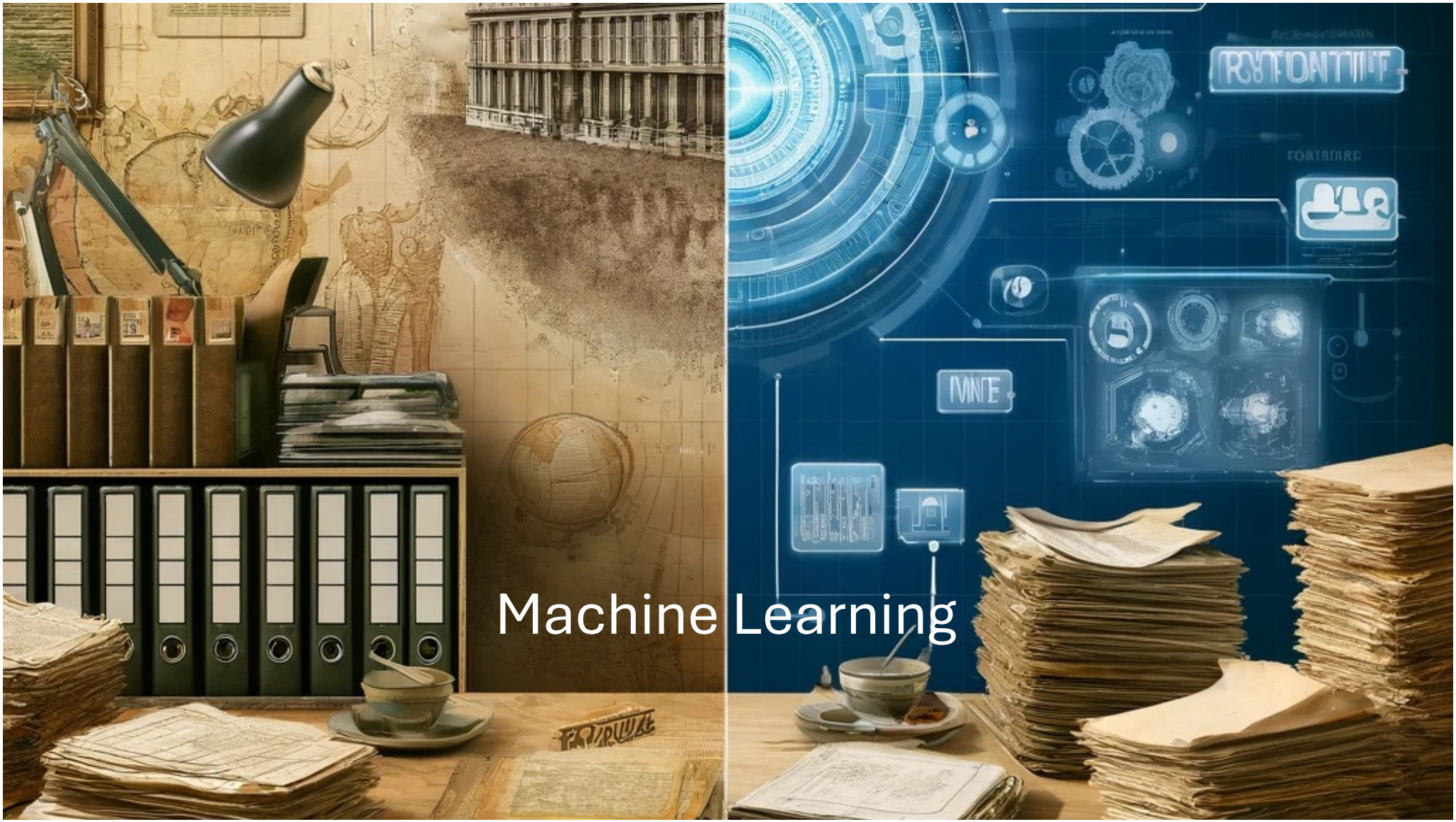
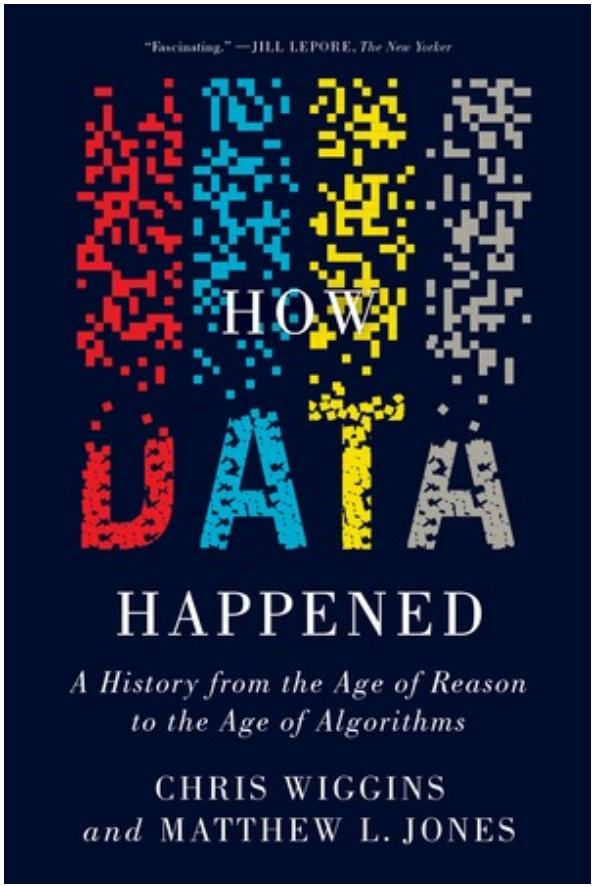

