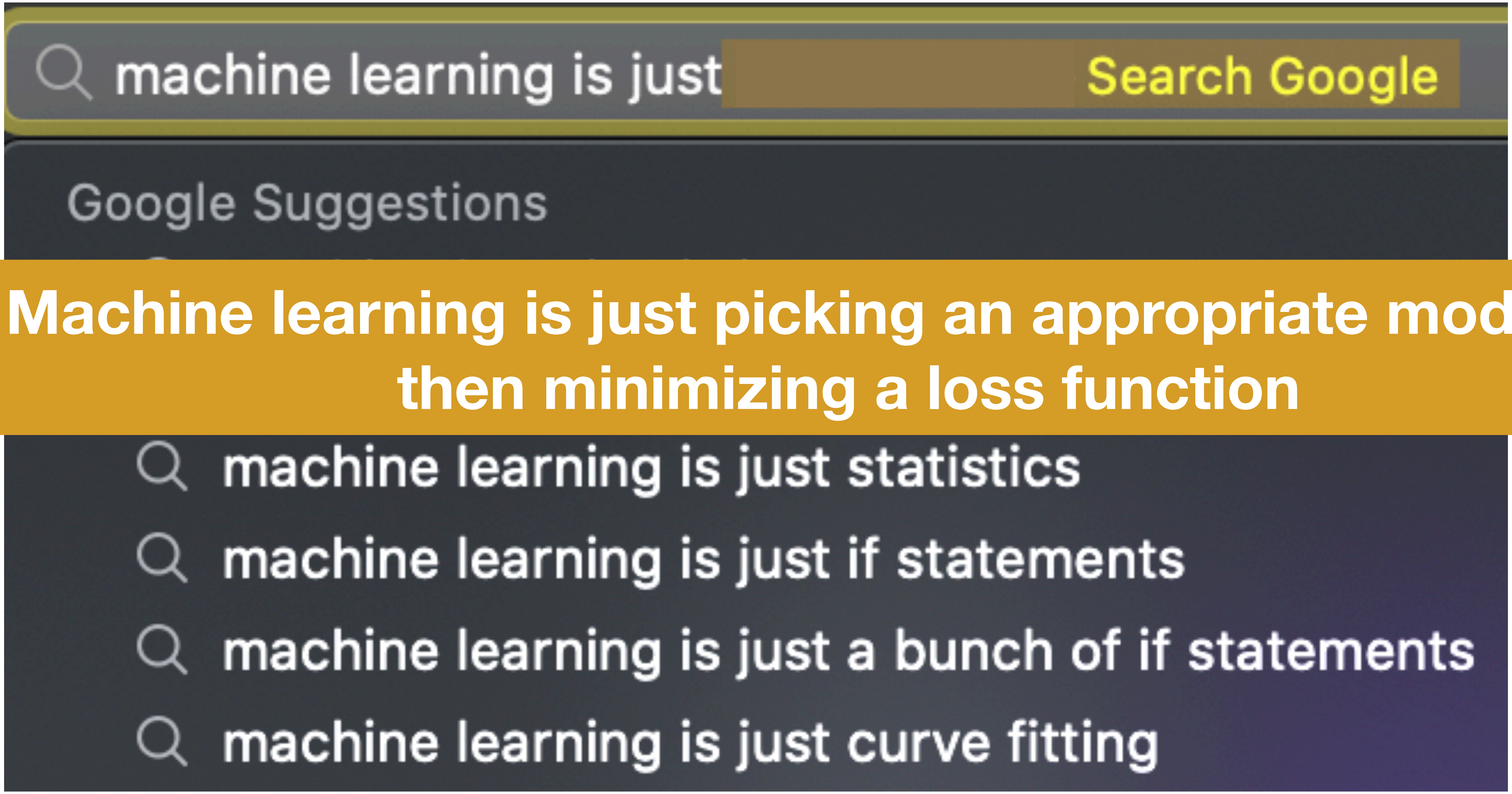
Handwriting Recognition Example:



- Task T: ?
- Performance measure P : ?
- Training experience E: ?

What the heck is Machine Learning?

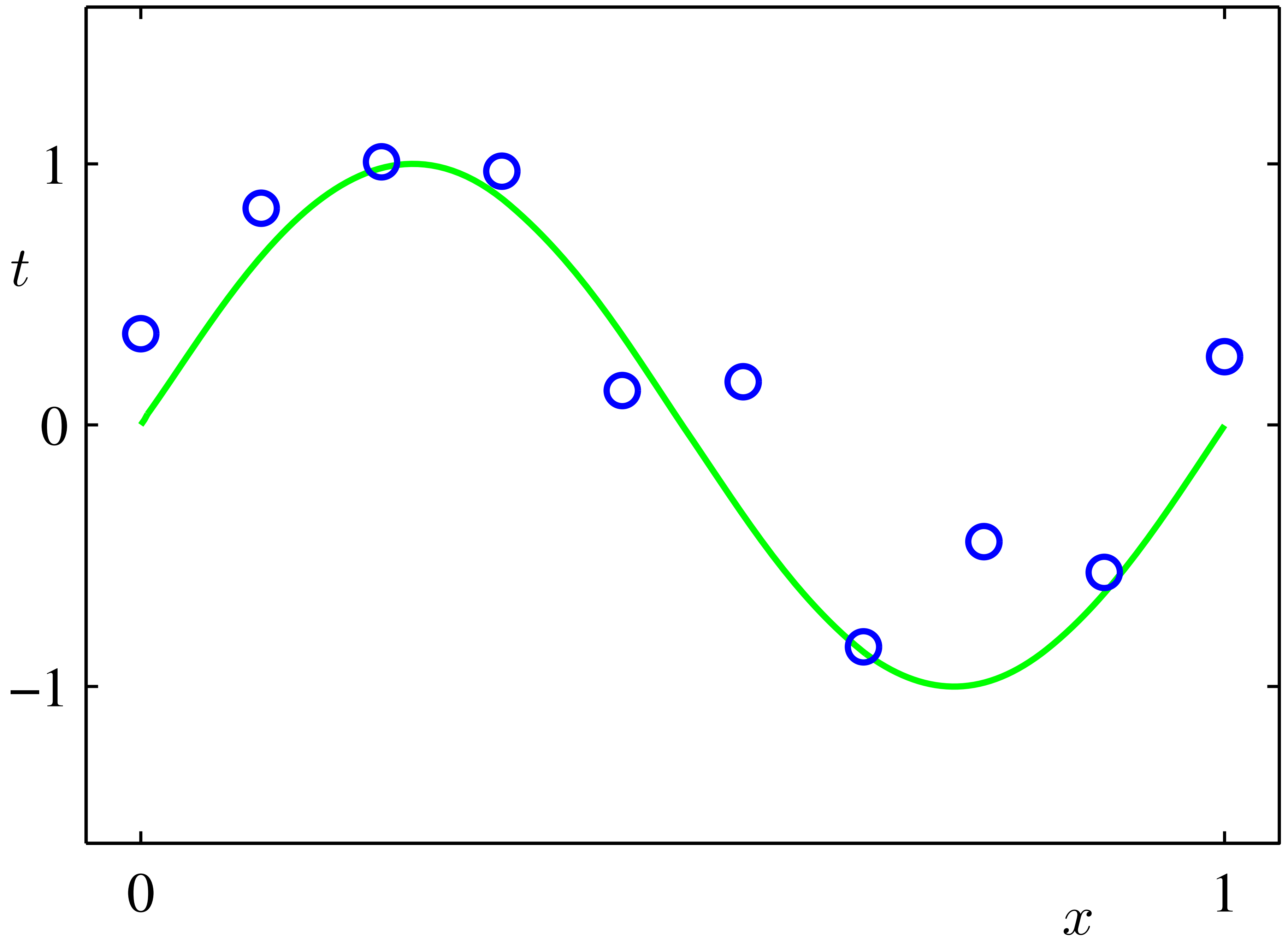
Machine learning is just picking an appropriate model and then minimizing a loss function

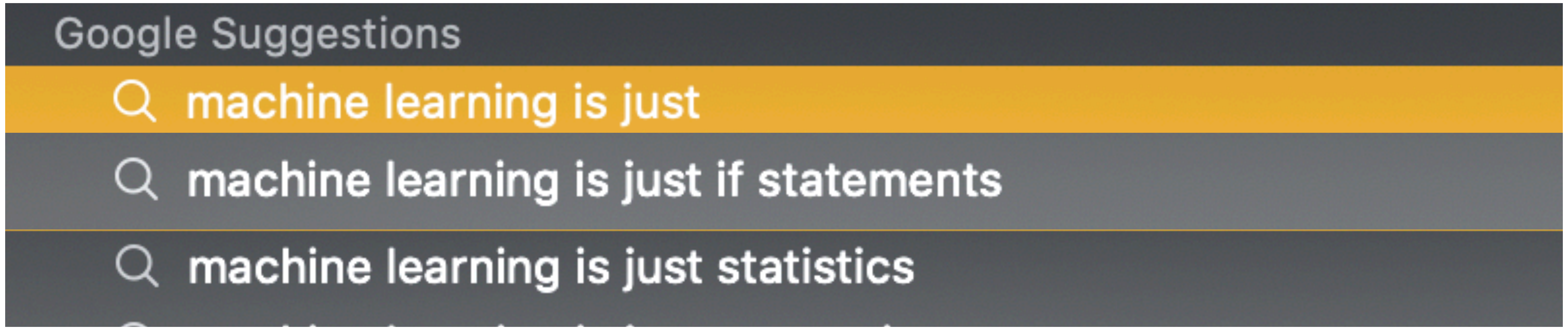


$$y(x, \mathbf{w}) = \mathbf{w} \cdot \mathbf{x} + b$$

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Plot of a training data set of $N = 10$ points, shown as blue circles, each comprising an observation of the input variable x along with the corresponding target variable t. The green curve shows the function $\sin(2\pi x)$ used to generate the data. Our goal is to predict the value of t for some new value of x, without knowledge of the green curve.





$$\text{if } (\%\texttt{george} < 0.6)\ \&\ (\%\texttt{you} > 1.5) \quad \begin{array}{l}\text{then } \texttt{spam}\\ \text{else } \texttt{email}.\end{array}$$

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* `spam` *and* `email`*.*

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

ML “Fields”

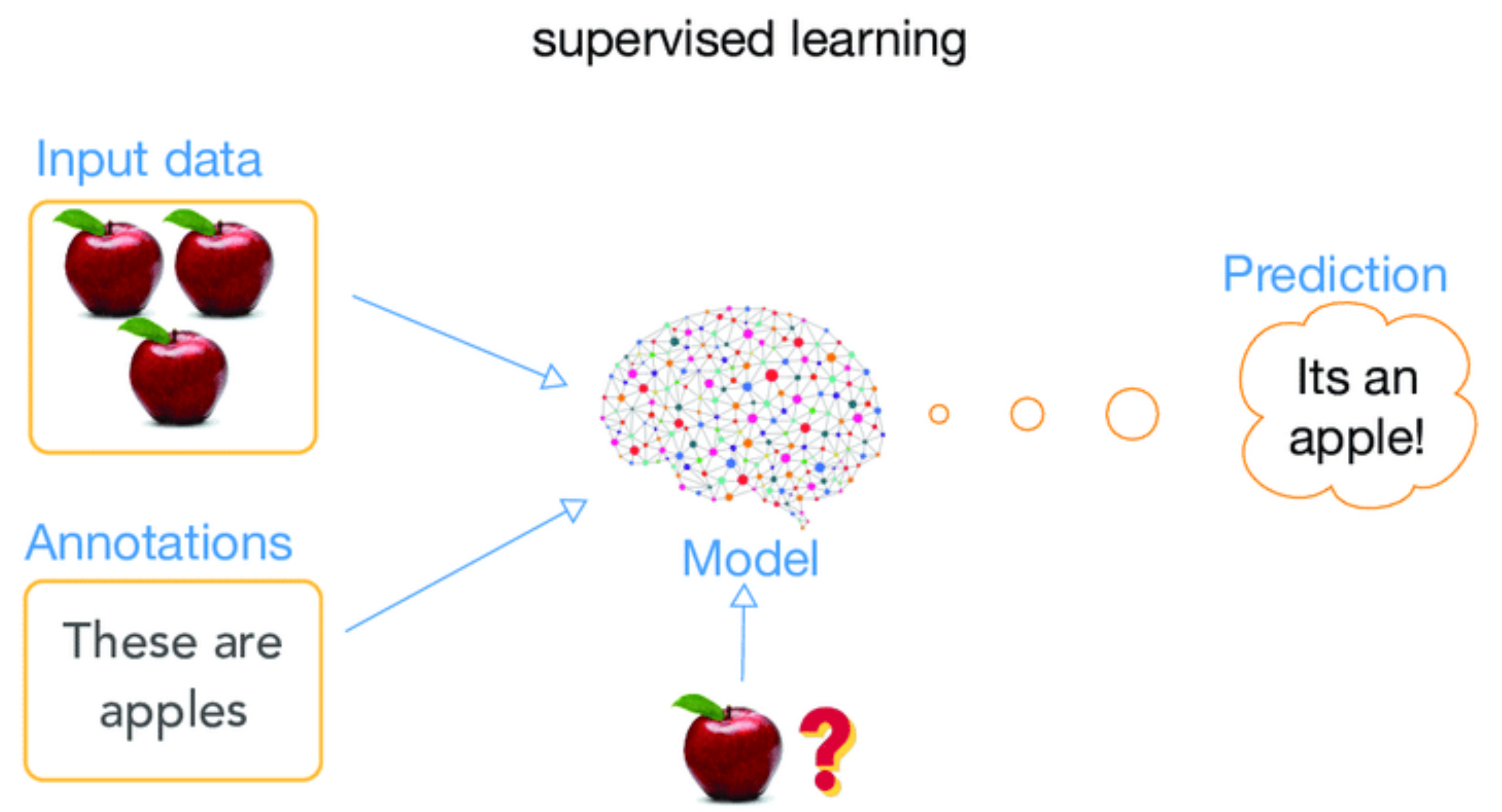
(Place Chart Here)

