EXPLORING AI

A LITTLE ABOUT ME!

EDUCATION

PhD Electrical and Computer Engineering (2023 - Present)
Research focused on new Attention mechanisms for speech and language processing

MS Statistics - Analytics (2021 - 2023)

Research focused on Deep Learning architectures for geospatial and climate applications

B.S. Statistics and Neuroscience (2017 - 2020)

Focus on theoretical statistics and machine learning for computational neuroscience

RESEARCH/INDUSTRY

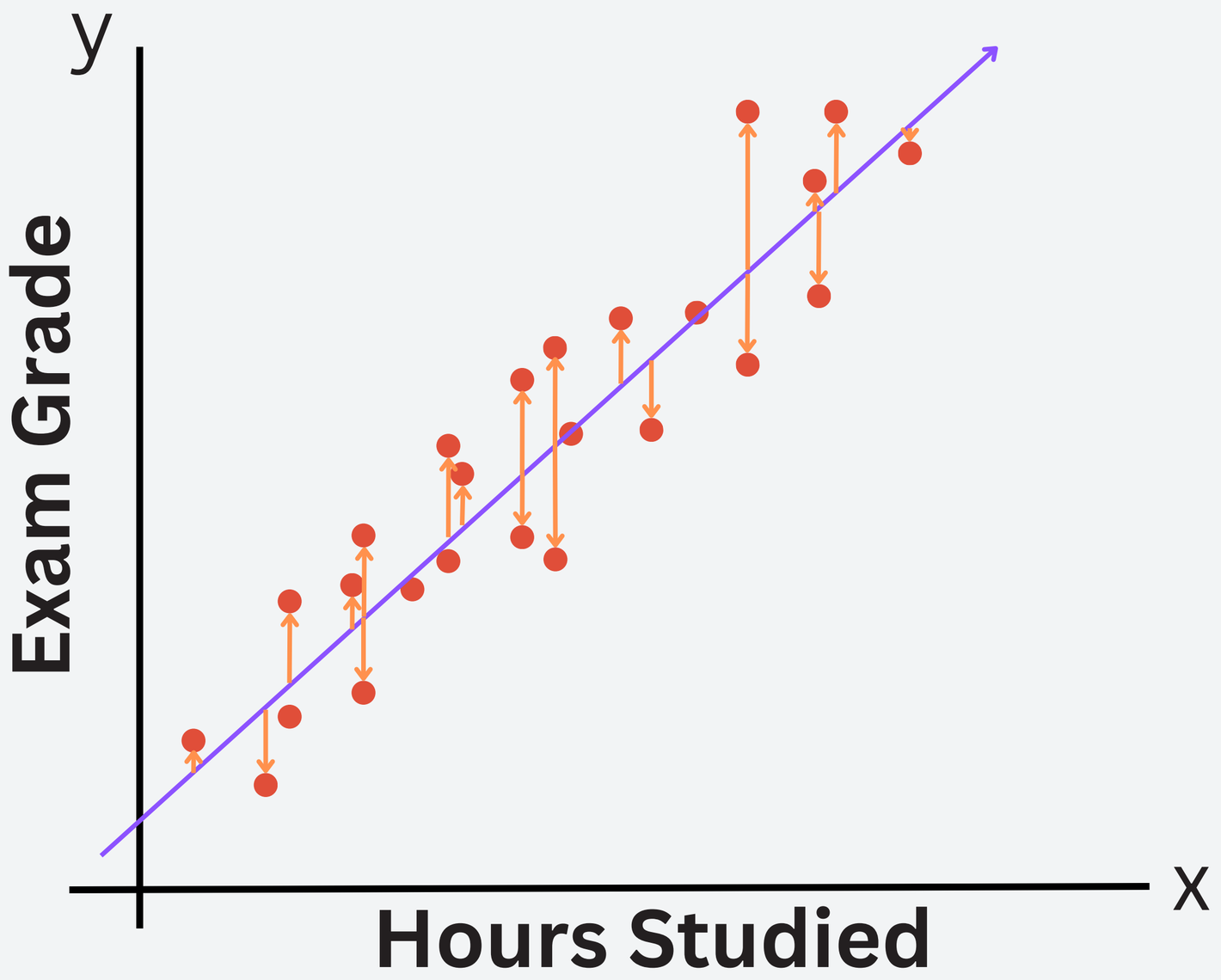
Sandia National Laboratories

NCSA

SILICON HILLS TECH

nano™

LINEAR REGRESSION



Model: $y = mx + b$

Input: x

Output: y

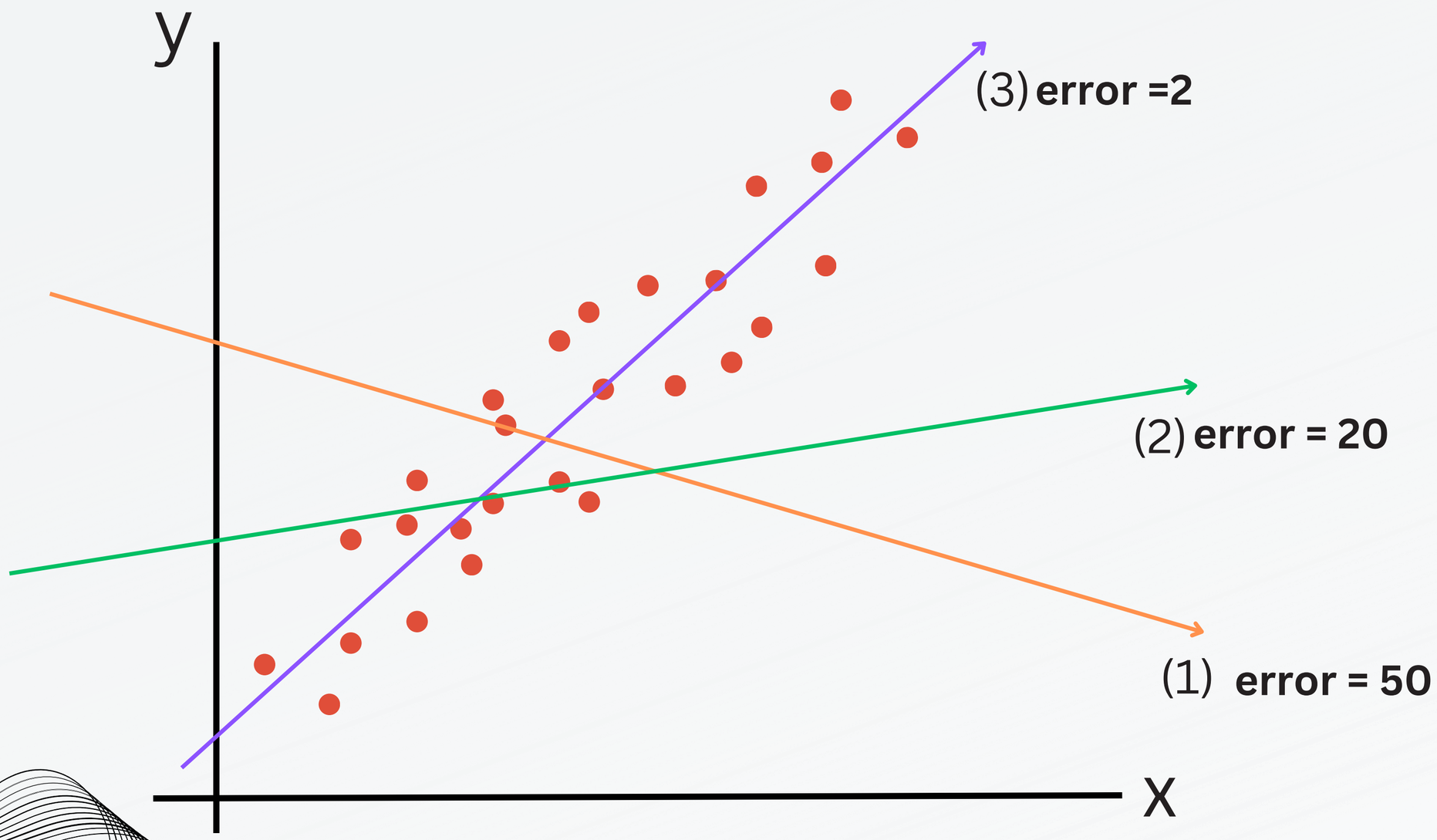
Loss: $\sum (y_{true} - y_{pred})^2$

Parameters: m, b

m (slope) - Grade increase for every hour studied

b (intercept) - If I study 0 hours what grade will I get?

OPTIMIZATION

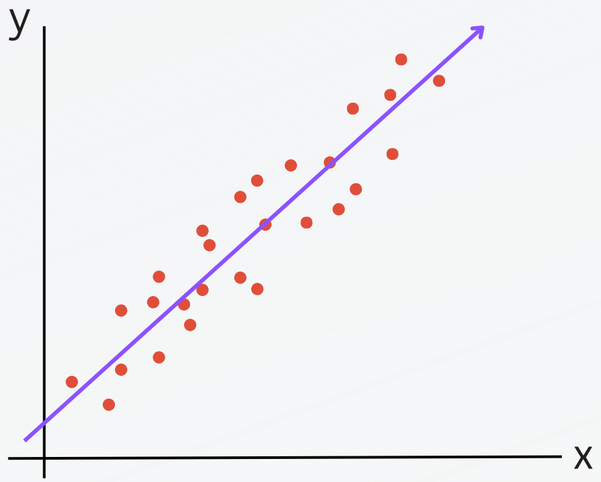


We dont know
the best Slope **(m)** and
Intercept **(b)** so the best
we can do is
start with random values
and take
little steps in the
right direction!