Machine Learning

Learning outcomes

- History of machine learning
- Epistemology of machine learning
- How can we use it for social scientific research?
- Under the hood of important ML techniques
- Transformers!

History of machine learning

- Traces its origins back to the mid-20th century, early pioneers like Alan Turing.
- NSA pursued Signal Intelligence (SIGINT) Analysis: had a lot of data, and less interested in “it works” than “doing it right”
- The field gained momentum in the 1950s and 1960s with the development of simple neural networks and the exploration of algorithms like the perceptron.
- "AI winters" of the 1970s and 1980s due to limited computational power and high expectations.
- The revival of machine learning came in the late 1980s and early 1990s, fueled by the advent of more powerful computers.
- Revolution in 2000s: GPU processing. Data from platforms. Deep learning.
- 2022 ChatGPT revolution



Machine learning and Surveillance Capitalism

- Machine learning become the foundation for platforms:
  - Online targeted advertising.
  - Behavioral modification.
  - Maximizing engagement on social media
- These are based on prediction logic – they don’t care why it works.
- “Will this individual buy the product if we show them the advertisement?”

Machine learning is what makes your personal data valuable

What is machine learning?

- Boundary between statistics and machine learning is blurry
- Some methods are both, such as linear regression
- Less about techniques than about the culture and aims

$$Y = f(X) = X^T\beta,$$

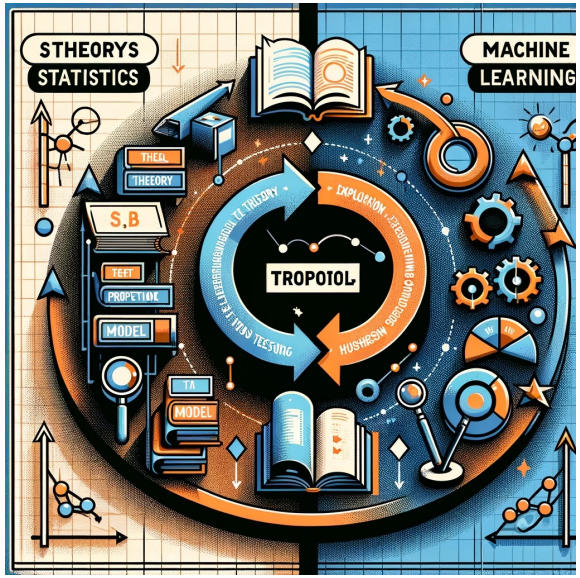
- ML focuses on Y, statistics focuses on β.

The Two Cultures

- Statistics:
  - *Understand* how an outcome is related to inputs.
  - Test theory
- Machine learning:
  - The aim is prediction in new unseen data
  - Models don't need to be understandable

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant = $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Machine Learning as an approach



Statistics:

Deductive. Linear approach: Derive testable propositions from clear theory, test using models.

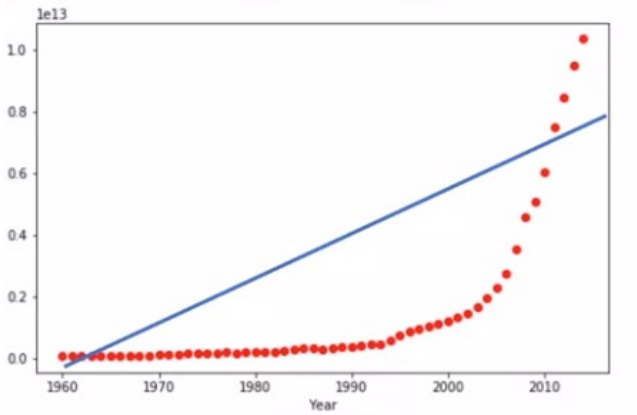
Machine learning:

Abductive. Grounded theory. Iterative: explore and develop theory.

Generate the research questions, concepts, and hypotheses that can later be rigorously tested with new data.