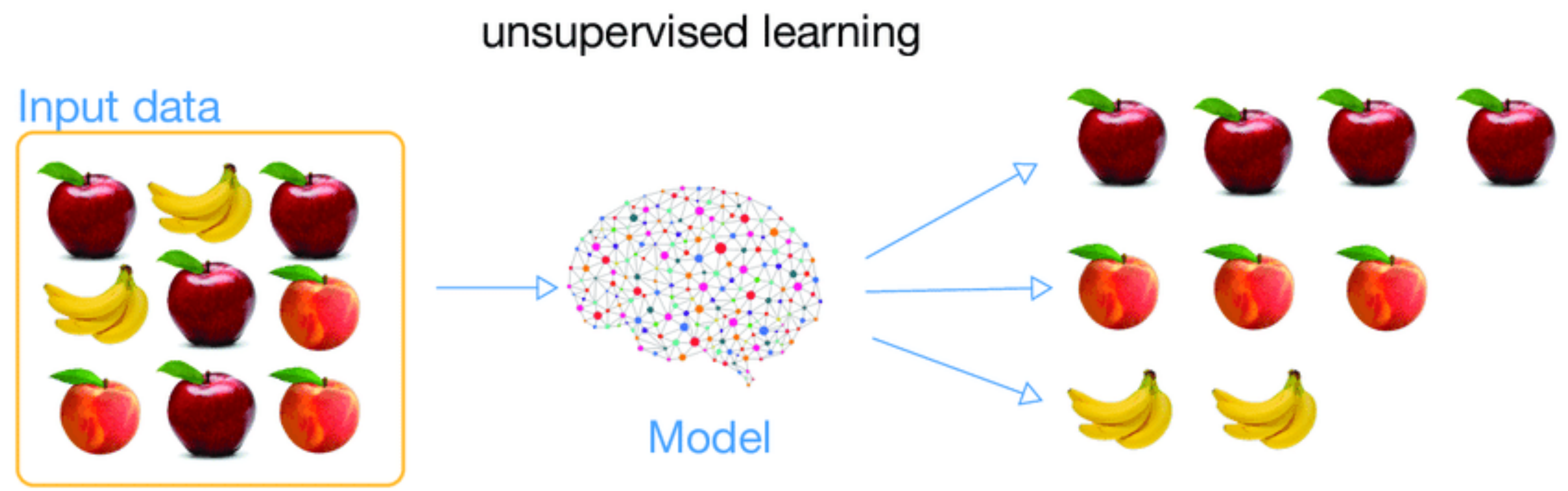
Supervised

- Labelled data
- Direct feedback
- Predict, classify, or fit a model

Unsupervised

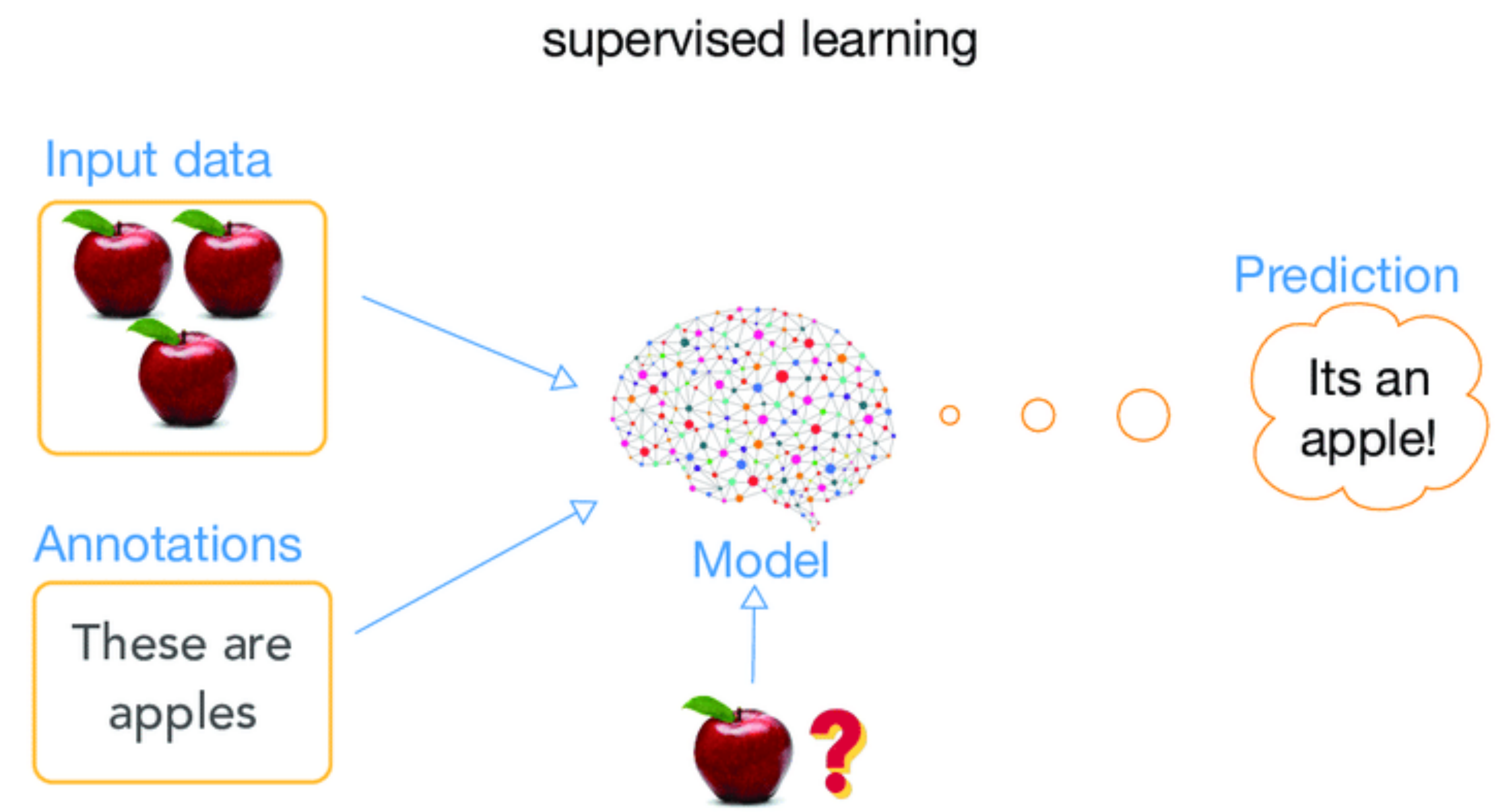
- No labels
- No feedback
- Find hidden structure using a model