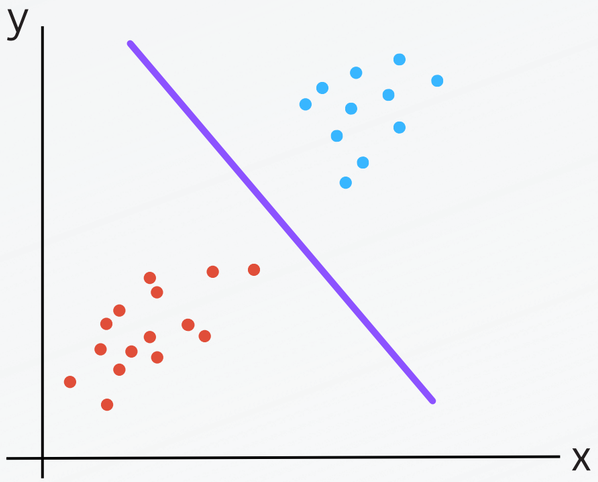
TYPES OF MACHINE LEARNING

Supervised

Training models when each input is paired with a known output



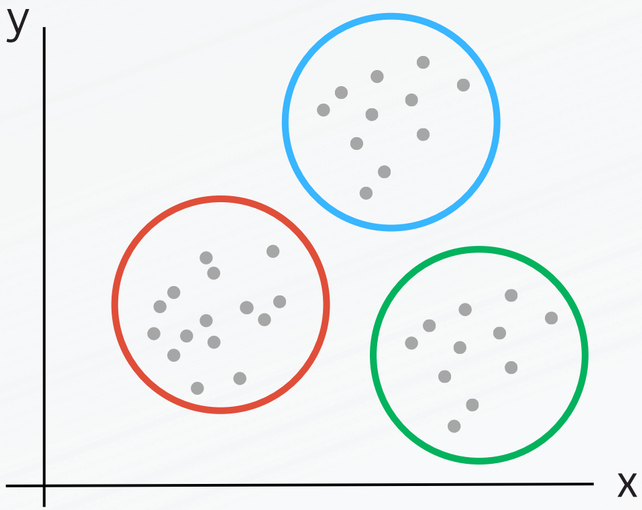
Regression



Classification

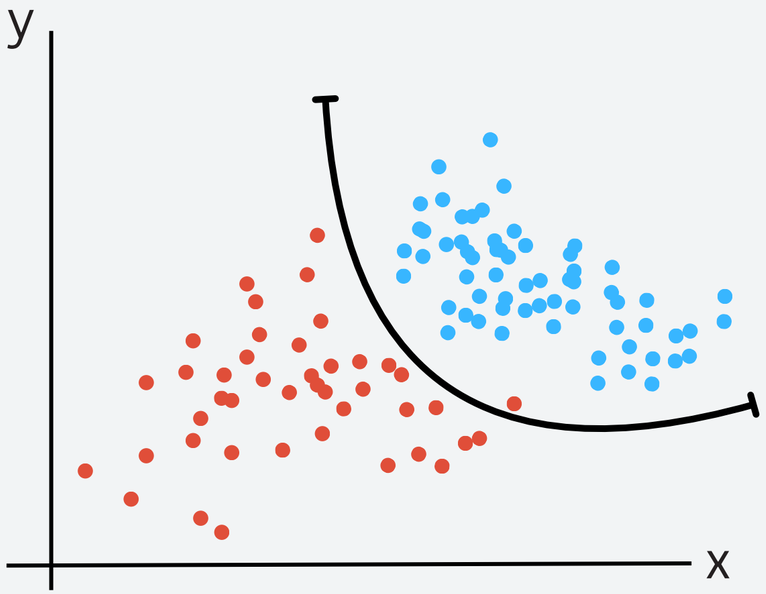
Unsupervised

Training models when only input is known with no known output

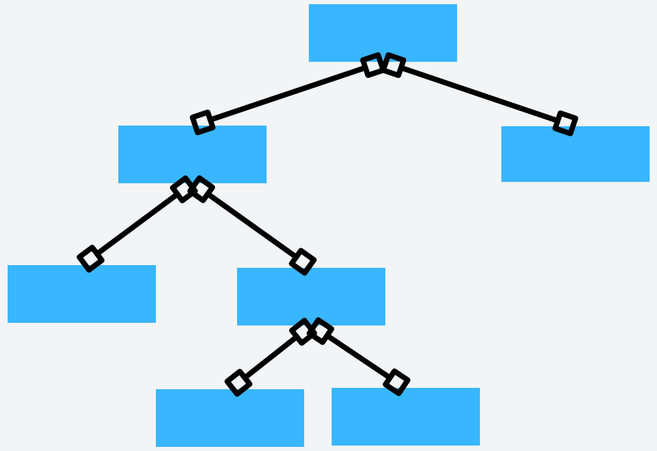


Clustering

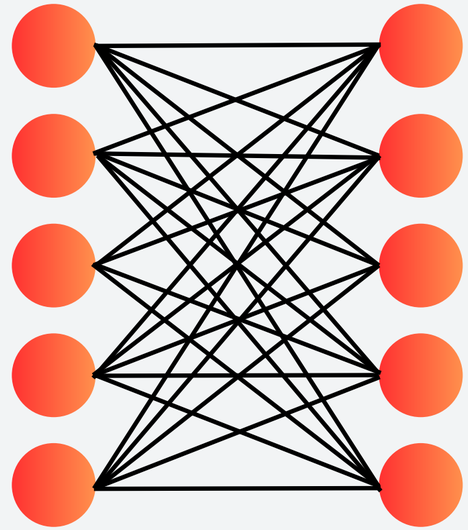
SUPERVISED LEARNING



Support Vector Machine

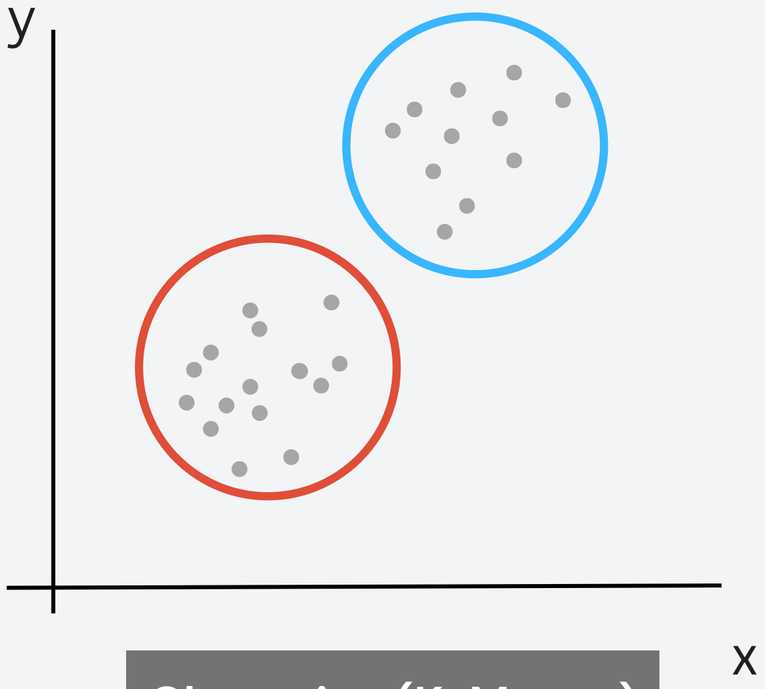


Decision Trees

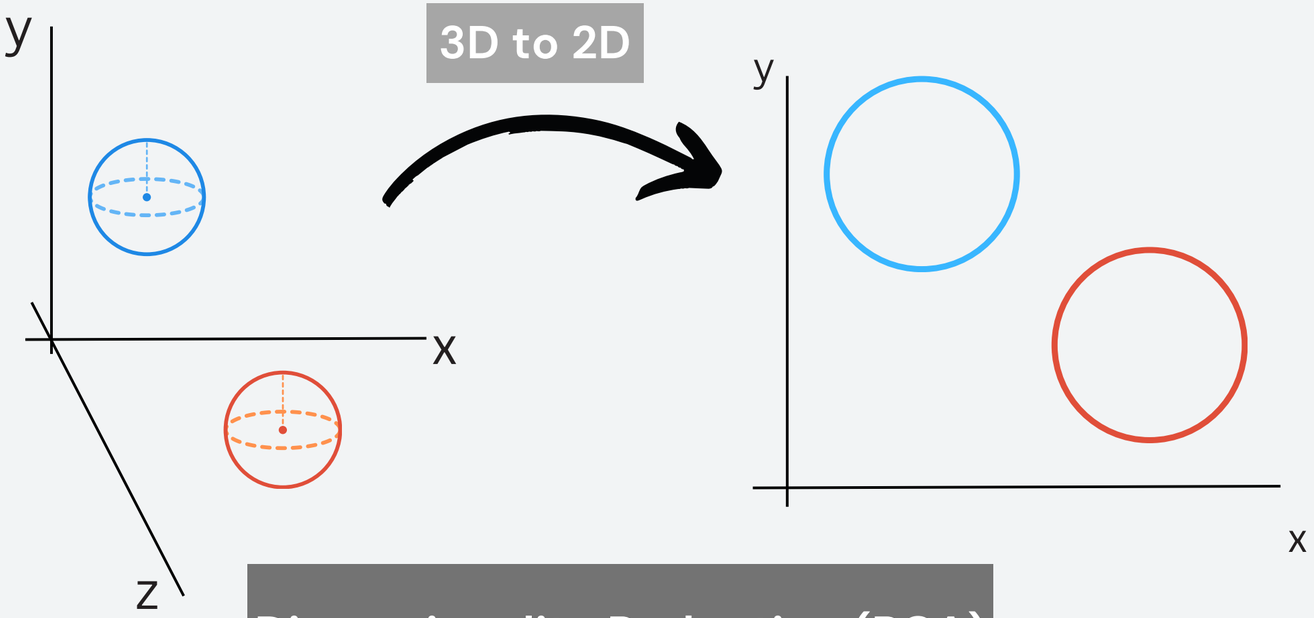


Neural Nets

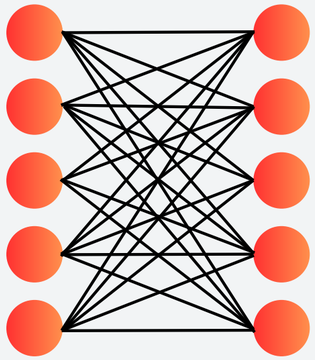
UNSUPERVISED LEARNING



Clustering (K-Means)



Dimensionality Reduction (PCA)



Neural Nets

NEURAL NETWORKS ARE A GENERAL PURPOSE LEARNER

Images

Language

Audio

WHEN TO USE NEURAL NETWORKS?