Advantages with ML

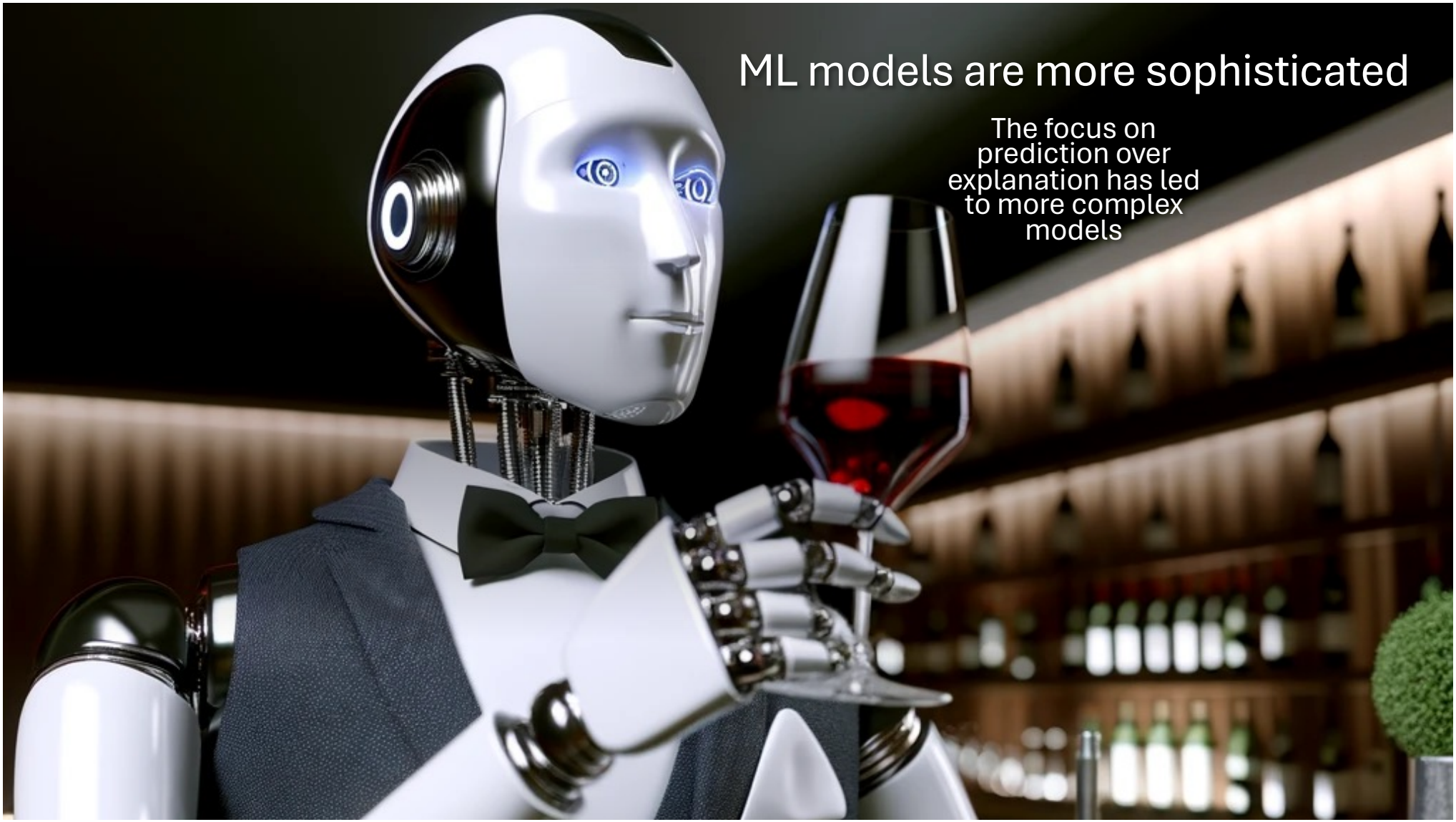
Reality is not linear!