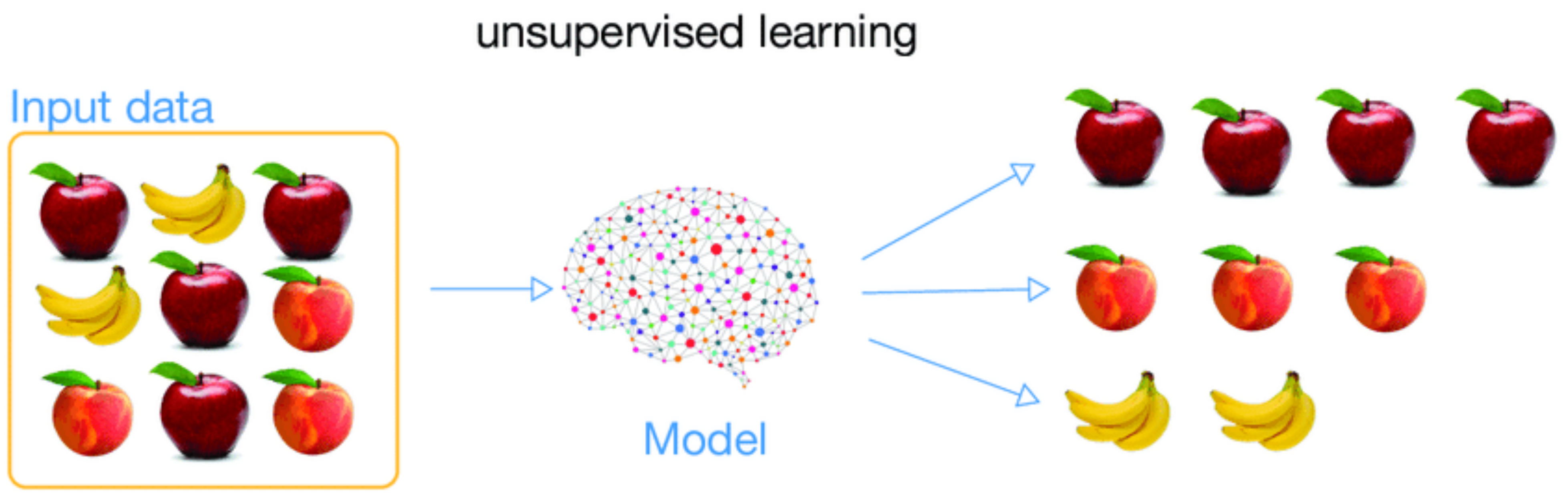
Supervised

- Labelled data
- Direct feedback
- Predict, classify, or fit a model



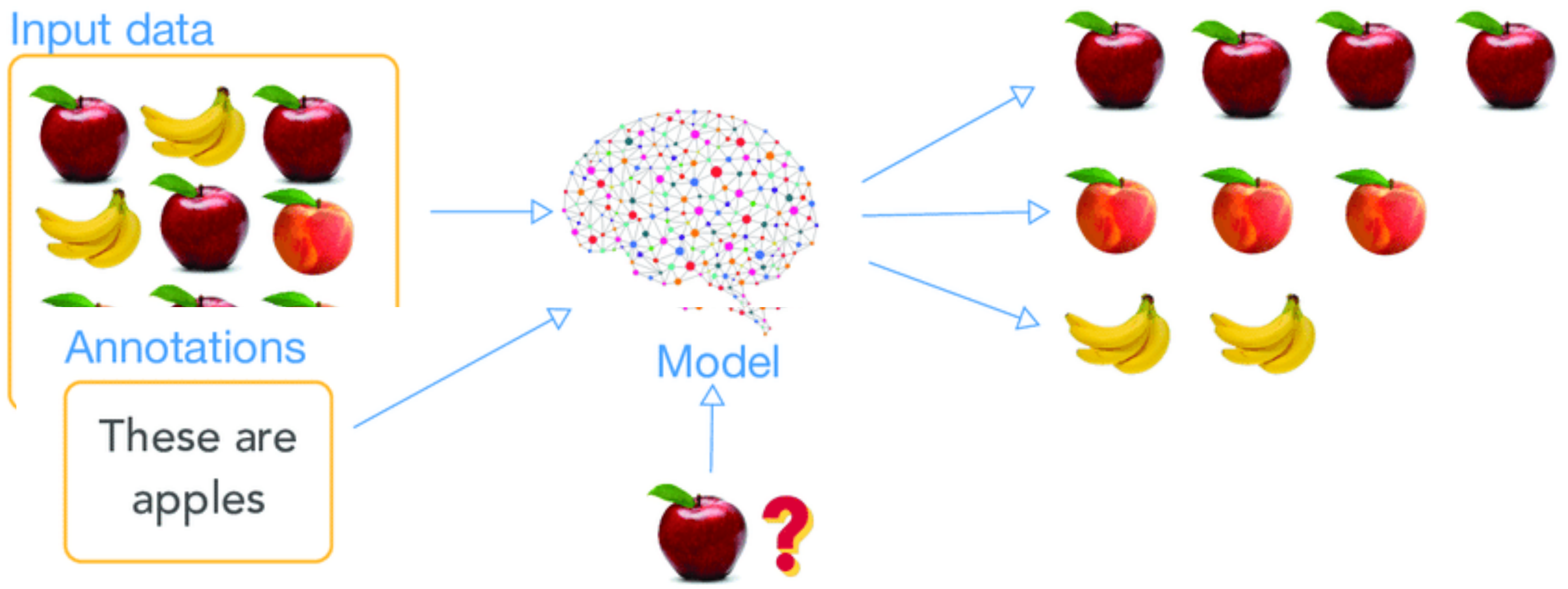
Unsupervised

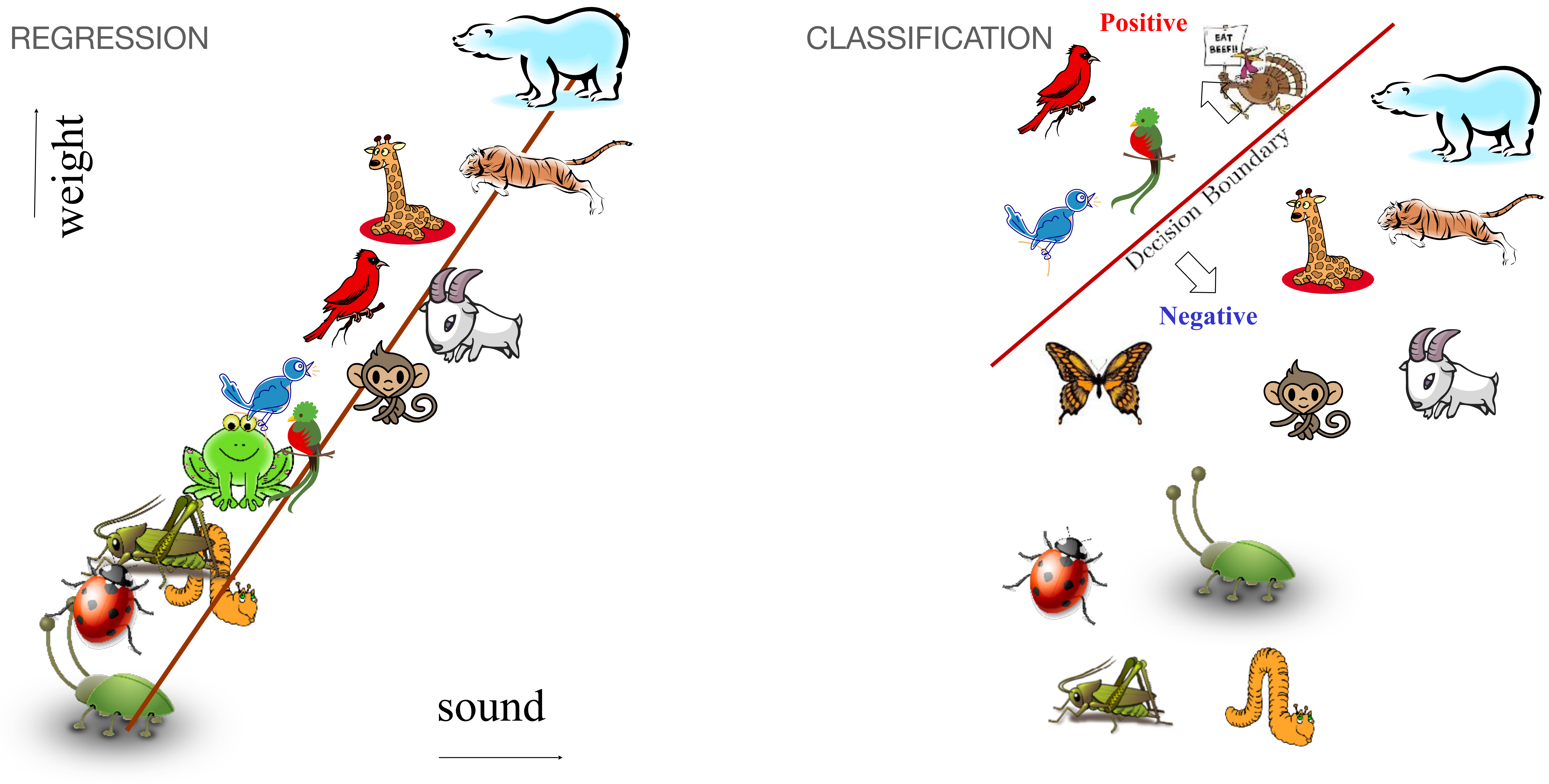
- No labels
- No feedback
- Find hidden structure using a model



Reinforcement learning

- Feedback via reward (only label)
- Learns the series of actions that lead to reward



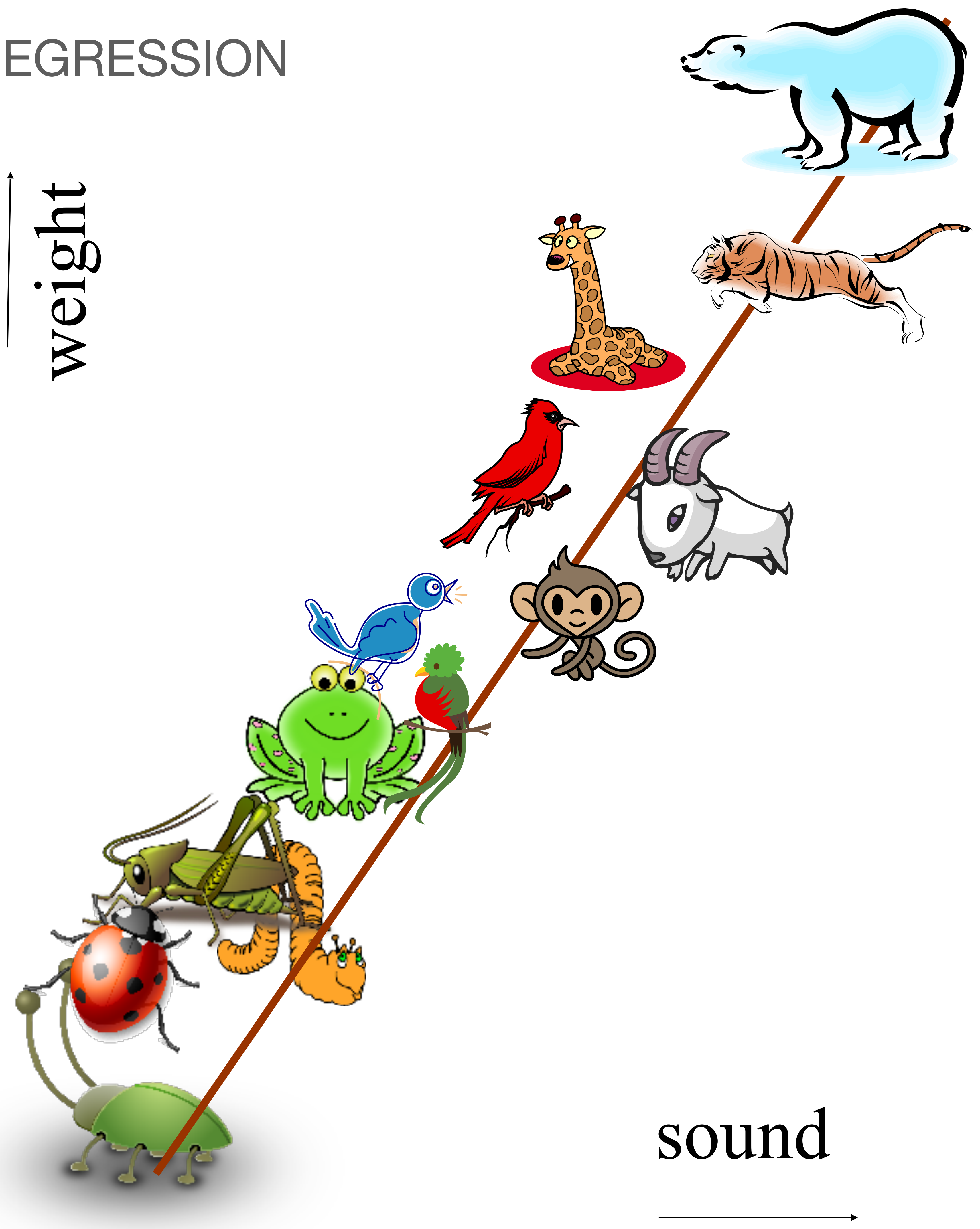


REGRESSION
weight
sound
CLASSIFICATION
Positive
EAT BEEF!!
Decision Boundary
Negative

Regression

In regression, the "ground-truth" labels are continuous values that are ORDINAL

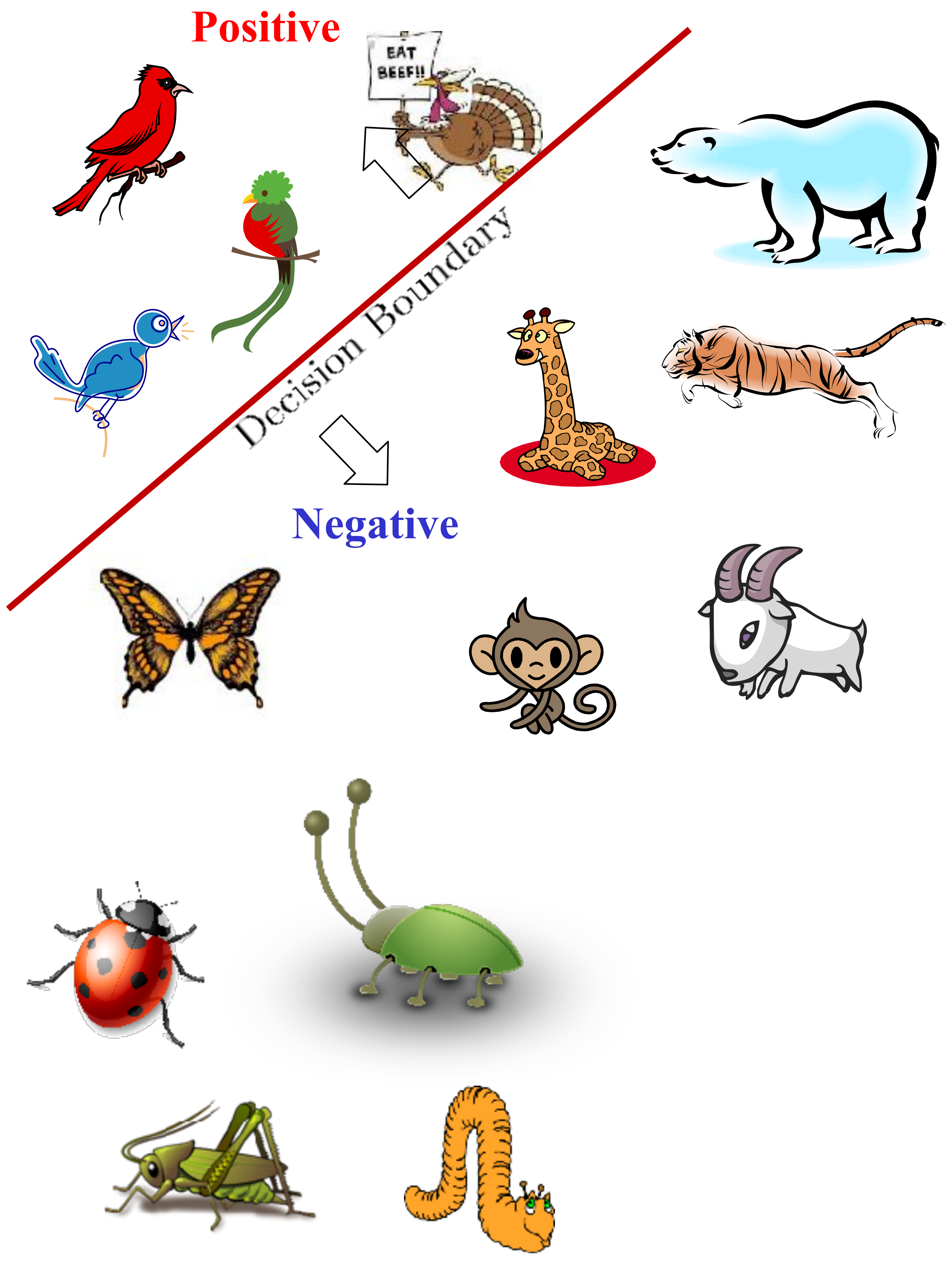
REGRESSION



That is, you can compare $y_1 = -1$ and $y_2 = 1$ to state for example $y_1 < y_2$ $(-1 < 1)$.

$$y \in \mathbb{R}$$

Classification



In classification, the "ground-truth" labels are categorical labels that are NOT ORDINAL

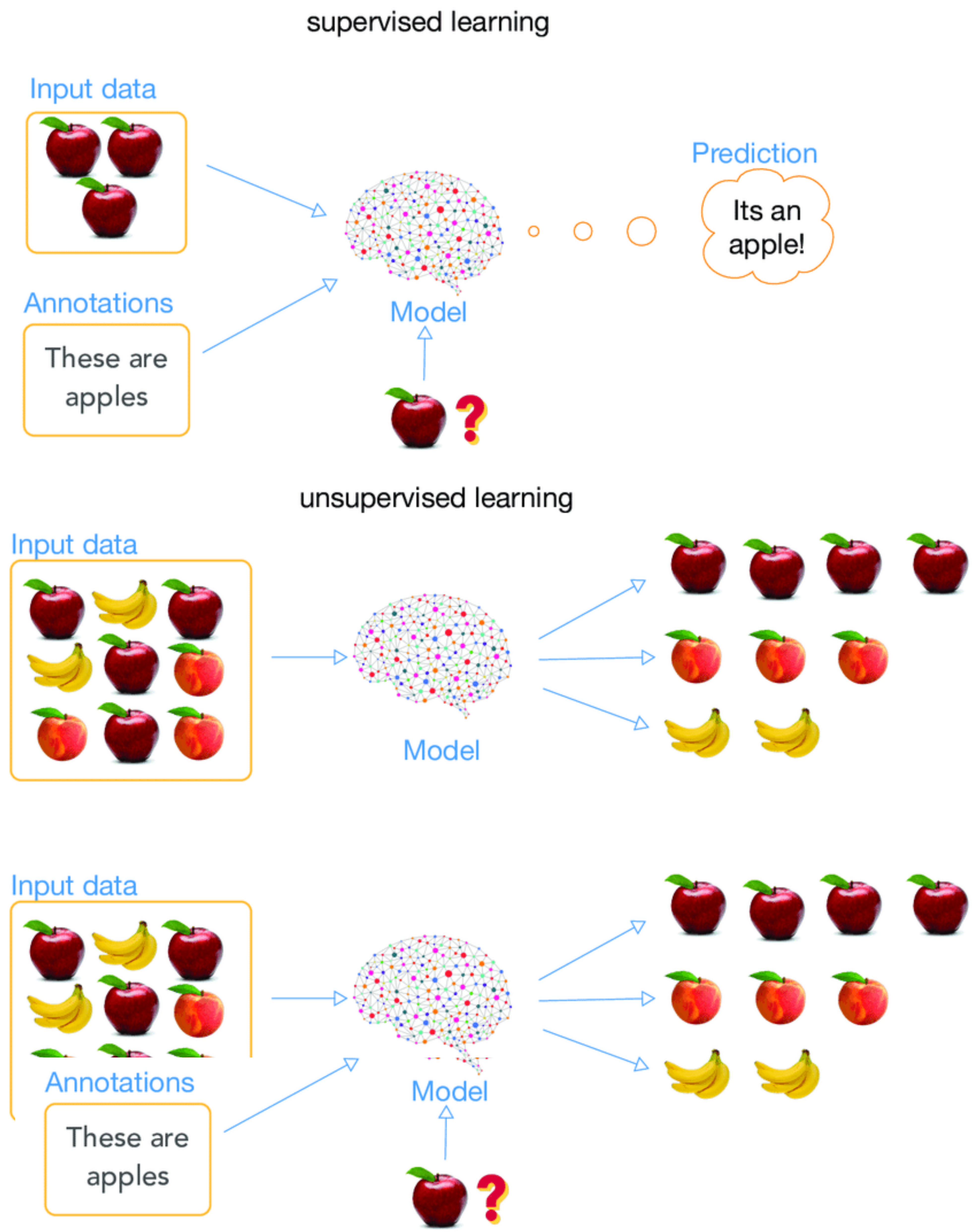
That is, you can not compare $y_1 =$ "$bird$" and $y_2 =$ "$non - bird$") to state for example:

"bird" < "non-bird" (even we use $+1$ and -1 to denote their labels respectively).

$$y \in \{-1, +1\}$$

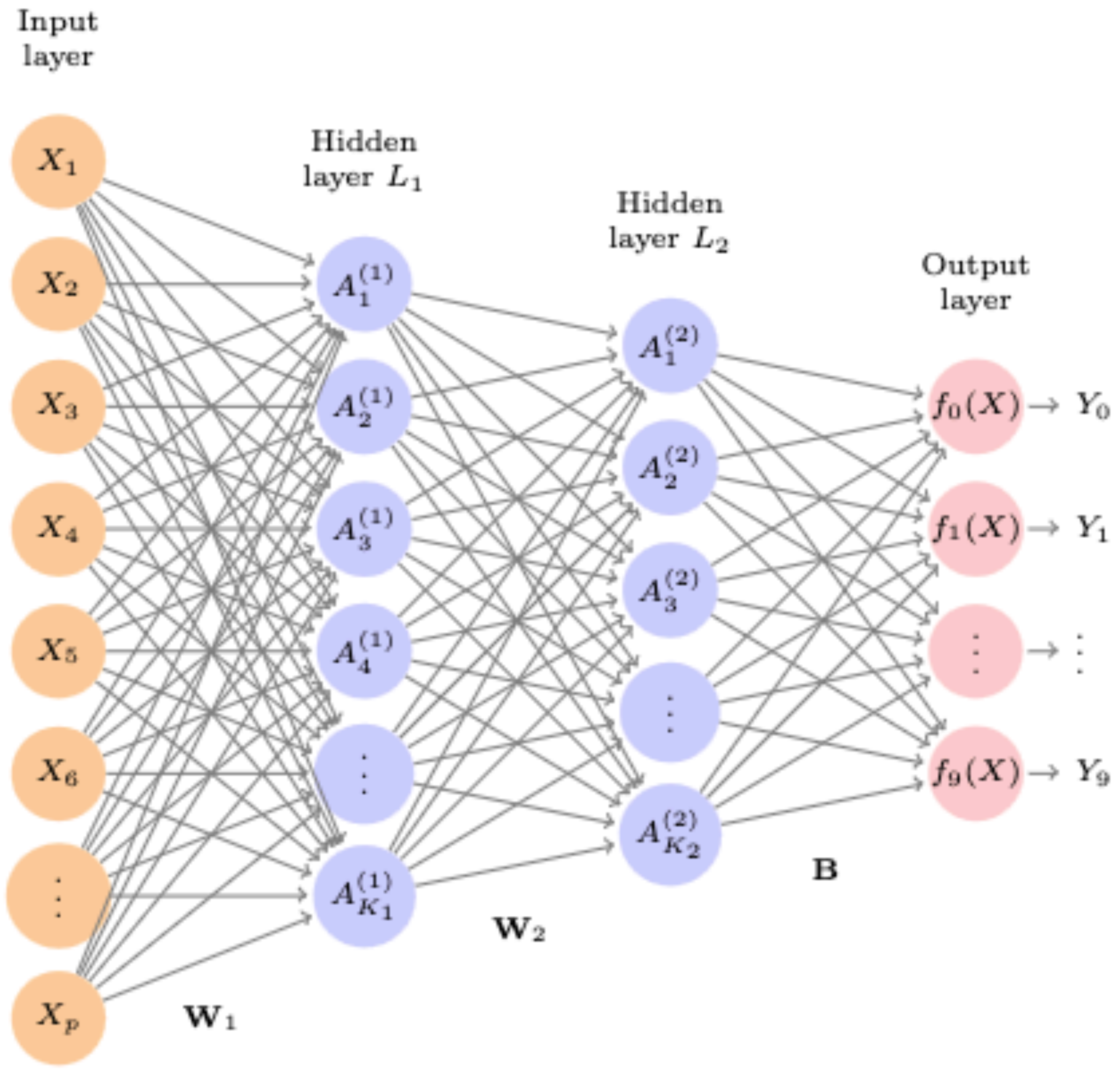
Neural Networks

- Mostly supervised and self-supervised. Some unsupervised
- Learn to predict/classify

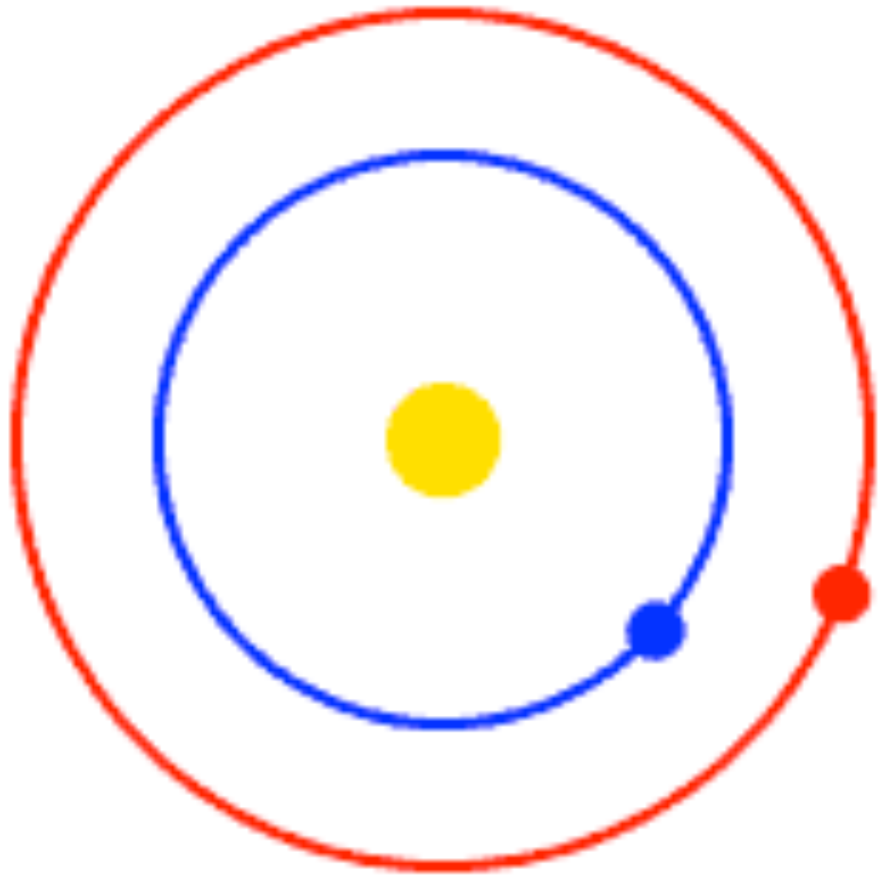


Neural Networks

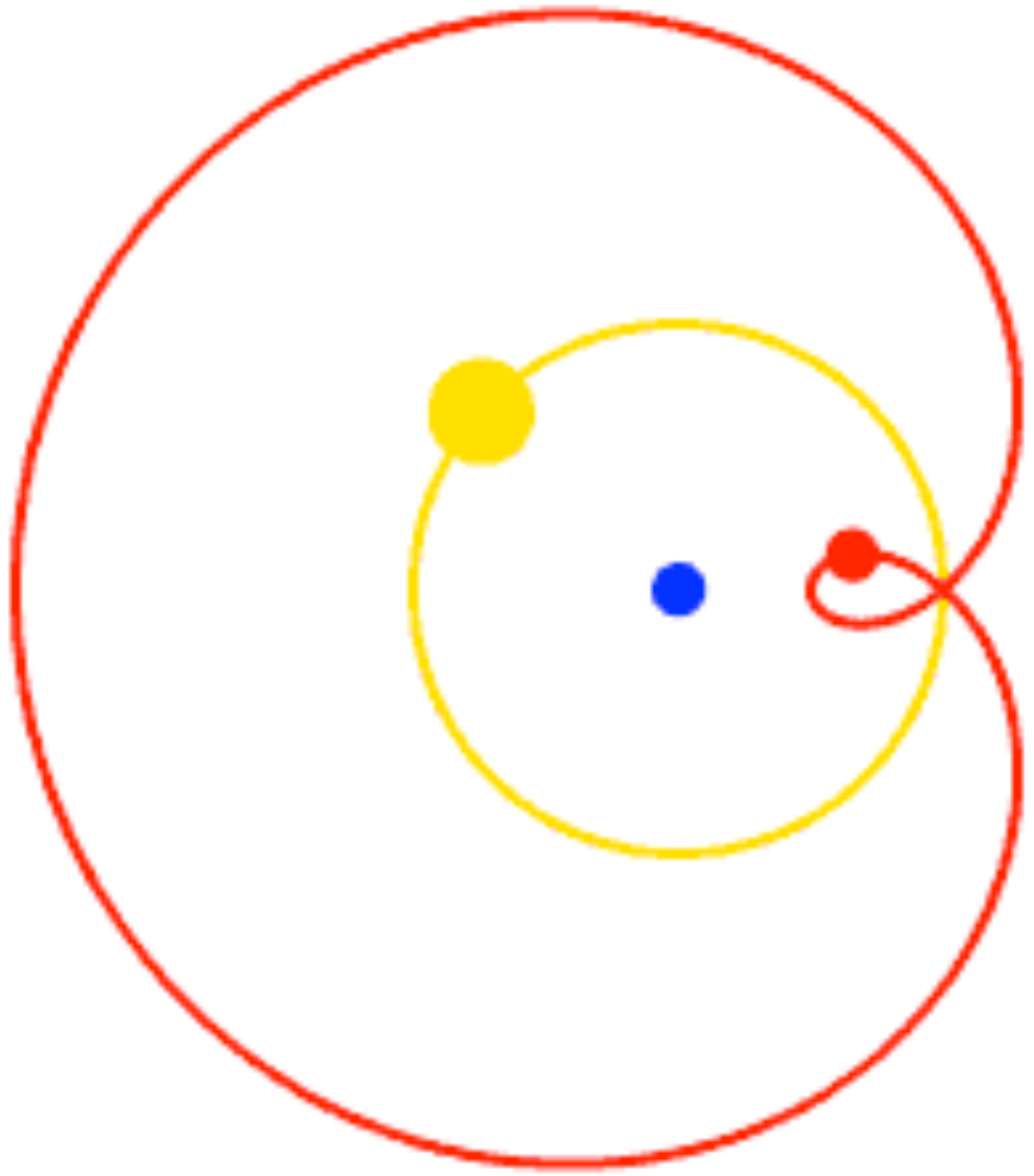
- Mostly supervised and self-supervised. Some unsupervised
- Learn to predict/classify



Predict / classify or model?