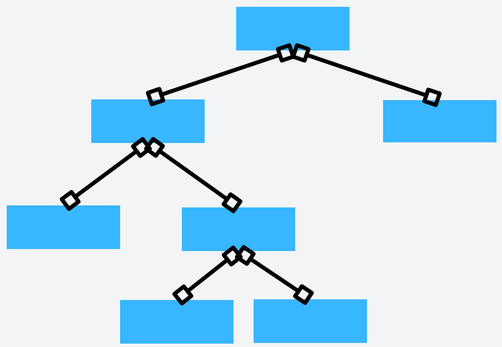
- Large Datasets
- Non-Tabular
- High Dimensional

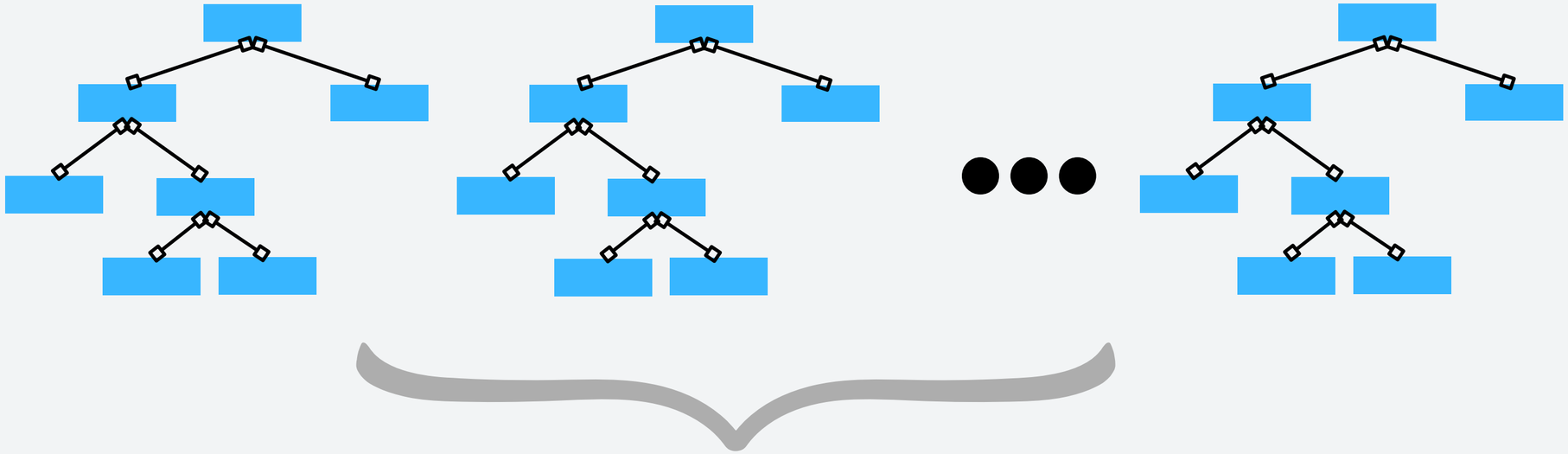
WHEN IS TRADITIONAL ML IS BETTER?

- Explainability
- Limited Compute
- Robust

COMPLEXITY RUINS EXPLAINABILITY

DECISION TREES ARE EXPLAINABLE... BUT RANDOM FORESTS ARE NOT





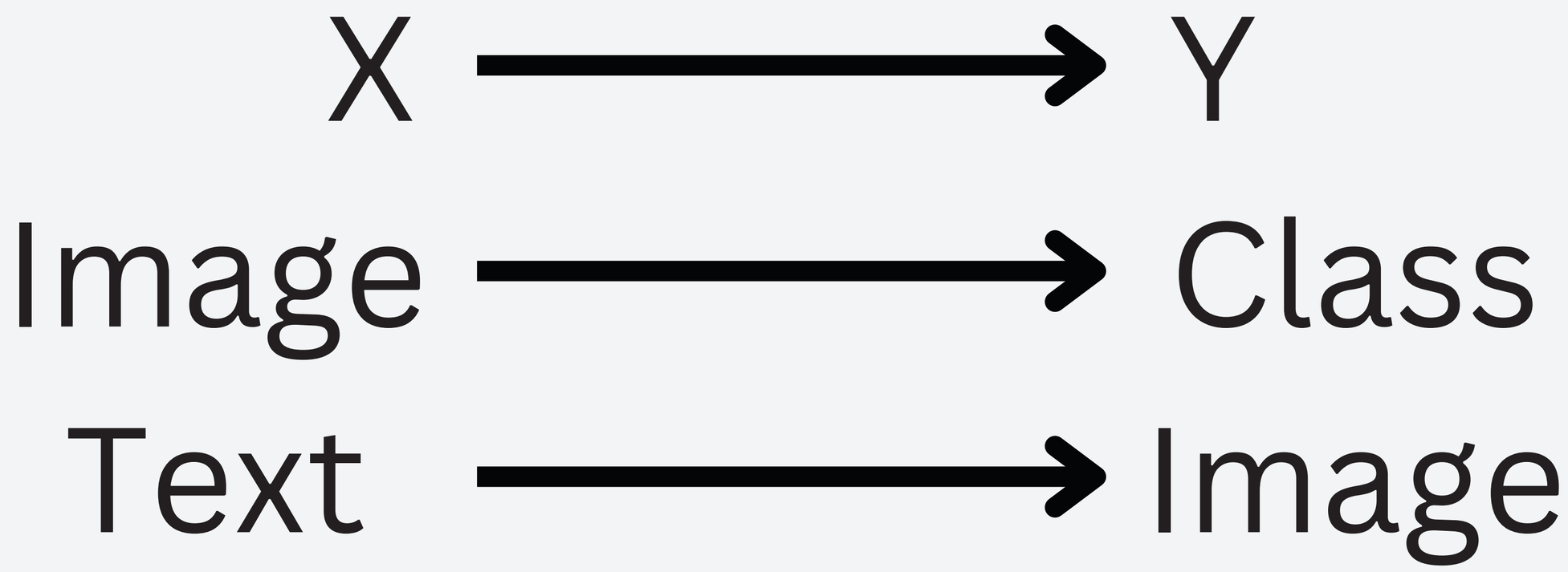
Weighted Decision

Decision Trees

Random Forest

NEURAL NETWORKS: UNIVERSAL FUNCTION APPROXIMATOR

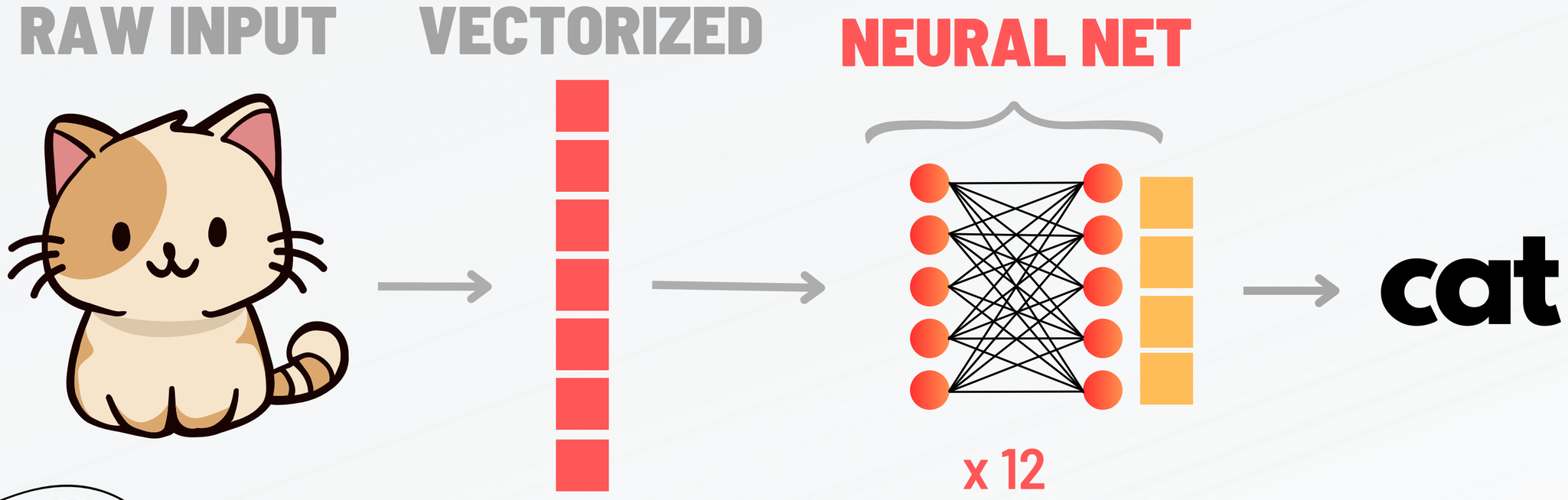
As a Neural Network Gets Larger, it can approximate any arbitrary function!