Advantages with ML

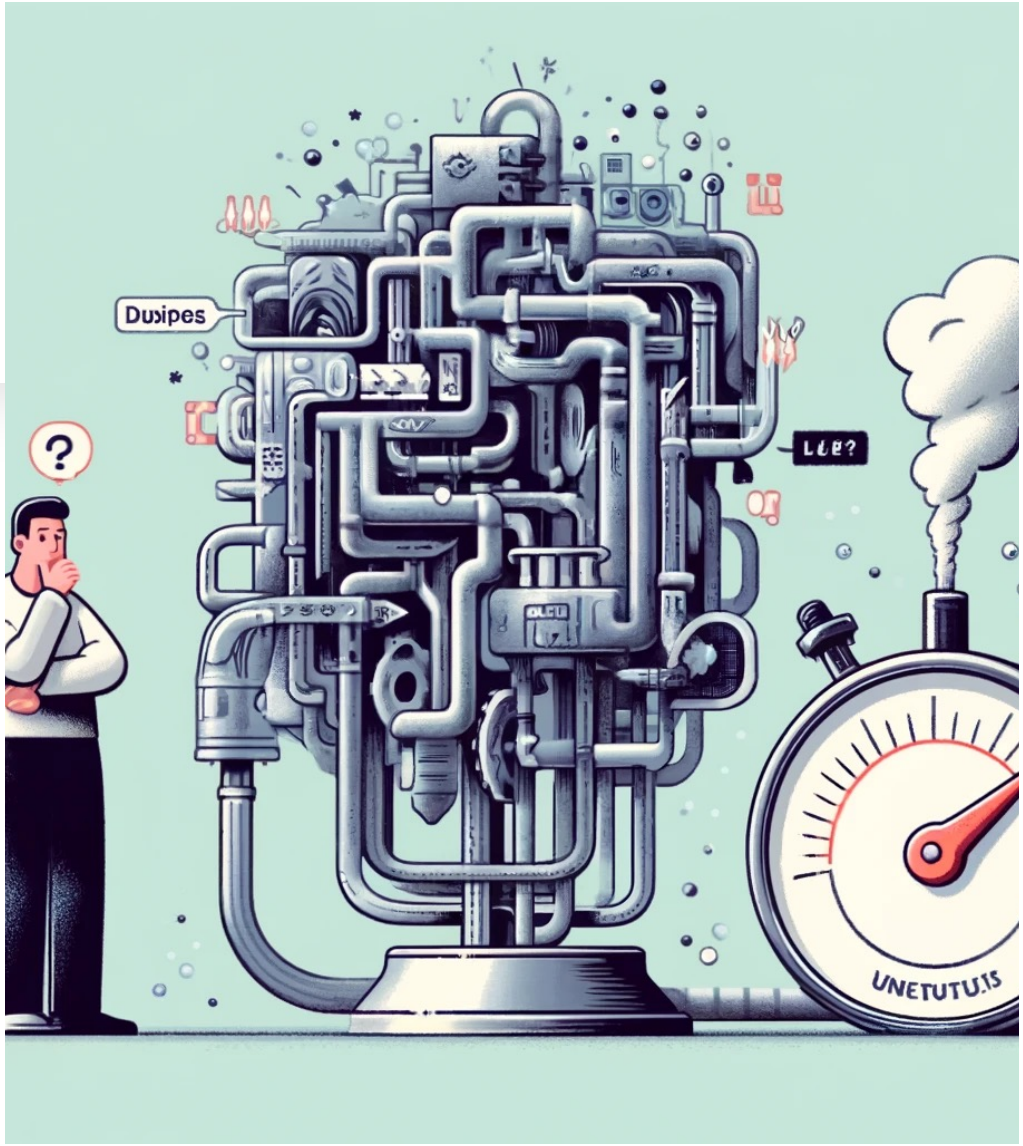
Population heterogeneity and causal complexity

- A cause doesn't always have the same effect
- Social scientific theory usually contains "if's" and "then's", but they cannot be captured by quantitative models.
- *On average* is often a terribly destructive way of dealing with data


ML models are more sophisticated
The focus on prediction over explanation has led to more complex models

Downsides of ML

- The predictions can be an opaque function of the inputs: hard to answer the *why?*
- The models may not easily provide meaningful estimates of uncertainty
- Estimation often more computationally intensive



The epistemology of machine learning

1. From *variables* to *patterns*
2. From *rules* to *associations*
3. From *surveys* to *sensors*

Unsupervised models

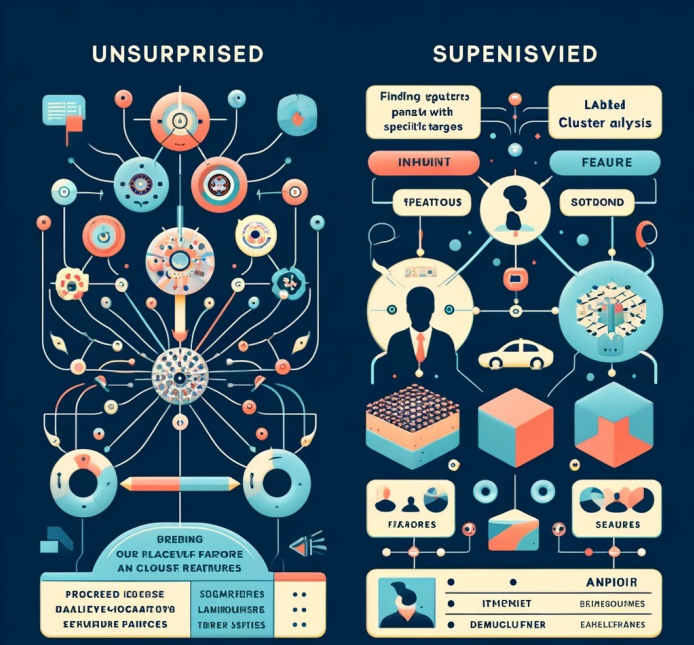
Identify pattern in data, or reduce dimensions.

- No target output to predict, no teacher showing the algorithm what it should aim for, and no immediate measure of success
- E.g. topic modeling, cluster analysis.

Supervised models

Mapping between features and outcomes.

- Uses labelled training data.
- E.g. identifying race from profile picture.