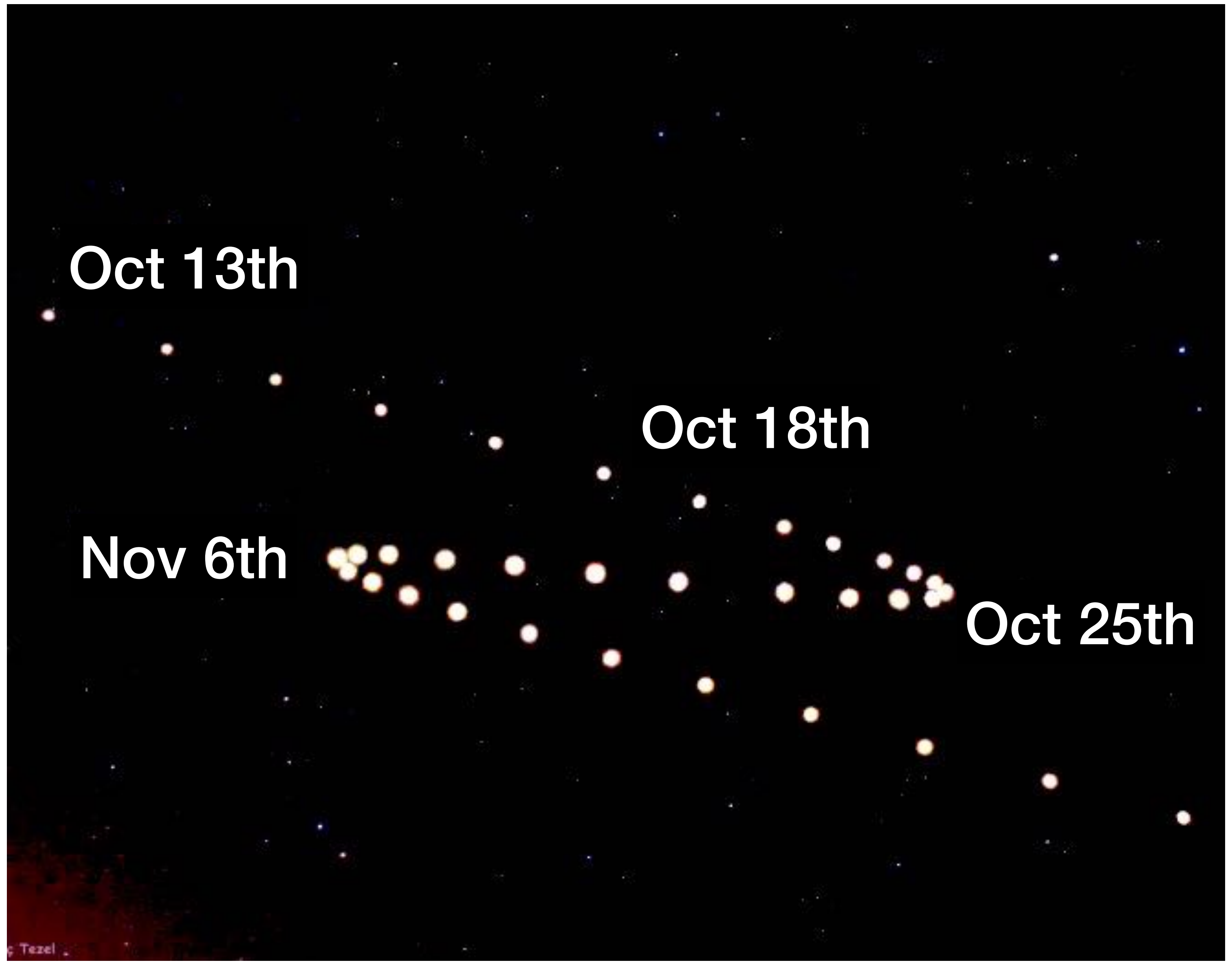
Heliocentric



Geocentric

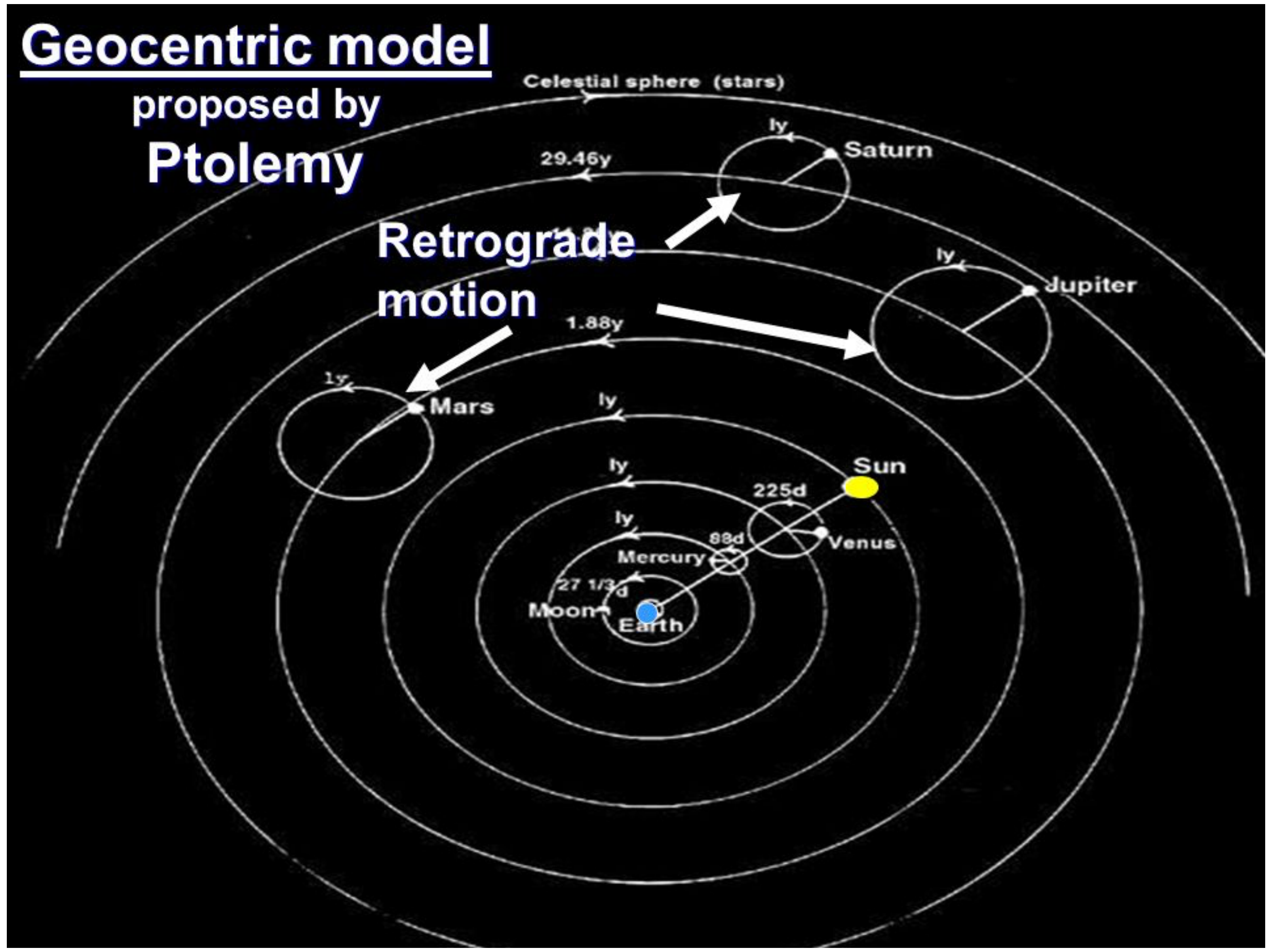
What is the goal?

Prediction



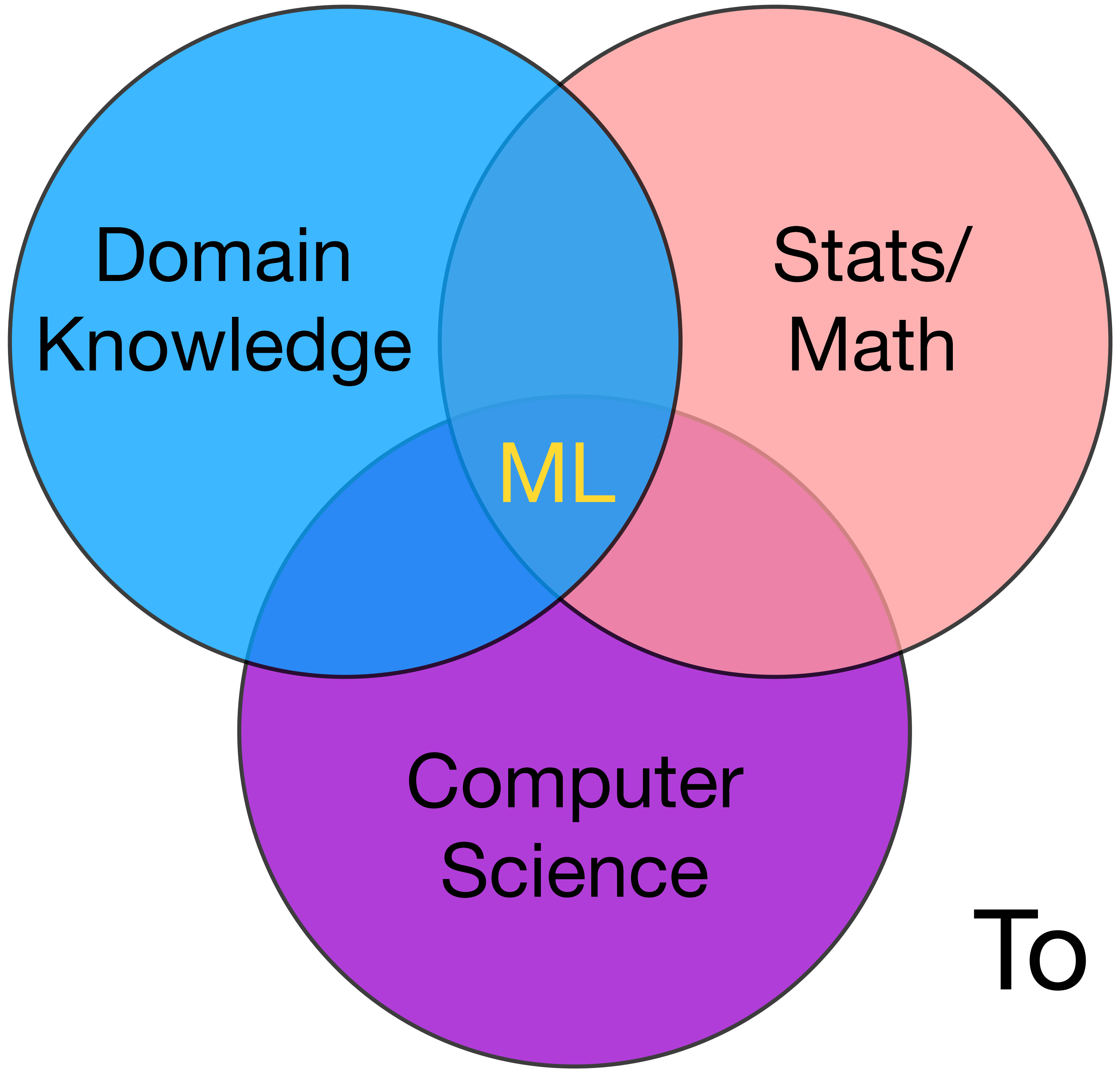
Only care that prediction y has low error

Modeling



We care that model w is an accurate representation of the real thing

What are you getting in to?



To do ML right, you need:

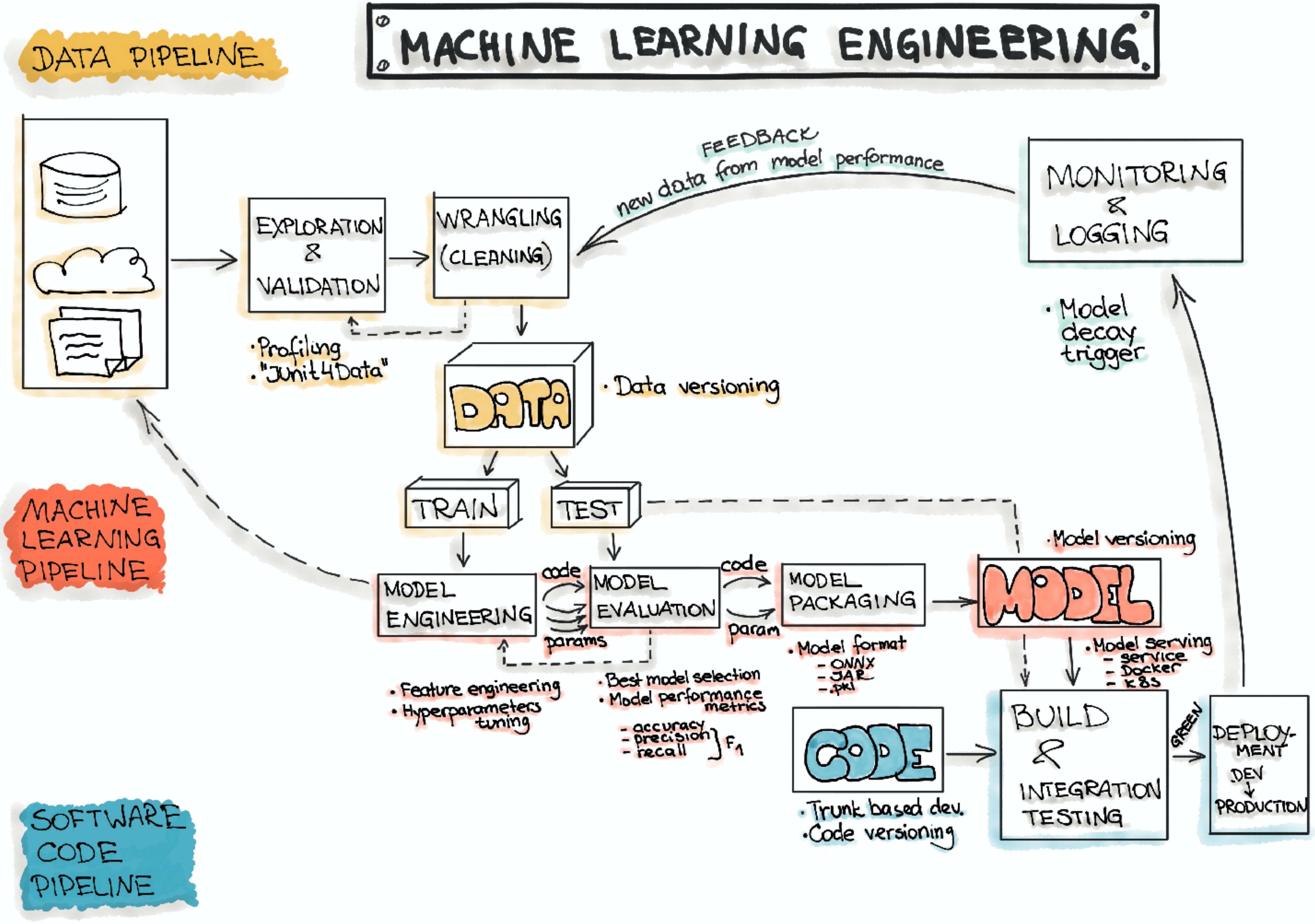
-Math/Stats

-Computer Science

-Intuition

The process and art of machine learning
(At a small scale)

- Acquire data and knowledge about the subject
- Curate & clean the data
- Explore the data
- Make useful transformations of the data
- Split the data in different ways according to what you need to do
    - e.g., Train/test set split, Cross-validation, Nested cross-validation
    - More complexity of model-building, Less data —> more elaborate data splitting
- “DO ML”
- Evaluate and interpret your results


MACHINE LEARNING ENGINEERING
DATA PIPELINE
MACHINE LEARNING PIPELINE
SOFTWARE CODE PIPELINE
EXPLORATION & VALIDATION
· Profiling
· "Junit4Data"
WRANGLING (CLEANING)
DATA
· Data versioning
FEEDBACK
new data from model performance
MONITORING & LOGGING
· Model decay trigger
TRAIN
TEST
MODEL ENGINEERING
code
params
MODEL EVALUATION
code
param
MODEL PACKAGING
MODEL
· Model versioning
· Feature engineering
· Hyperparameters tuning
· Best model selection
· Model performance metrics
- accuracy
- precision
- recall
F1
· Model format
- ONNX
- JAR
- .pkl
· Model serving
- service
- Docker
- K8S
CODE
· Trunk based dev.
· Code versioning
BUILD & INTEGRATION TESTING
GREEN
DEPLOYMENT
DEV
PRODUCTION

Challenges of ML

- Data
  - Insufficient
  - non-representative
  - poor quality
- Testing how good your model is
  - ML is too powerful - overfit!
  - Generalization error: the errors on new data not used in training
- Selecting a good enough/better/best ML
  - “Hyperparameter tuning”
  - “Model selection”
- How do you even measure performance? What’s the metric?