ANYTHING APPROXIMATOR... BUT AT WHAT COST?

The penalty we pay for powerful models is the lack of explainability... we cannot assign reasons typically to why a decision was made by a model

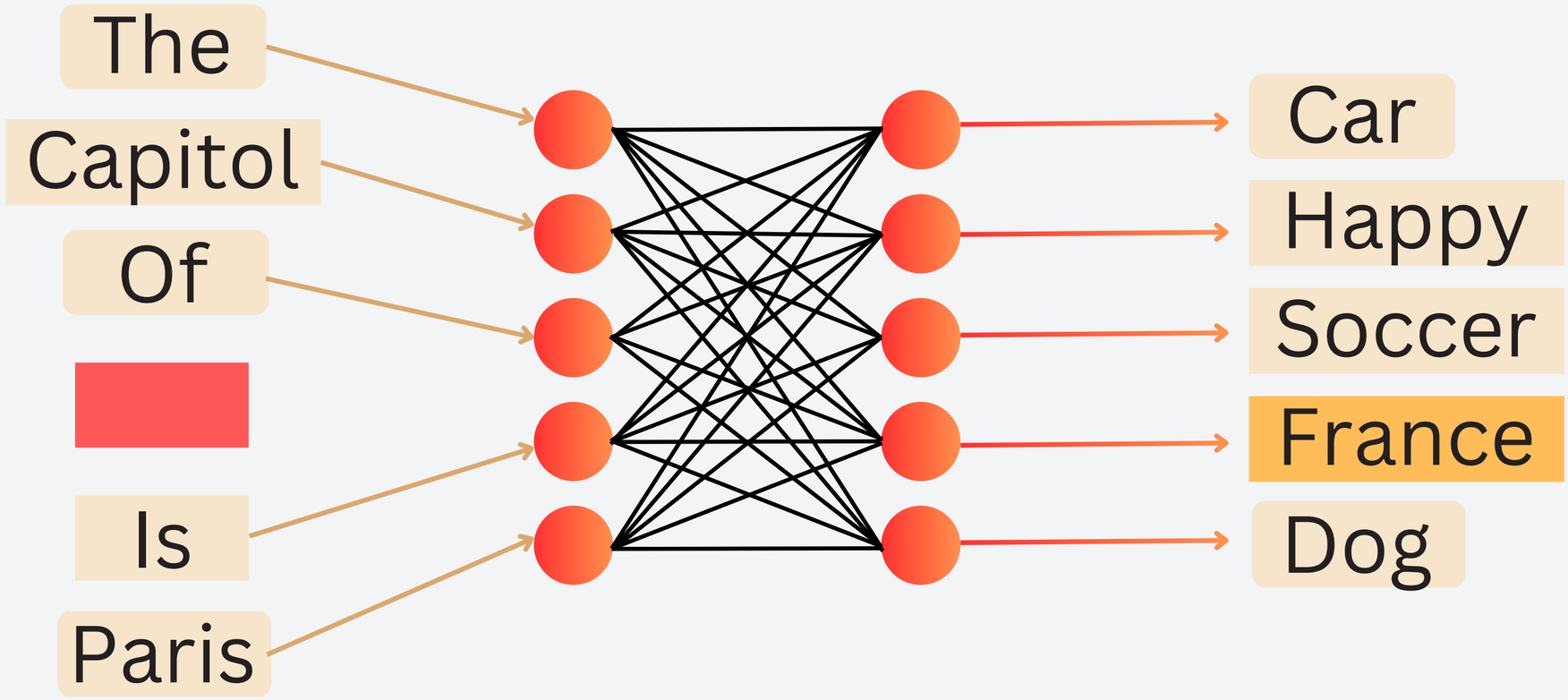
Knowing which neurons fire in a brain in exactly which order cannot completely explain what I am thinking about, it is kind of the same issue with neural nets!



HOW NEURAL NETS LEARN?

REPRESENTATION THROUGH EMBEDDINGS...

SKIP-GRAM MODEL FOR LANGUAGE



WHAT ARE EMBEDDINGS?