Reinforcement learning: learn-by-doing

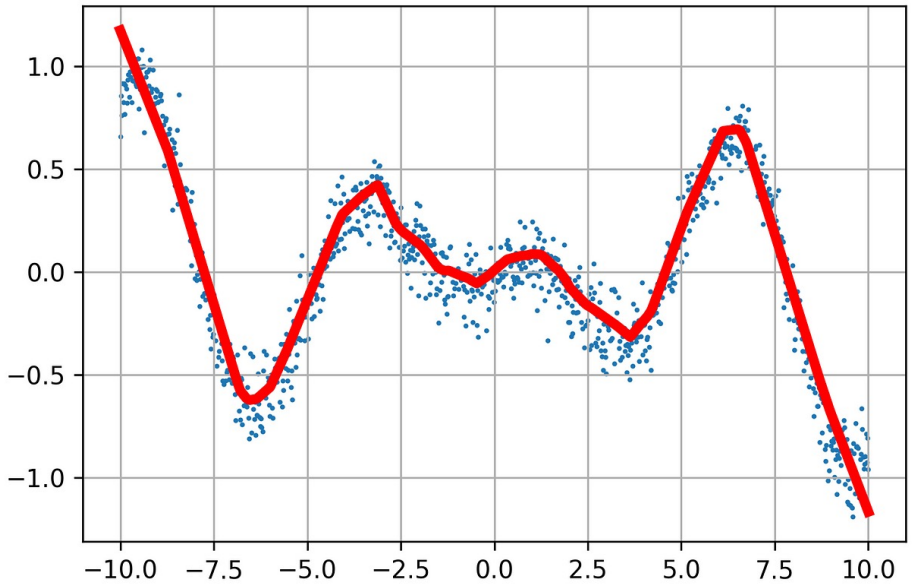
- Train AI for skills we don’t fully understand: we know how to walk, but it’s awfully hard to explain.

- Trial-and-error
- The better the AI walks, the more reward they get.
- Several techniques and algorithms underpin the actual learning process by which agents make decisions to maximize cumulative rewards.

Fitting supervised models

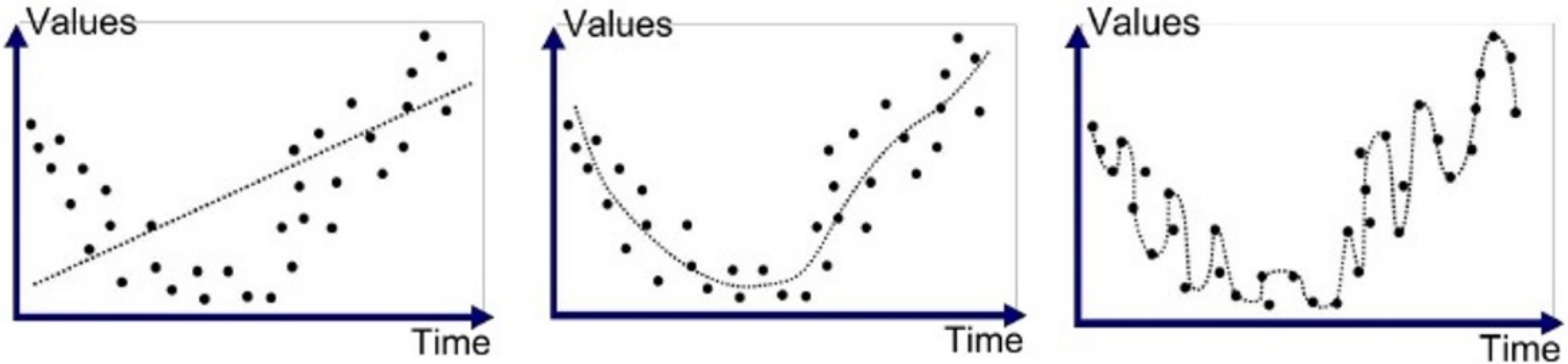
Optimization of parameters with respect to an objective function.
Such as good-old OLS model (Ordinary Least Squares) $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \sum_{i}^{n} (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i)^2.$

What distinguishes machine learning methods is that they can fit substantially more flexible functions.



Assessing model performance

Prediction error within the data set is not a good indicator of prediction in new data