That math
part?

Basic notation

INPUT DATA

We use x (lower case) to denote a feature value (scalar).

The ith input data sample is represented as a vector using bold $\mathbf{x}$:

$\mathbf{x}_i = (x_{i1}, ..., x_{im}) \in \mathbb{R}^m$: A row vector of m elements.

$$\mathbf{x}_i = (22, 1, 0, 160, 180)$$

The entire dataset is represented by a set (the sequence in which each data input $\mathbf{x}_i$ usually doesn't matter.

$S = \{\mathbf{x}_i, i = 1..n\}$: A set S with n samples. i goes from 1 to n.

Or we can write it as a matrix, when we need to do some linear algebra :)

Basic notation

PREDICTION

We use y (lower case) to denote a binary classification.

$y = -1$ (or sometimes we use $y = 0$) is referred to as the negative class.
$y = +1$ is referred to as the positive class.

Given a data sample $\mathbf{x}_i = (x_{i1}, ..., x_{im})$,

we want to predict $y_i = -1 \ or \ +1$?

OR… y is just a real number we want to predict

Given a data sample $\mathbf{x}_i = (x_{i1}, ..., x_{im})$,

we want to predict $y_i \in \mathbb{R}$?

Basic notation

MODEL PARAMETERS

Model: $\mathbf{w} = (w_1, ..., w_m) \in \mathbb{R}^m$ (in the same dimension of input $\mathbf{x}$)

bias: $b \in \mathbb{R}$ (scalar)

Data sample $\mathbf{x} = (x_1, ..., x_m) \in \mathbb{R}^m$,

$$\mathbf{w} \cdot \mathbf{x} + b \qquad (w_1, w_2, .., w_m) \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_m \end{pmatrix} + b$$

"·" refers to as the dot product between two vectors

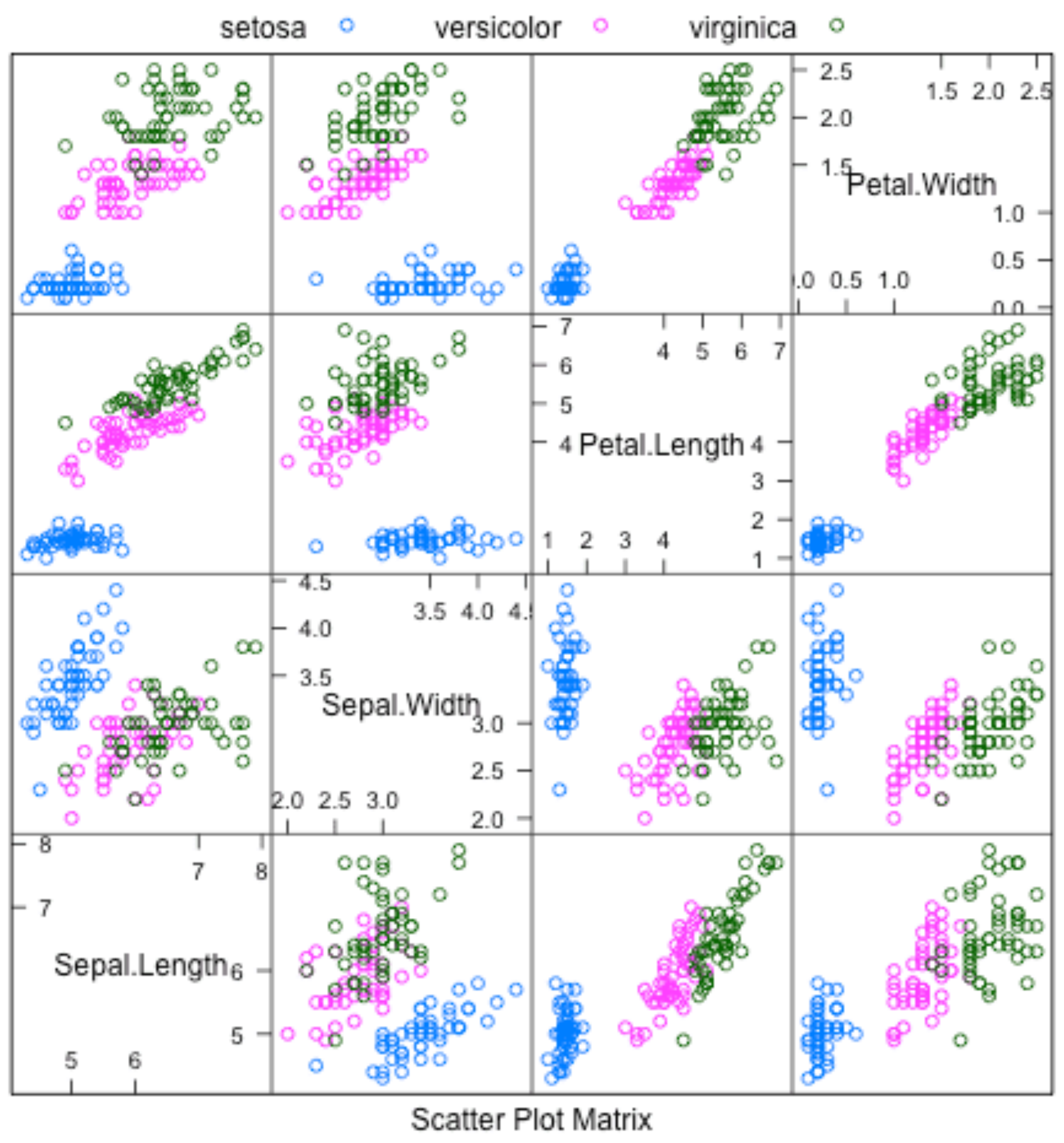
Alternative notation 1: $< \mathbf{w}, \mathbf{x} > +b$

Alternative notation 2: $\mathbf{w}\mathbf{x}^T + b$ ($\mathbf{w}$ and $\mathbf{x}$ are row vectors).

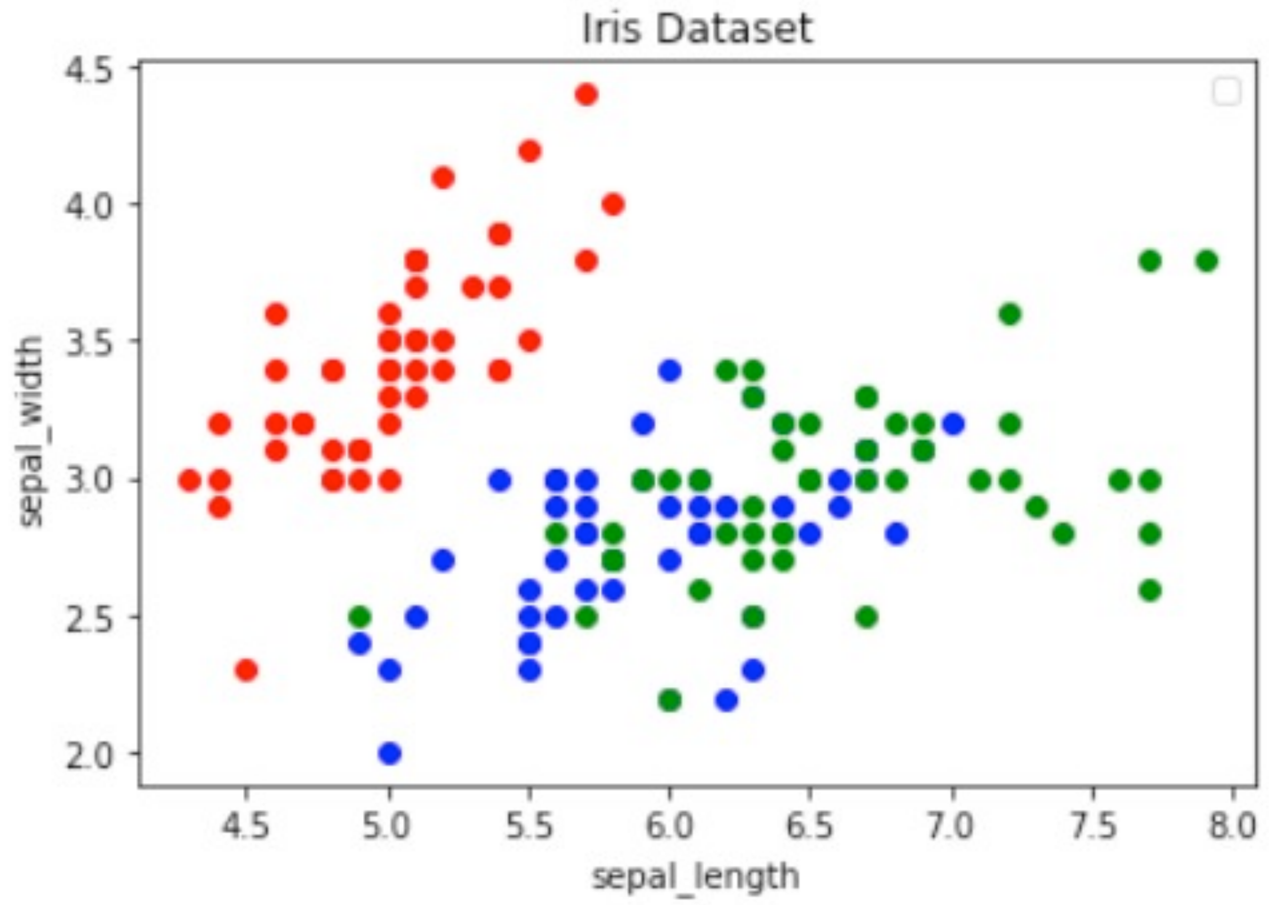
$\mathbf{w}^T\mathbf{x} + b$ ($\mathbf{w}$ and $\mathbf{x}$ are column vectors).

That coding part?

```r
#This is the function with the new shorthand
x <- 3
y <- 4
adding <- \(x,y){
  z<-x+y
  print(z)
}
adding(x,y)
```