France

Of

Happy

Is

Paris

Car

The

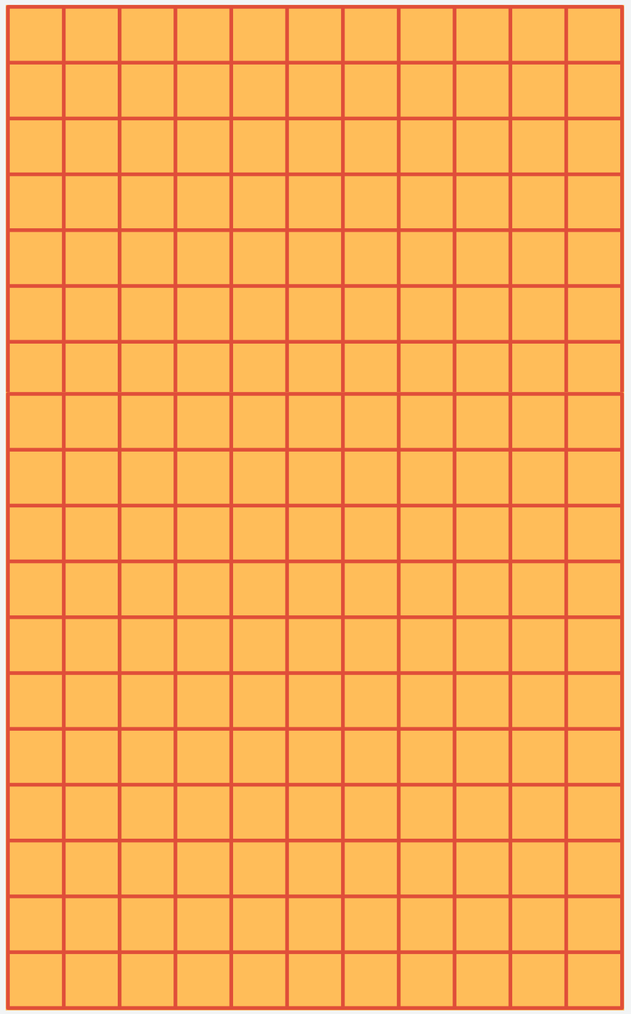
⋮

Soccer

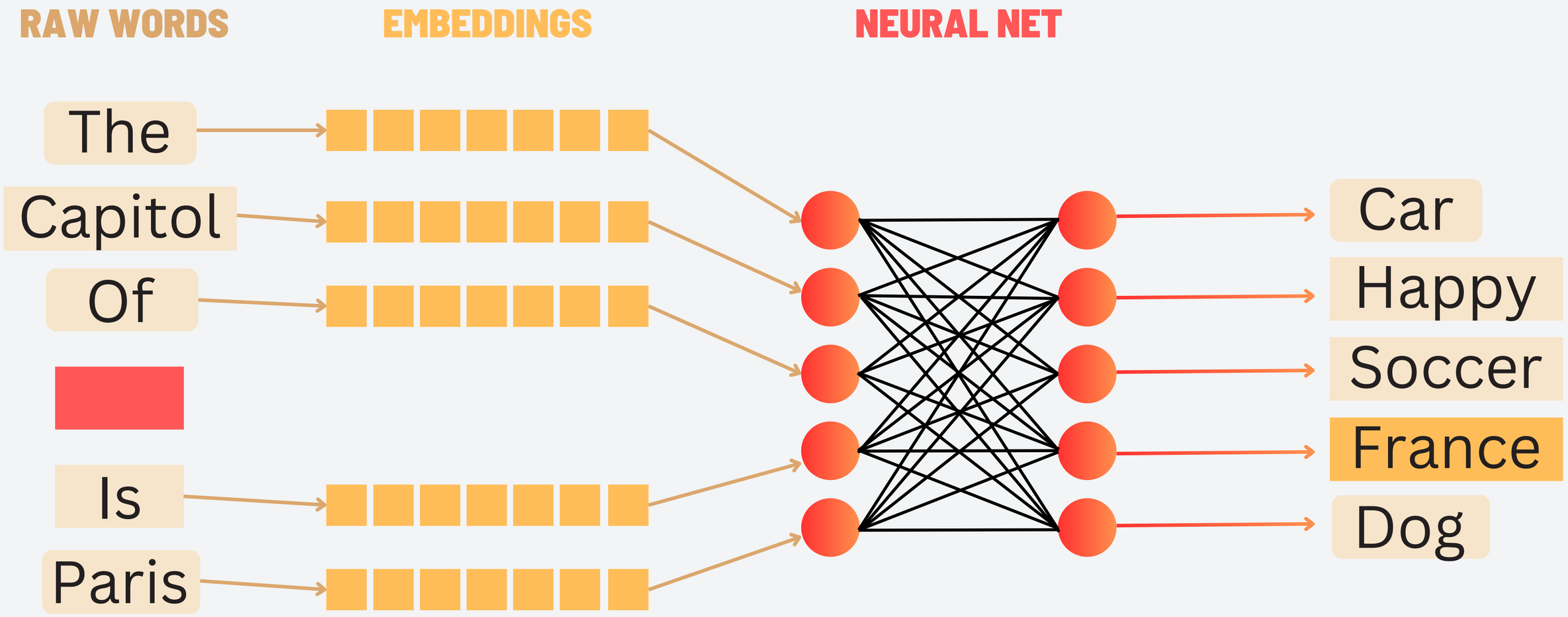
Capitol

OF UNIQUE WORDS

EMBEDDING SIZE

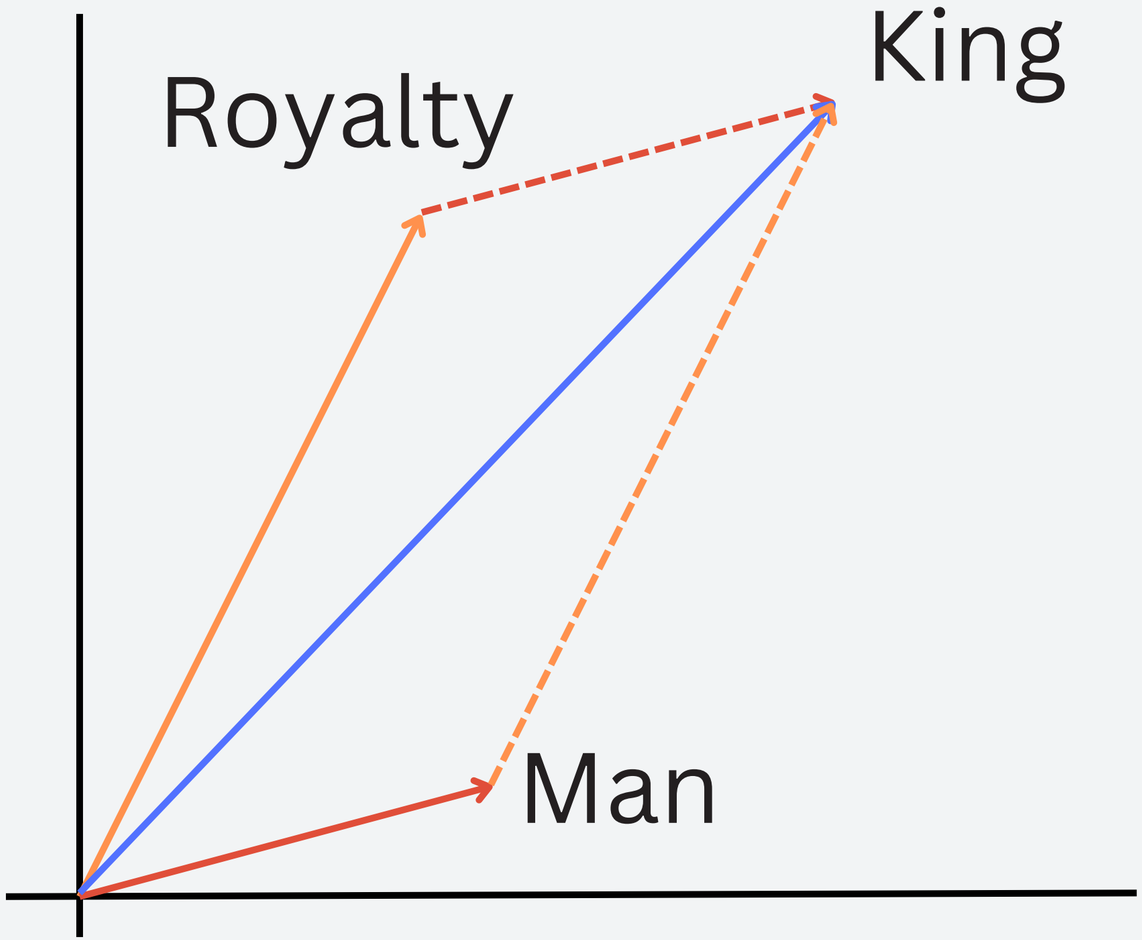


SKIP-GRAM MODEL FOR LANGUAGE



EMBEDDING ARITHMETIC

SOMETIMES EMBEDDINGS ARE INTERPRETABLE

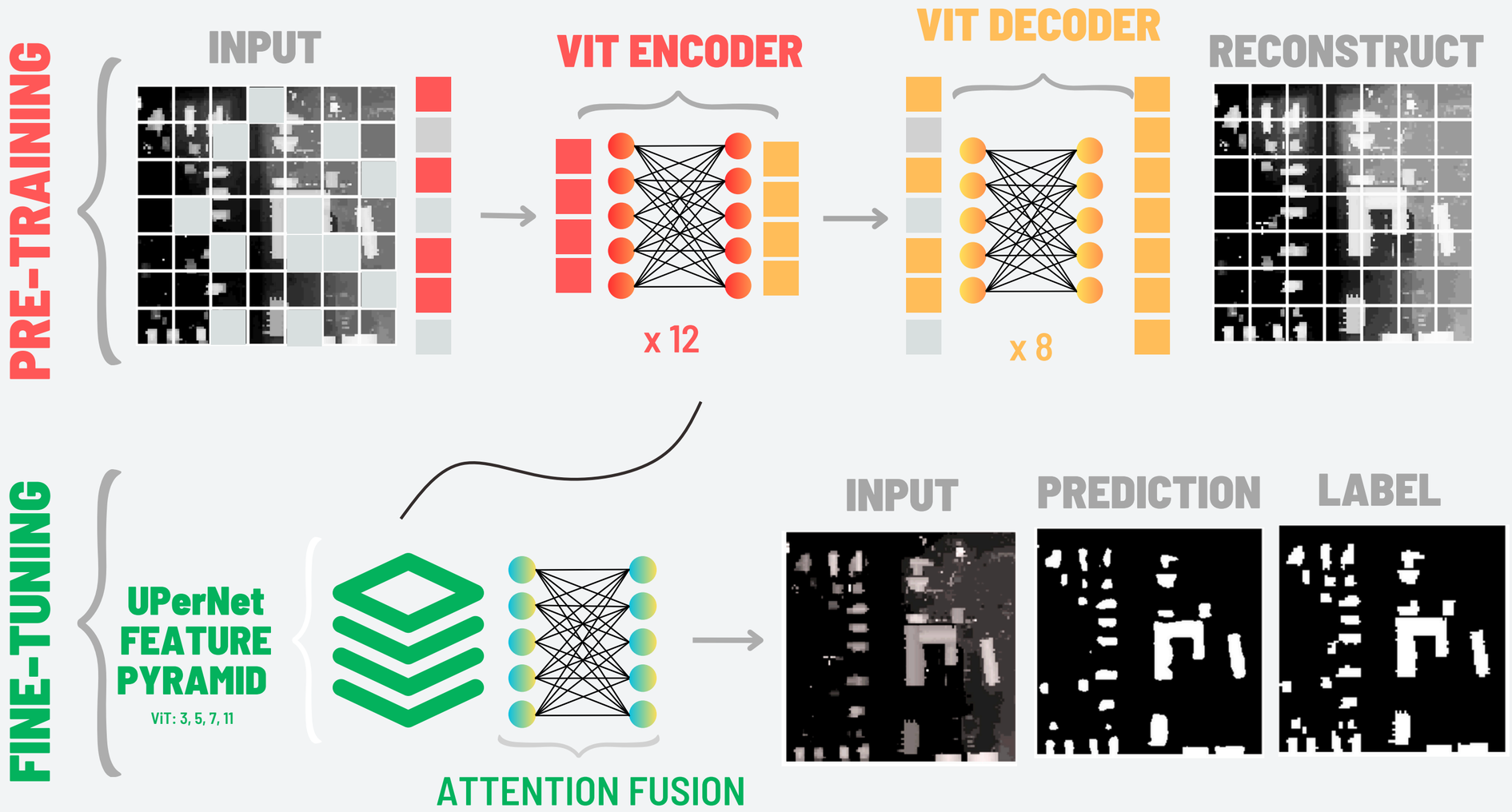


Arithmetic

Man + Royalty = King

King - Man + Woman = Queen

STATE OF THE ART MODELS ARE (MOSTLY) UNINTERPRETABLE -> SEMANTIC SEGMENTATION



STATE OF THE ART NEEDS LOTS OF COMPUTE AND DATA

STABLE DIFFUSION

~1 BILLION PARAMS

256 X A100 GPU

$600,000

5 BILLION IMAGES

GPT4

~175 BILLION PARAMS

25000 X A100 GPU

$100 MILLION

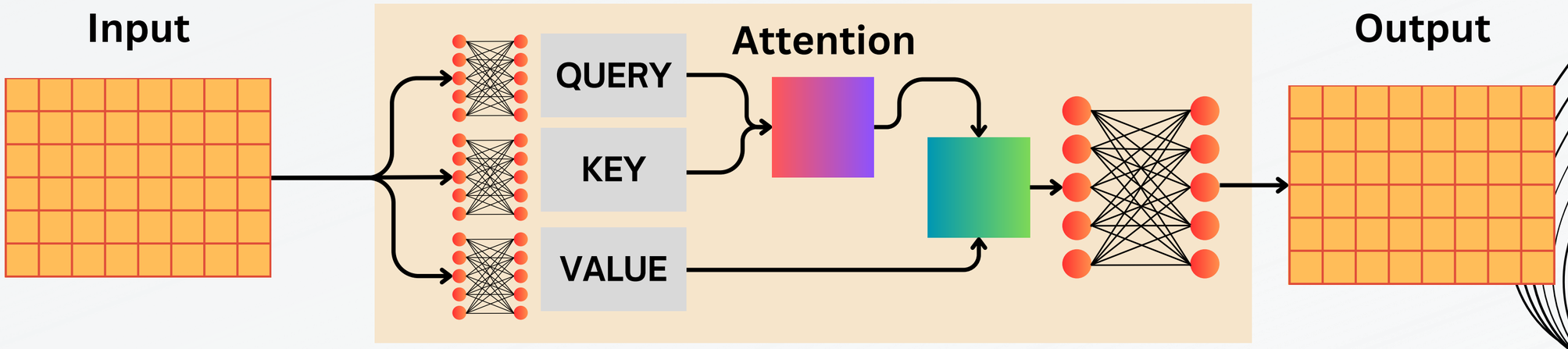
13 TRILLION TOKENS

WHISPER

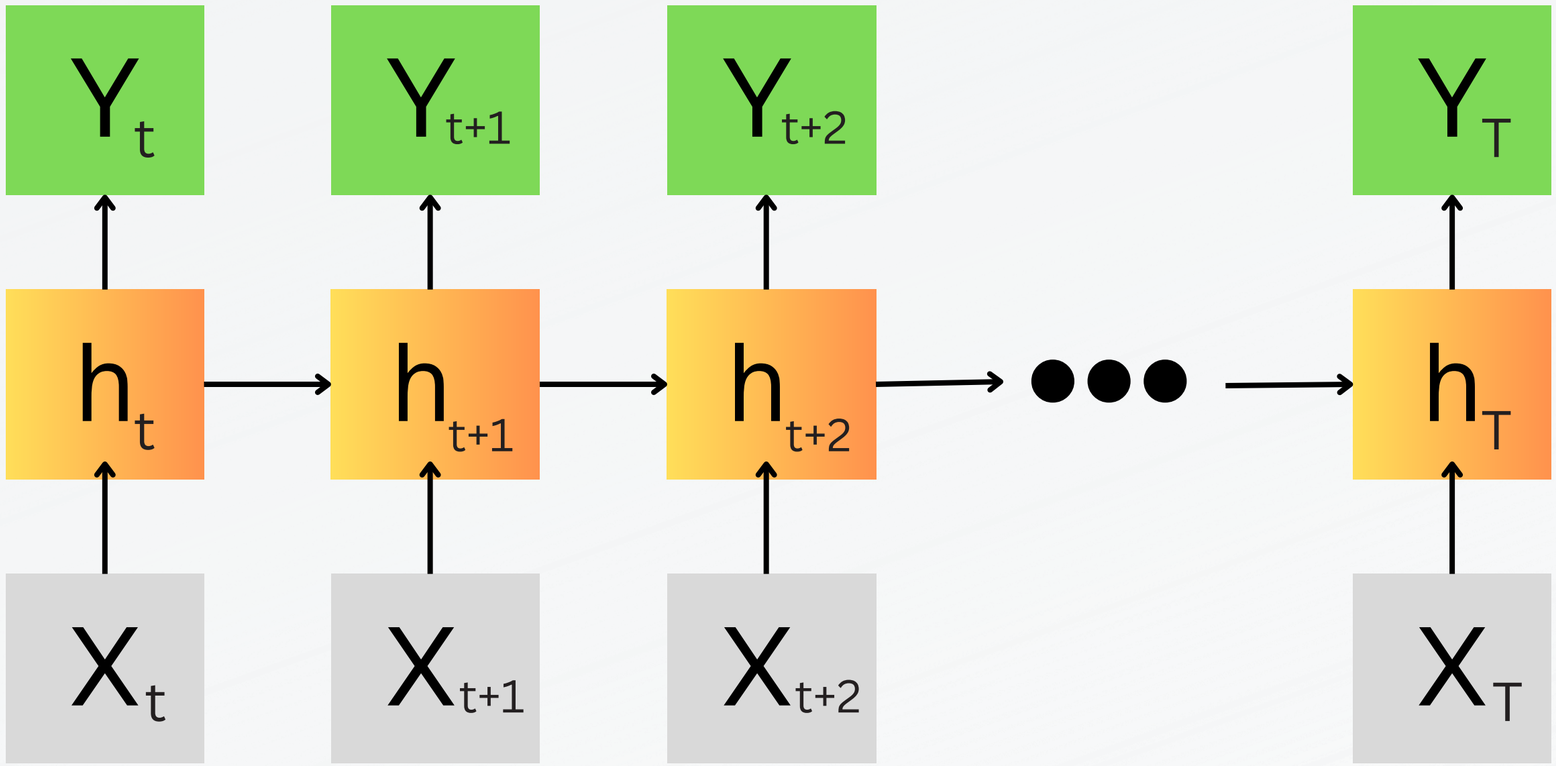
~1.5 BILLION PARAMS

680,000 HOURS OF AUDIO

STATE OF THE ART IS POWERED BY THE TRANSFORMER



PROBLEM WITH RECURRENCE



WHERE DO YOU SEE LONG SEQUENCES?

TEXT

To be or not to be—that is the question, Whether 'tis nobler in the mind to suffer The slings and arrows of outrageous fortune, Or to take arms against a sea of troubles And, by opposing, end them. To die, to sleep— No more—and by a sleep to say we end The heartache and the thousand natural shocks That flesh is heir to—'tis a consummation Devoutly to be wished. To die, to sleep— To sleep, perchance to dream. Ay, there's the rub, For in that sleep of death what dreams may come,

GENES

cgtgcgaatcttctatcaacgattagttgttctgaggatacgcagcacaggcatgaaatacaagatcg aatgtgcccagagttcaatgcacacagtatatctatggcccacatccccccactggtgtaccgattac ttcacccttcggaggacacatctgctcactttaggaggtcggcaggaacctctgaaaagtgactggt agggataacgcagaaagcgctgtacgcgtctgttgatttgcacaaagatgggtagaaacagtctccc ccctgtcttaggttgaatgagcagggaccgccacccatacgggattacaggtcgcatgtacagccc ccgttagcccattagaccccgagtctagagaacctagagcagaaccacgggcactaaggtacacc atgagcgtttccaccggatgtactagataaacggccggtttctggccaggctcacgtcggagcgcat acttgtcta