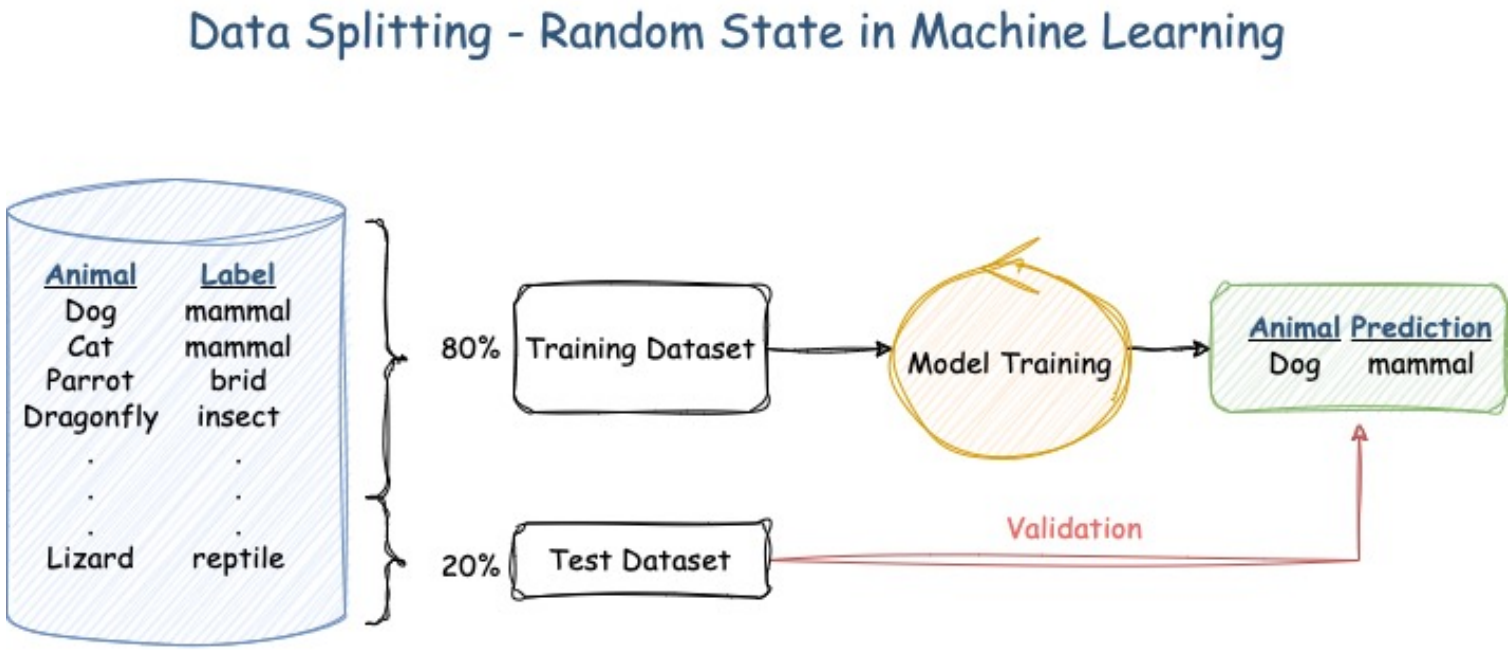


Underfitted

Good Fit/Robust

Overfitted

Sample splitting



Regularization: Keep it simple, stupid model

We can punish the model for being complex, as a way of keeping it as simple as possible.

- This is called regularization.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \underbrace{\sum_i^n \left(y_i - \boldsymbol{\beta}' \boldsymbol{x}_i\right)^2}_{\text{sum of squared residuals}} + \overbrace{\lambda \sum_{j=1}^p |\beta_j|^q}^{\text{complexity penalty}}$$

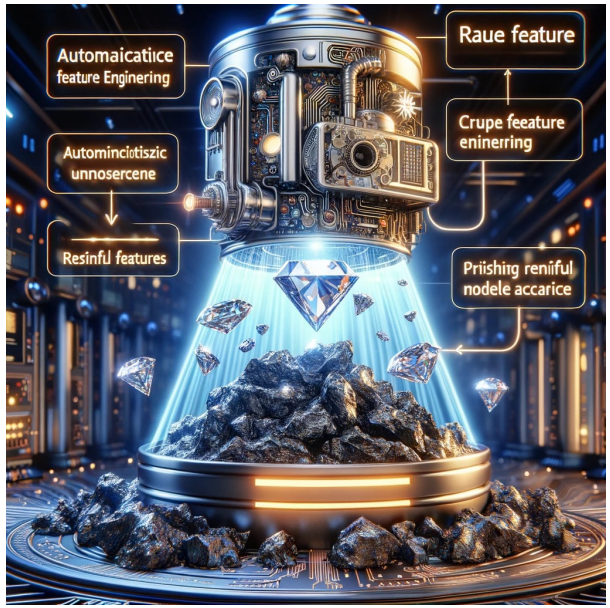
- λ: strength of penalty
- q: the type of regularizer (q=1 LASSO sets many coefficients to zero; q=2 Ridge regression, smooth push to zero)

Hyperparameter search

- What λ parameter will lead to the best results?
- We can optimize the parameters themselves!

Automatic feature engineering

- What variables/features should be included in prediction? How to preprocess?
- How to measure of latent concept like ideology or economic growth?