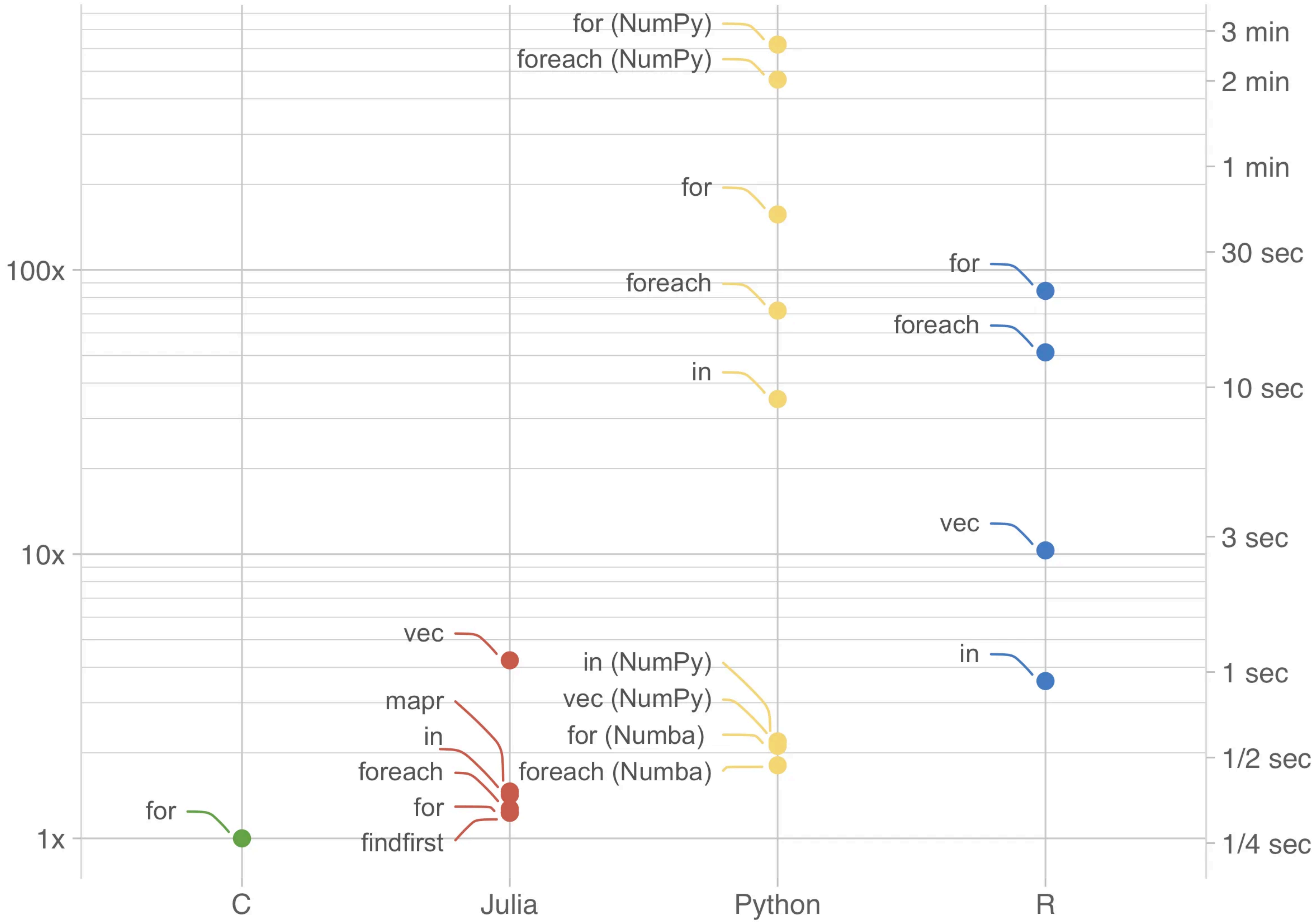



```python
# This is single line comment
a = 5
"""
-- This is multiple line comment --
   Asking user to enter a number that
   will be stored in variable b """
b = int(input("Enter a number b/w 0-10 : " ))

# Comparing two numbers to see which is bigger
if (a > b):
    print(" a is greater than b")
elif (a < b):
    print("b is greater than a")
else:
    print("a and b are same numbers")

#printing variable a and b
print("a:", a, "b:", b)
```

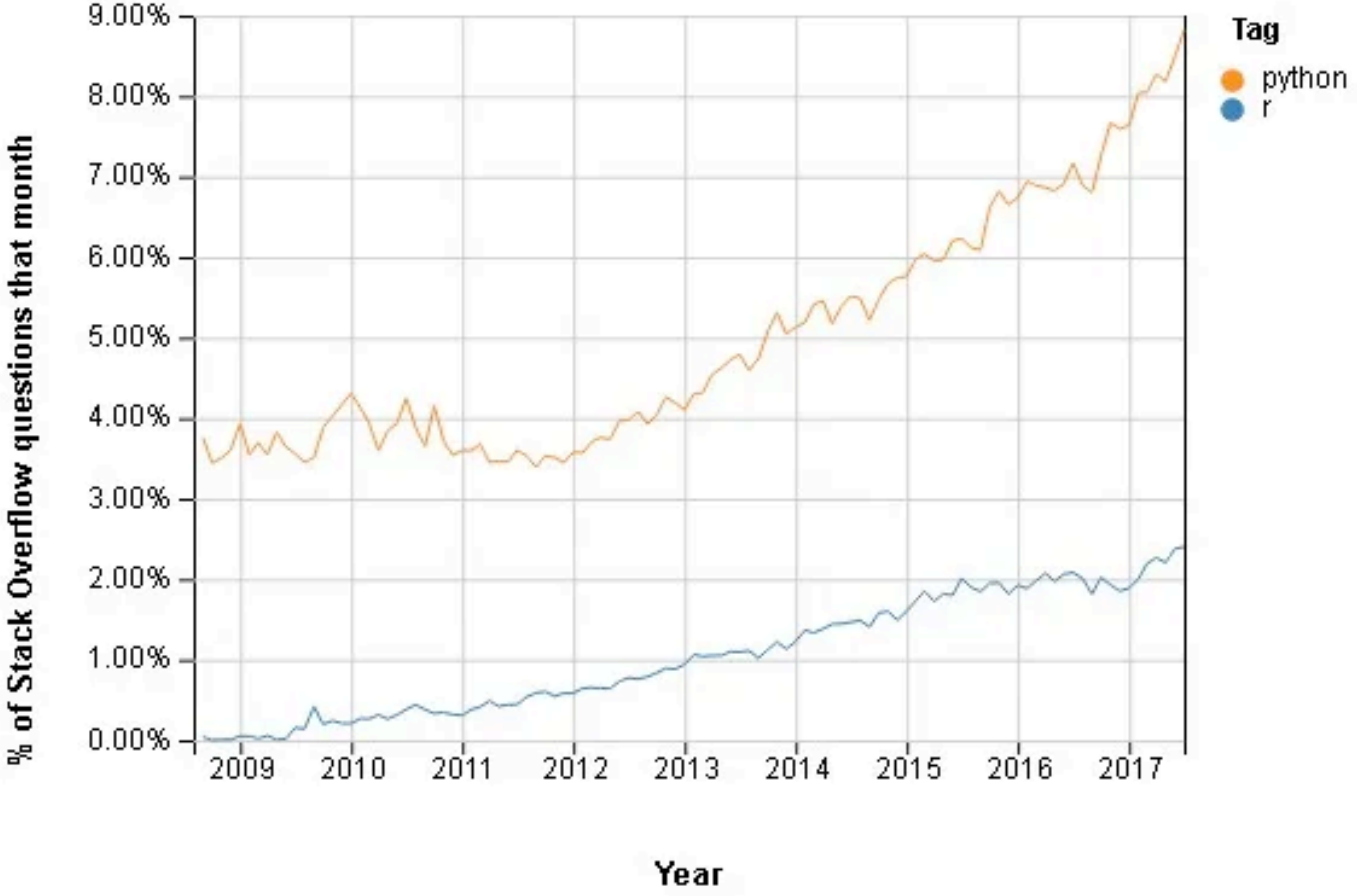


Sourced from:
https://warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/iris_lm/
https://www.analyticsvidhya.com/blog/2021/06/guide-to-data-visualization-with-python-part-1/
https://medium.com/tech-lounge/getting-started-with-python-21b6c3a324fd
https://medium.com/analytics-vidhya/how-to-use-the-new-function-symbol-in-r-programming-4-1-r-4-1-c7cd2ac27b98

CPU time (relative to C and absolute)



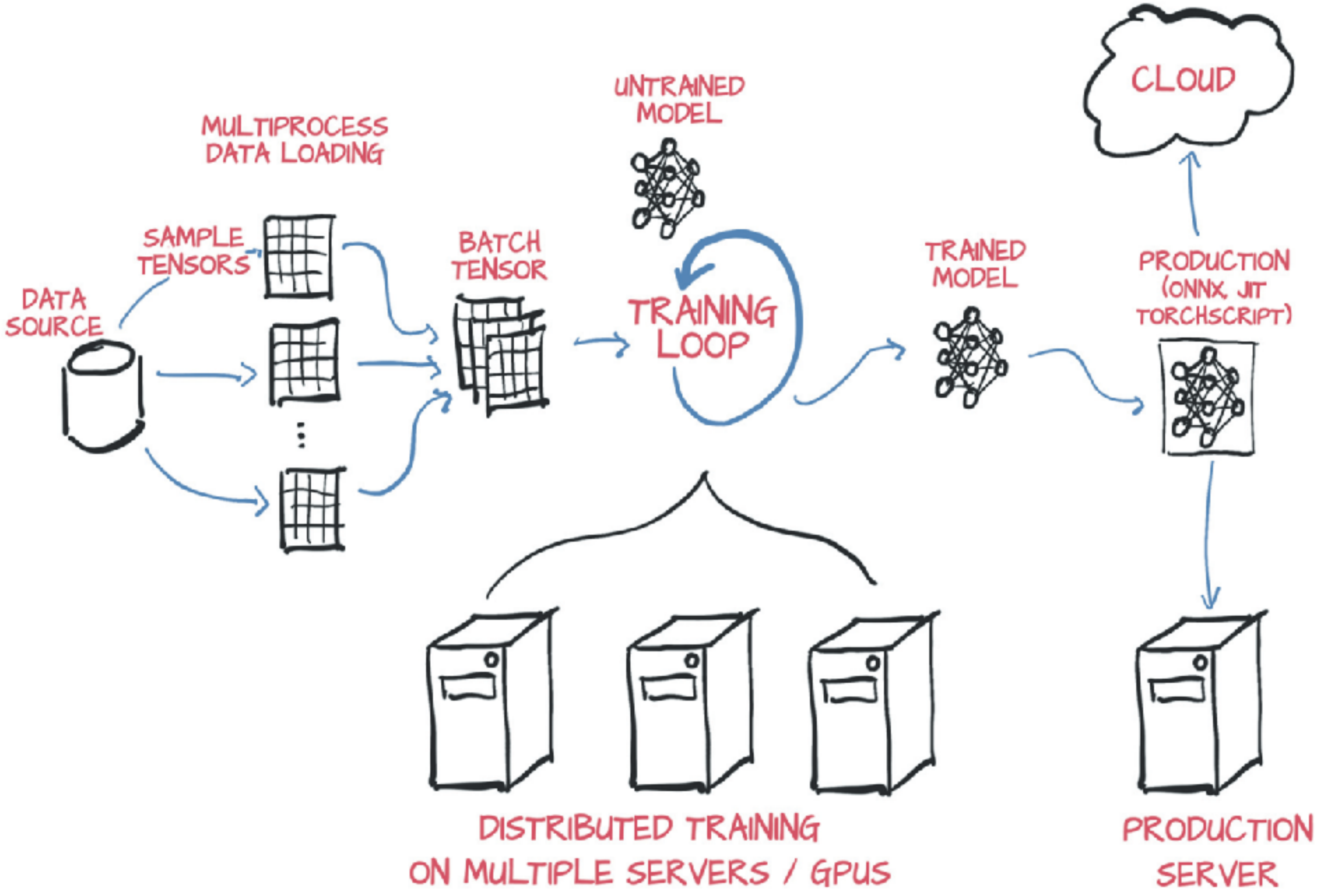
Sourced from:
https://towardsdatascience.com/r-vs-python-vs-julia-90456a2bcbab



Sourced from:
https://medium.com/@datadrivenscience/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197

DL popularity: from toolkits that allow people to quickly and easily train huge models

Google Cloud | PyTorch

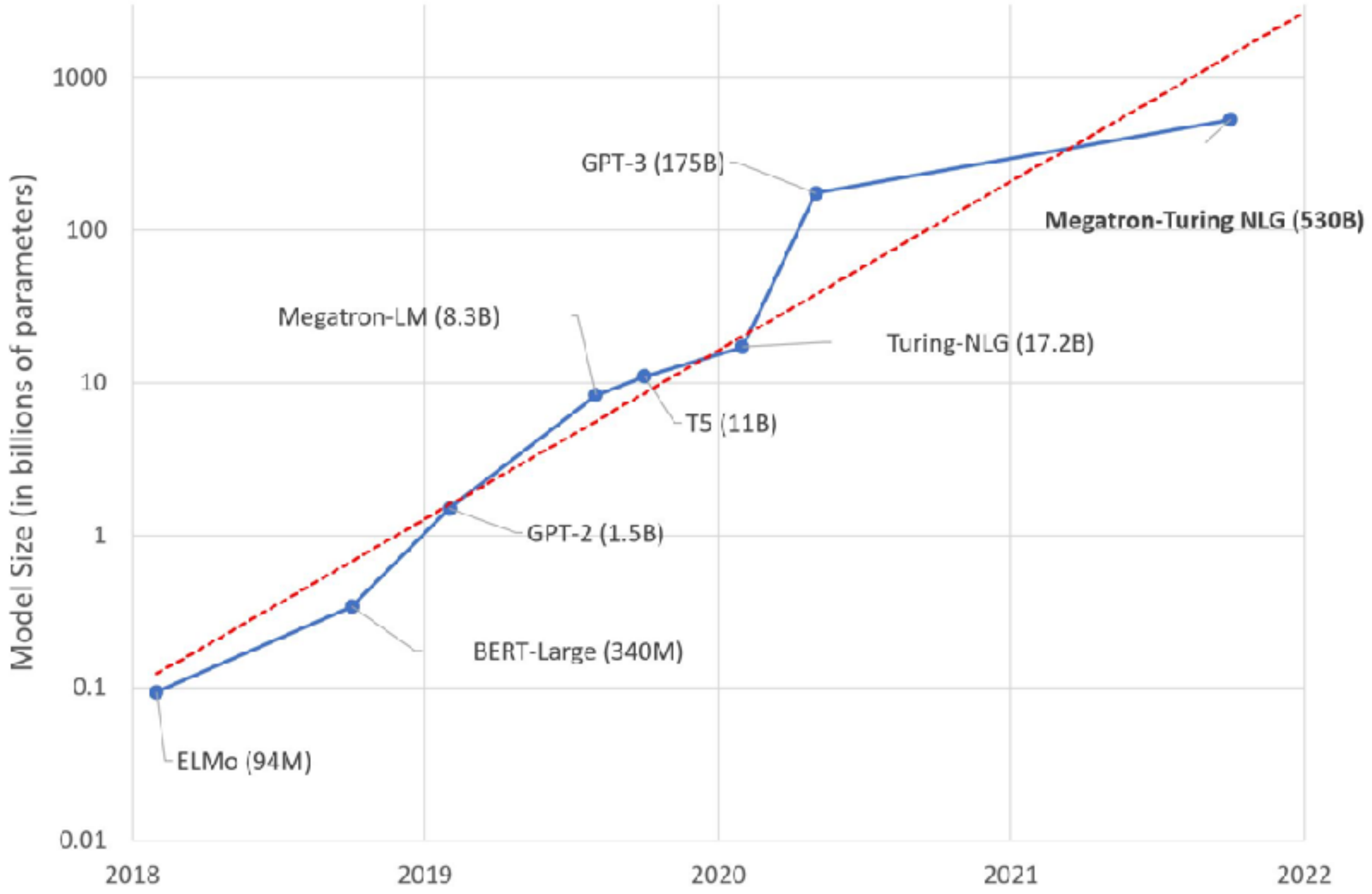


Deep-Learning U-Turn

DL isn't like other ML

- What is "deep", why is it different?
  - Huge datasets + networks take days or weeks to train on $\sim 10^2 - 10^3$ GPUs
- Leads to new mindsets
  - "The unreasonable effectiveness of DL"
  - "The data is the model"… better data > model tweaking
  - Dataset -> Population. Don't need fancy stats or resampling
  - Manipulating training procedure to prevent overfitting