RECURRENCE

Forgets Long Sequences

Computationally Cheaper

Very Slow to Train

Fast to Inference

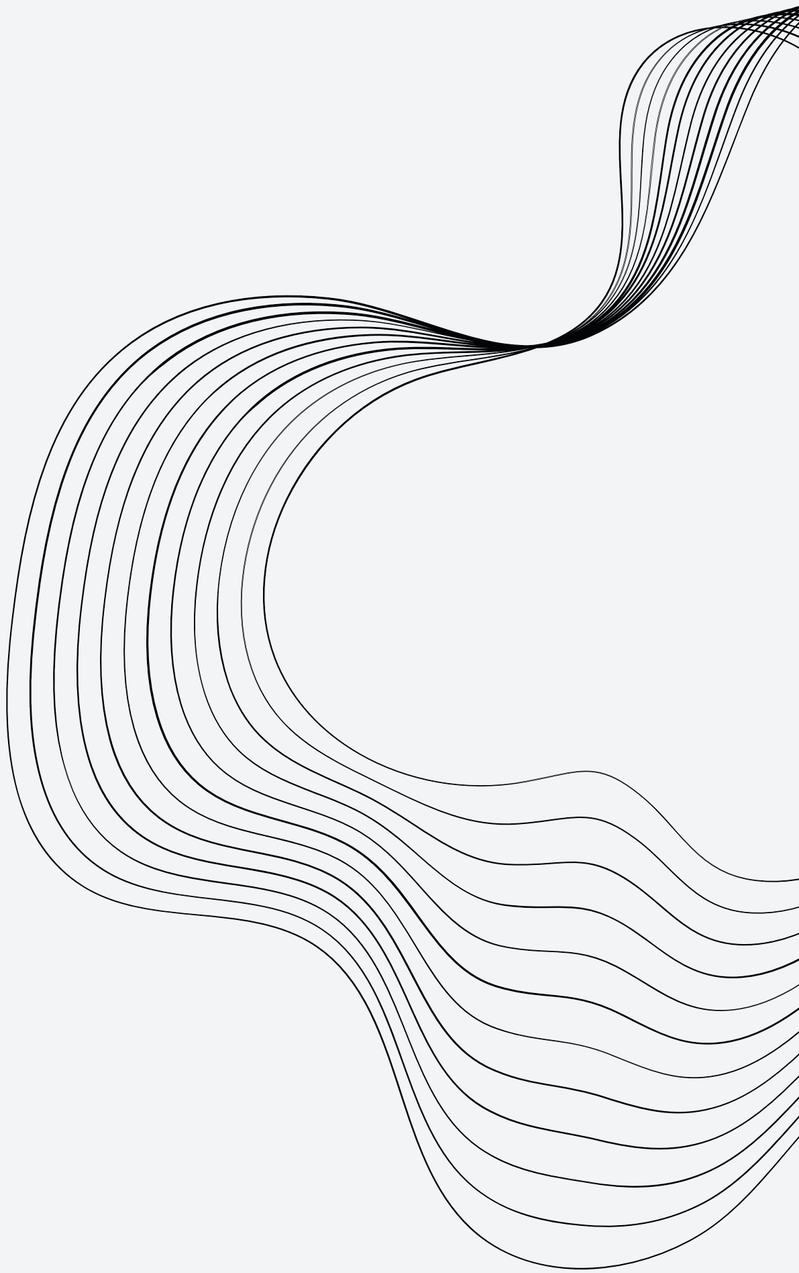
TRANSFORMERS

Extremely Fast to Train

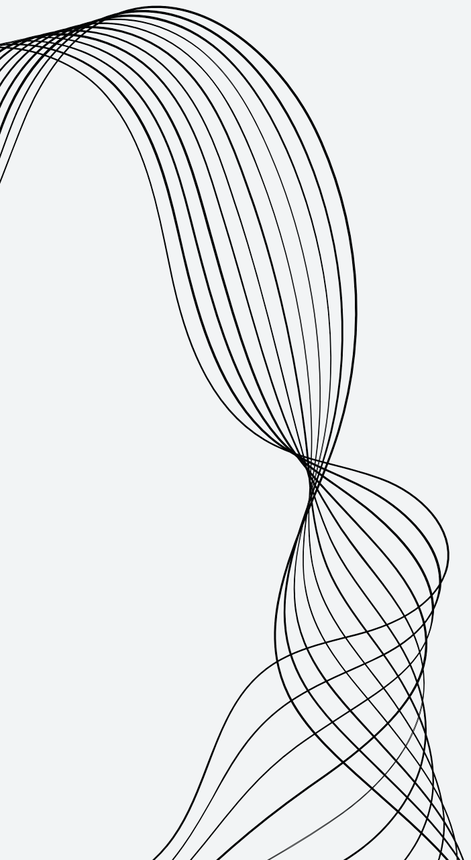
Full Sequence Attention

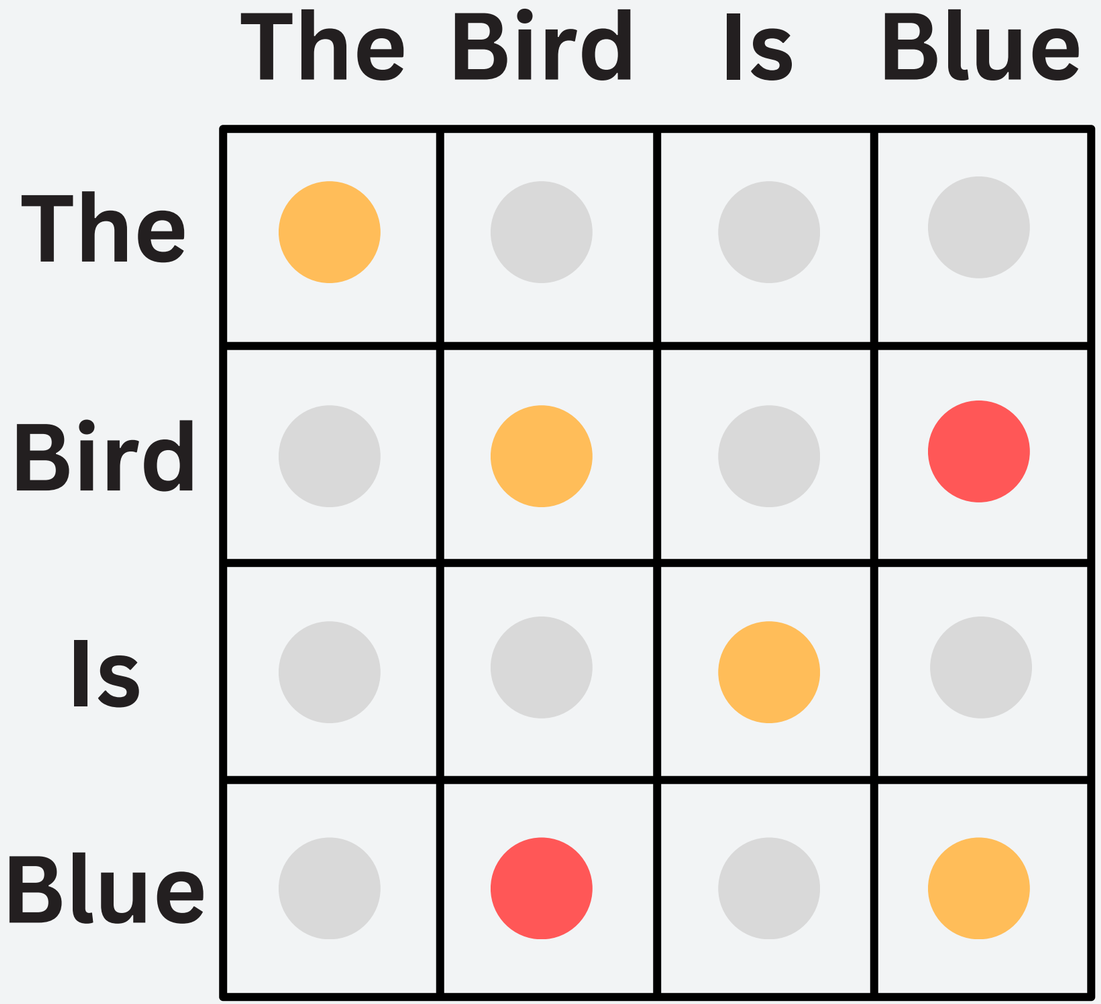
Slow to Inference

Massive Data Cost



THE SECRET IS ATTENTION





ALTHOUGH ATTENTION HAS A PROBLEM...