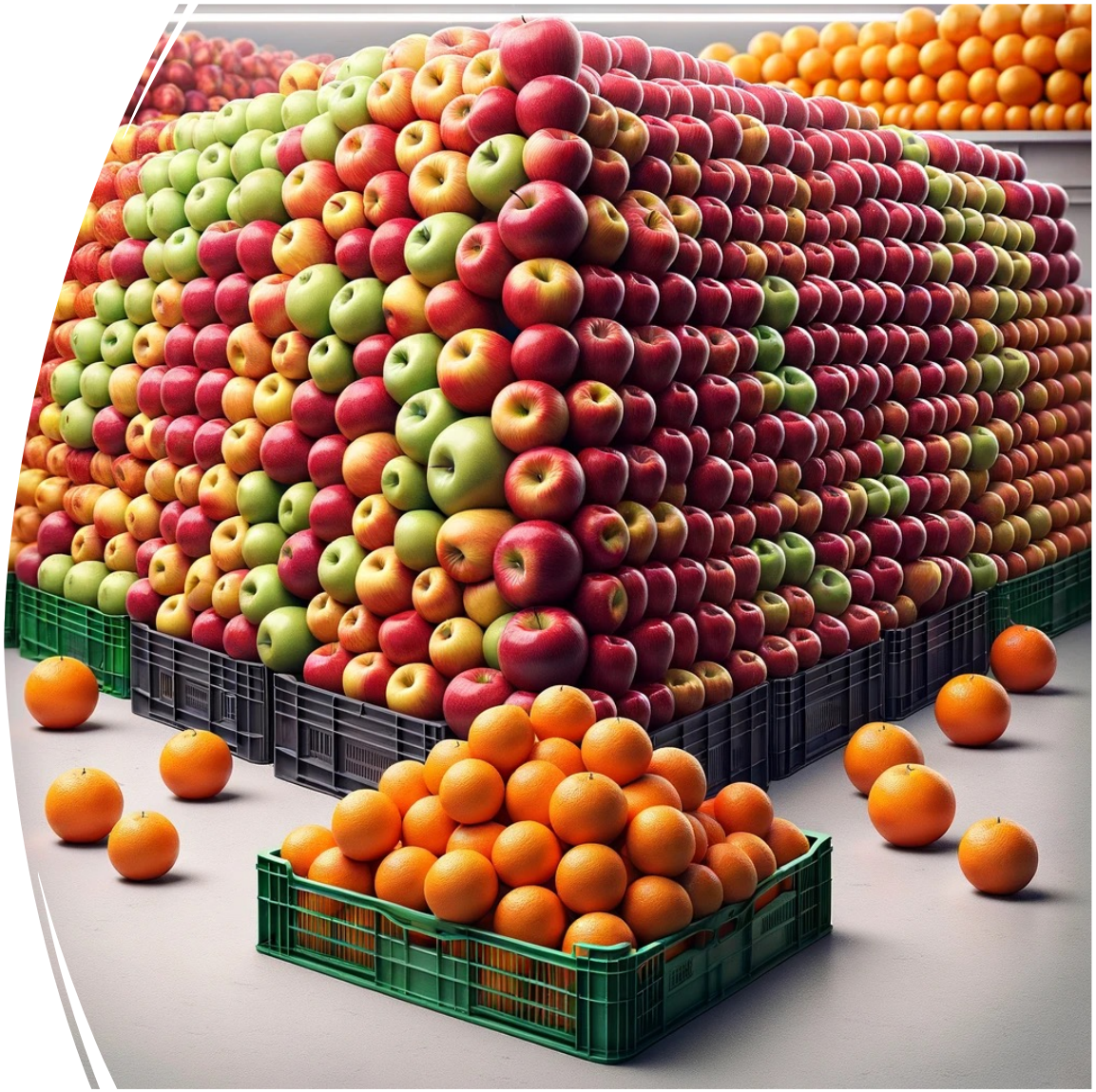


- Machine learning can figure it out automatically.
- Especially important in e.g. vision or sound, where a pixel says nothing on its own.

What does this mean for theory?

Evaluating performance

- You have pictures of 500 apples and 50 oranges.
- You apply a machine learning classification model?
- Your model reaches 92% accuracy.
- Is it a good model?



Performance metrics

Accuracy: Proportion that is correct. Misleading for imbalanced datasets.

Precision: The proportion of positive identifications that were actually correct. Important when the cost of false positives is high (e.g., spam detection).

Recall: The proportion of actual positives that were correctly identified.
Crucial when it's important to capture all positives (e.g., disease screening).

F1-score: The harmonic mean of precision and recall.

Macro: combines the scores from both classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

	precision	recall	f1-score	support
0	0.77	0.86	0.81	37584
1	0.84	0.75	0.79	37577
accuracy			0.80	75161
macro avg	0.81	0.80	0.80	75161
weighted avg	0.81	0.80	0.80	75161

What can we actually use ML for?