	GPT-2	GPT-3	ChatGPT
Training set size in GB	40	570	"13 million active daily users" implies >100GB per day at 5 pages/user

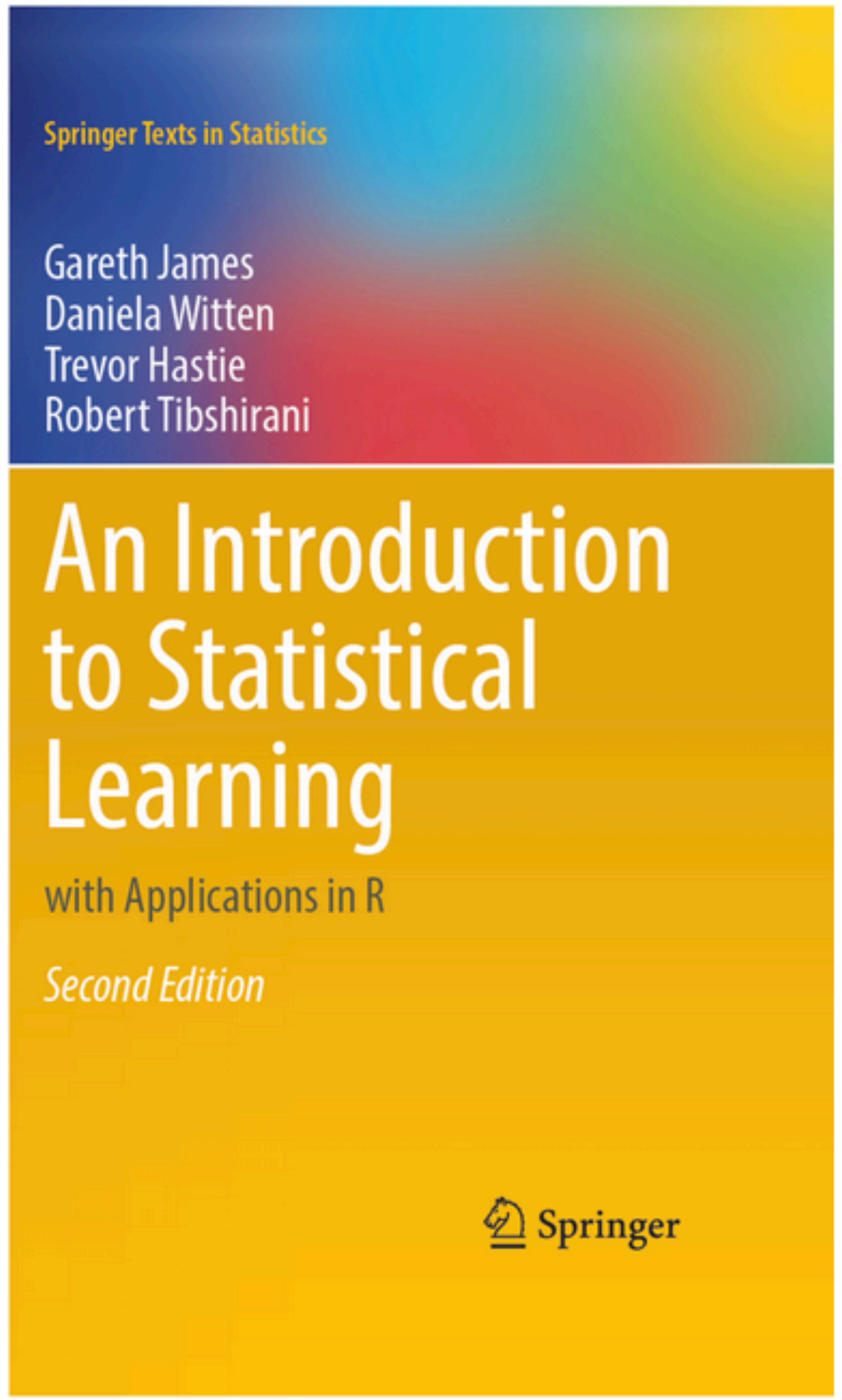


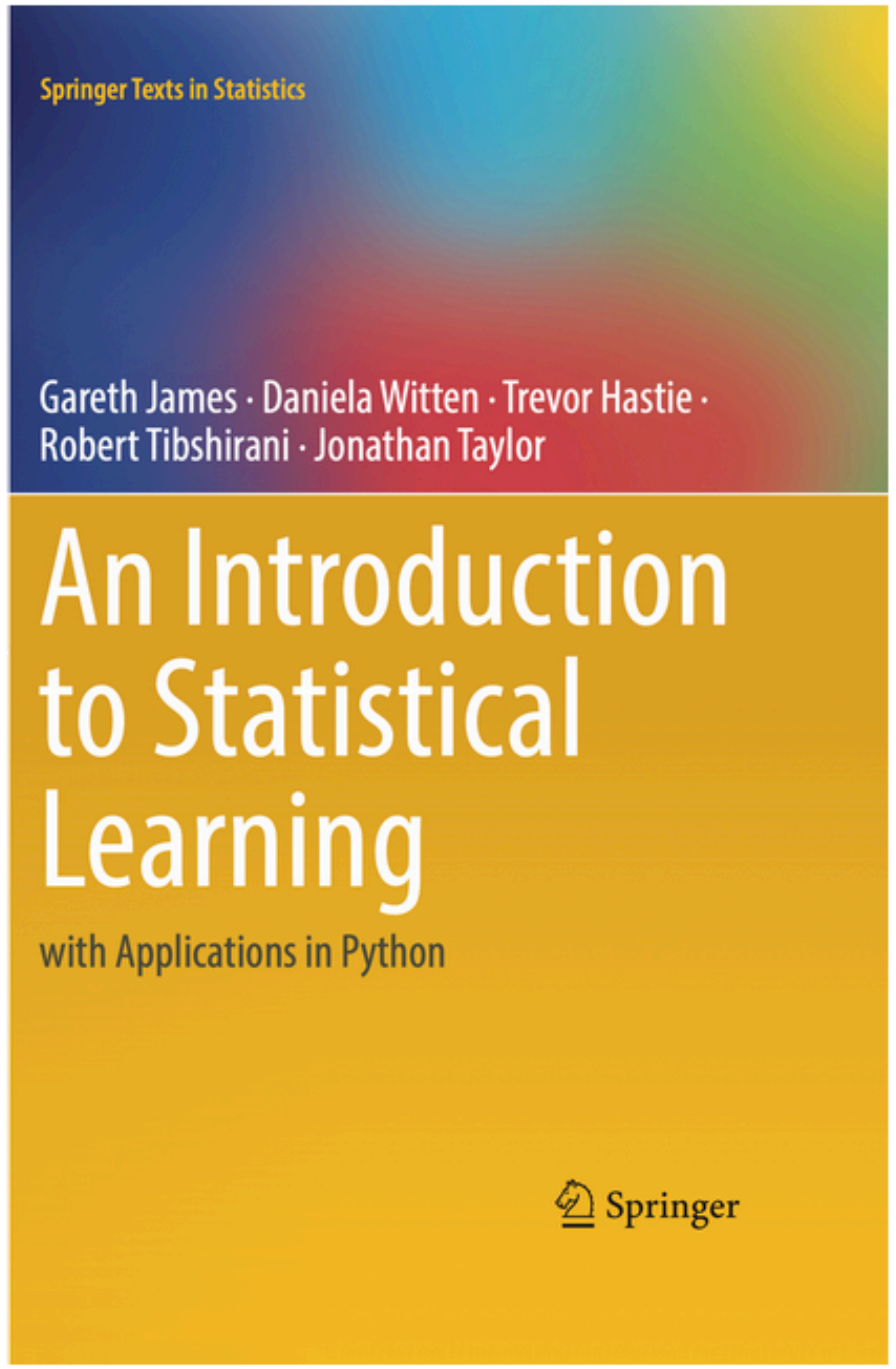
PaLM

LLaMa

ChatGPT

Where do I start?


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer


Springer Texts in Statistics
Gareth James · Daniela Witten · Trevor Hastie ·
Robert Tibshirani · Jonathan Taylor
An Introduction to Statistical Learning
with Applications in Python
Springer

University of Chicago
Machine Learning

MIT Open Learning

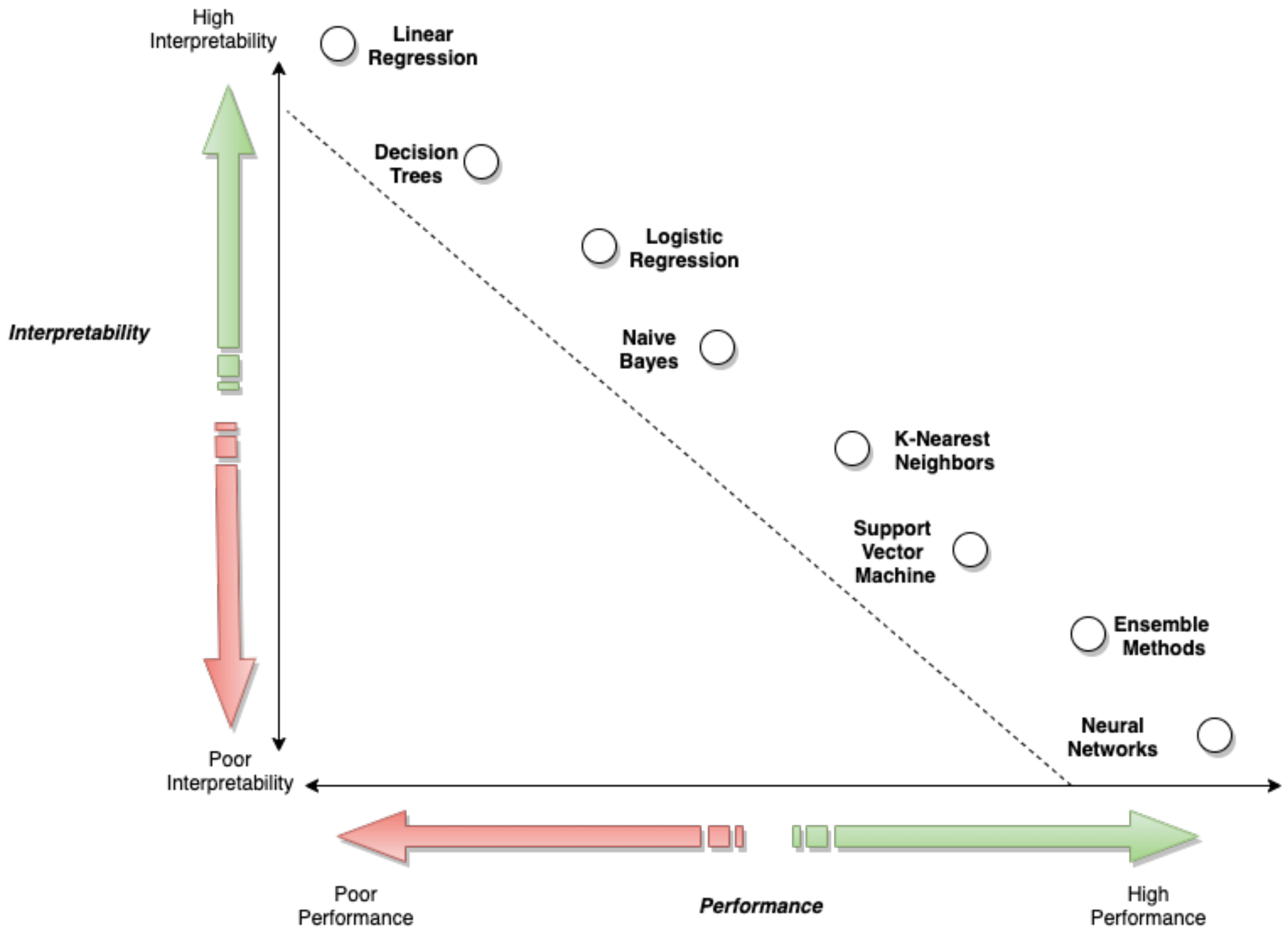
ML@B

Cornell University

arXiv.org

ML
MACHINE LEARNING
DEPARTMENT

HDSI
UC SAN DIEGO


High
Interpretability
Linear
Regression
Decision
Trees
Logistic
Regression
Interpretability
Naive
Bayes
K-Nearest
Neighbors
Support
Vector
Machine
Ensemble
Methods
Neural
Networks
Poor
Interpretability
Poor
Performance
Performance
High
Performance

Thank you!
+ Questions