T = 10

10 * 10 = 100

T = 100

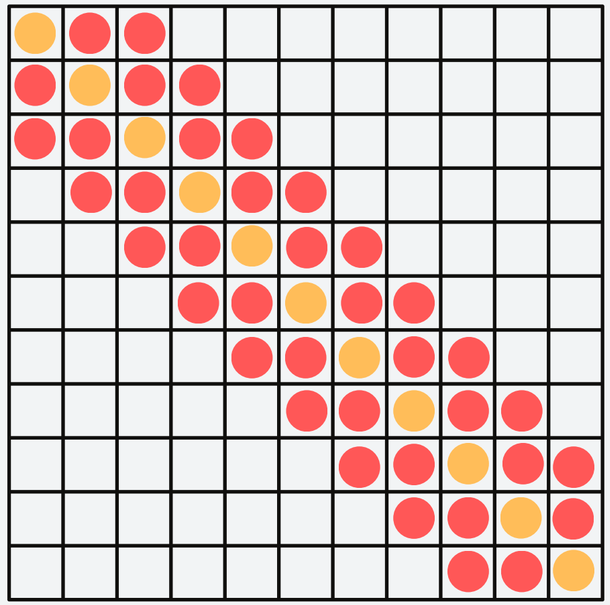
100 * 100 = 10000

T = 100,000

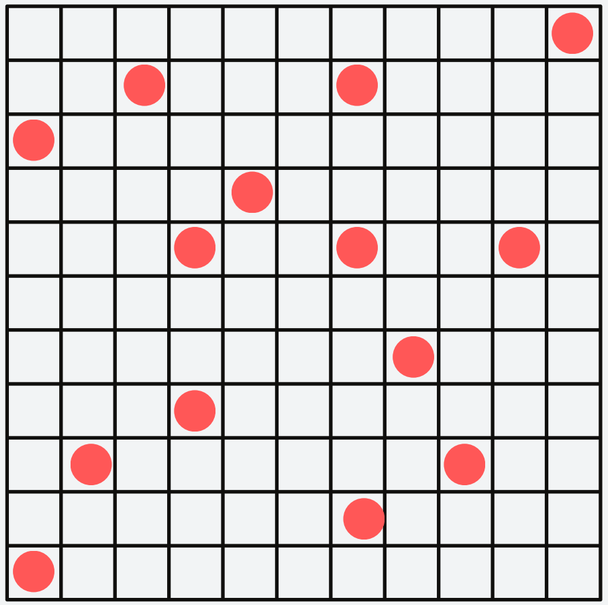
100000 * 100000 = 10 Billion

POTENTIAL SOLUTION: SPARSE ATTENTION

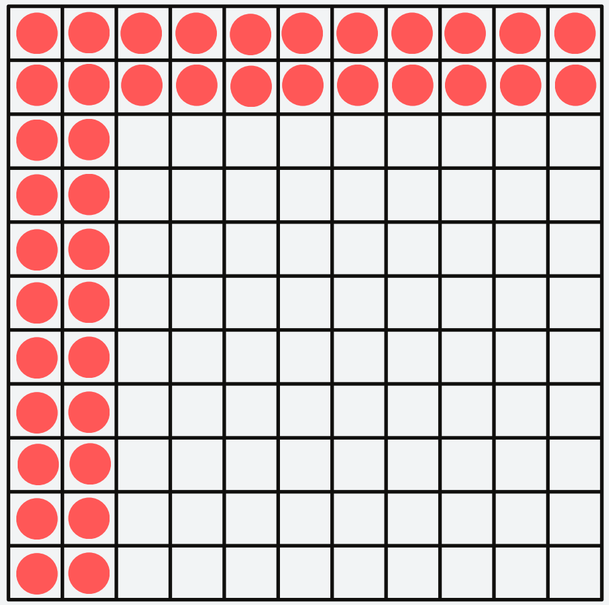
WINDOWED ATTENTION



RANDOM ATTENTION

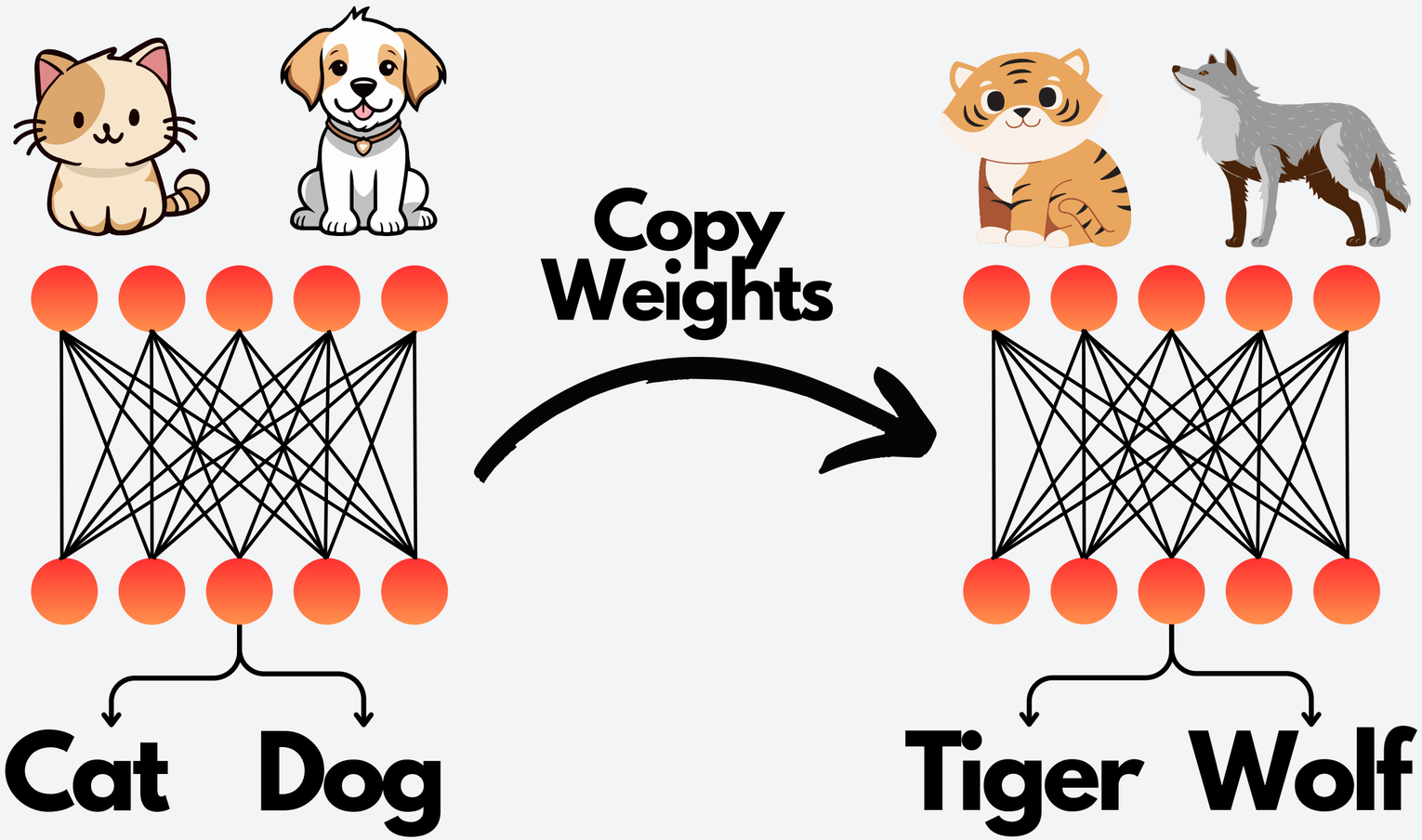


GLOBAL ATTENTION



IF THE DATA COST IS SO HIGH, HOW DO I USE IT?

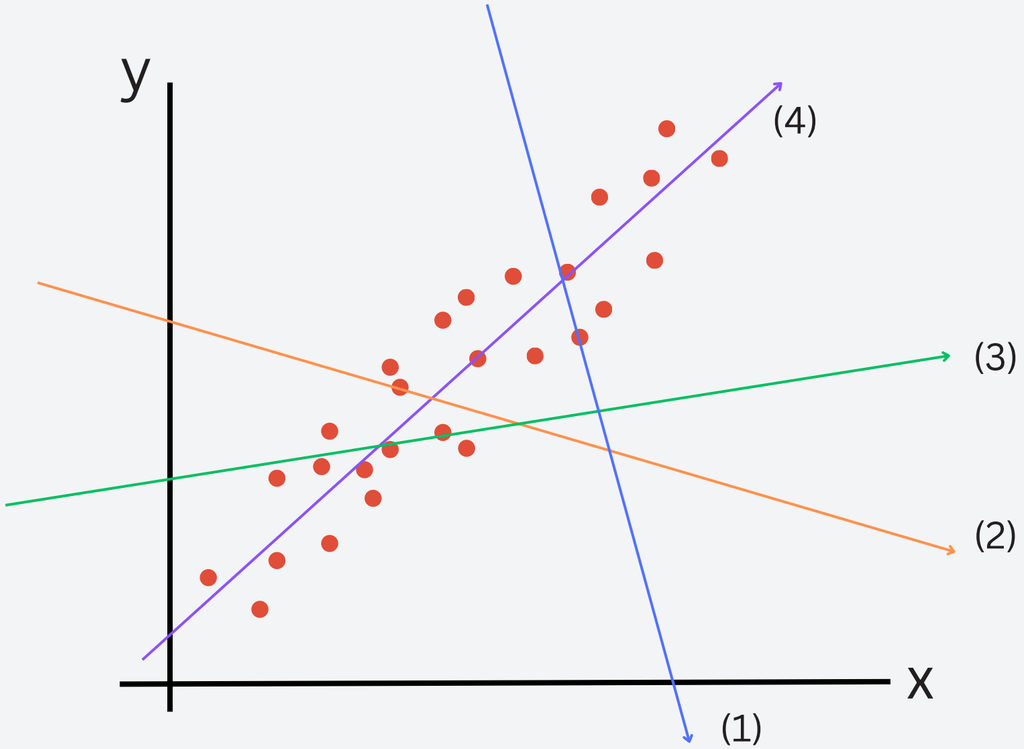
TRANSFER LEARNING



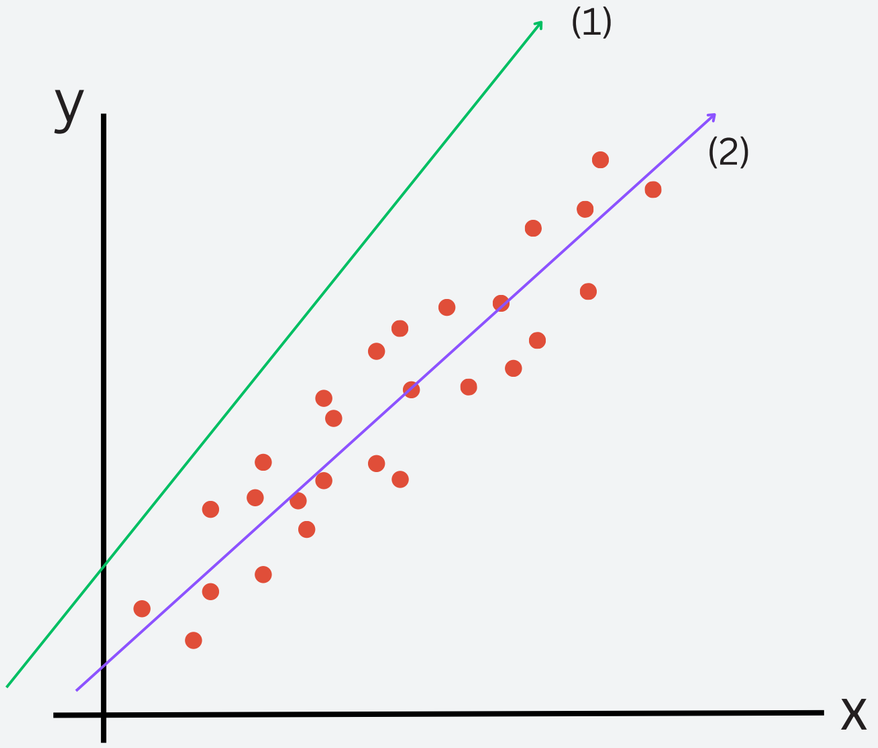
KNOWLEDGE IS TRANSFERABLE

Remember, Models are just bags of numbers to be optimized! If we don't know anything, the numbers start at random, but preinitializing with numbers that are already good at a similar task is preferred!

Random Weights



Pretrained Weights



TRANSFER LEARNING IS THE SECRET INGREDIENT

The vast majority of the worlds data is **Unlabeled**, the only way to deal with this type of data is Unsupervised Learning as we saw before.

TRAINING PATHWAY

Training modern Neural Networks typically perform a pathway of learning

Train on a **Large Corpus** of unlabeled data:

- Giant Image Dataset
- All of Wikipedia
- LibriVox Recordings

PRE-TRAINING

What do you want to actually do?

- Image Classification
- Question Answering
- Sentiment Analysis
- Speech Recognition

FINE-TUNING

This is new and used in Language Models today

- Human Feedback Loop
- Close domain gap (get model to understand specific domains like medicine)

INSTRUCT TUNING

CURRENT STATE OF THE ART

VISION	LANGUAGE	AUDIO	MULTIMODAL
Vision Transformer	GPT-4	Whisper	CLIP
DINO	Llama3	Wav2Vec2	Stable Diffusion
YOLO	Phi-3	Conformer	LLaVa
ConvNeXT	Mixtral	Wav2Vec2-XLSR	Segment Anything
SegFormer	Claude	RNN-Transducer	MatCha

ALMOST ALL OF THIS IS POWERED BY ATTENTION!!