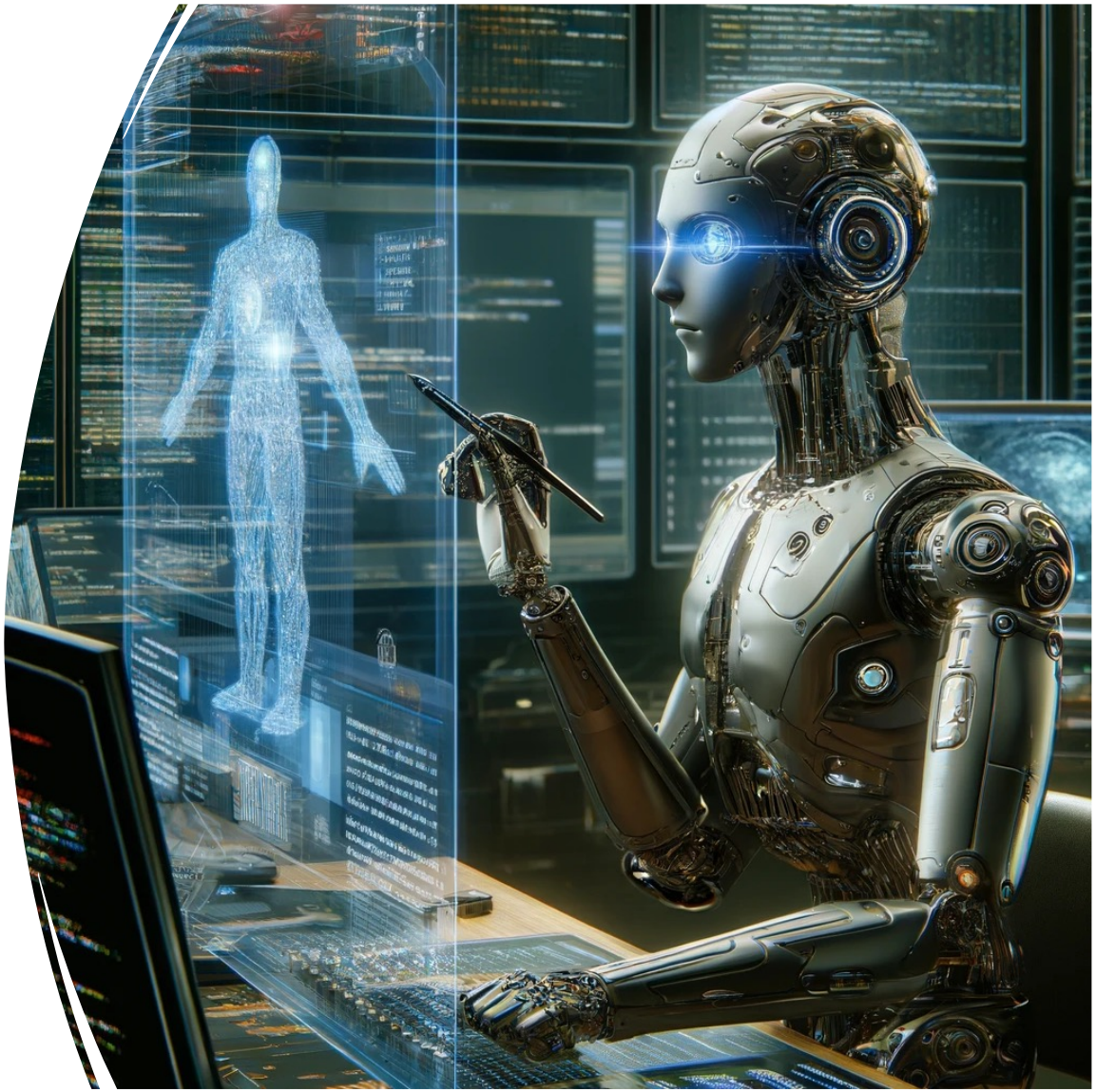
Improved regression models

- We can use ML as a form of statistics on steroids, with more complex modeling capabilities, handling non-linear relationships, and managing high-dimensional data
- **Interpretable Models**: Some machine learning models, such as linear regression with regularization (like Lasso and Ridge), still allow for relatively straightforward interpretation. Yay, regression tables!
- There are also ways to make sophisticated models interpretable. Bootstrapping can be used to estimate confidence intervals and standard errors, including Random Forests.
- However, these techniques might not always provide exact p-values or the traditional statistics familiar from ordinary least squares regression.

Coding and classification

One of the most widely used tools in the social sciences is hand coding: manually classifying observations into a set of categories that are determined before the analysis begins.

The models allow measuring virtually any latent aspect of data!

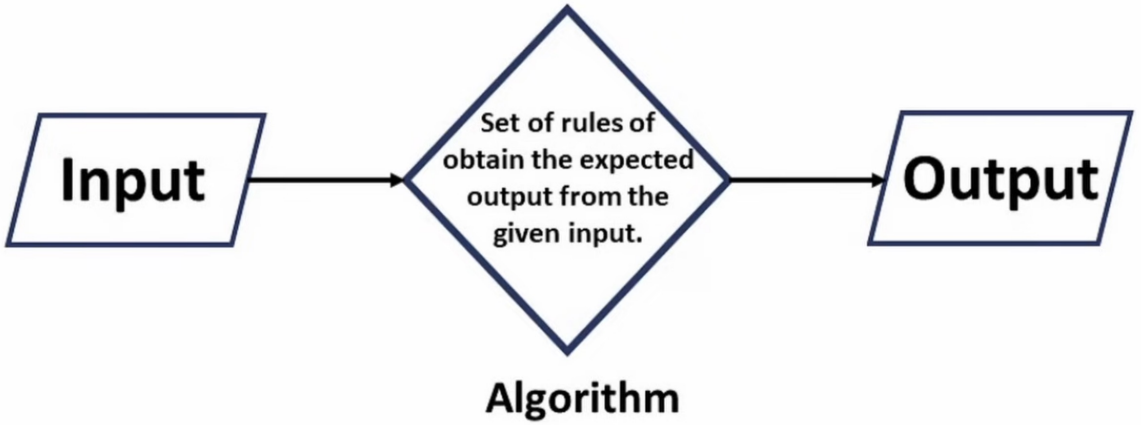


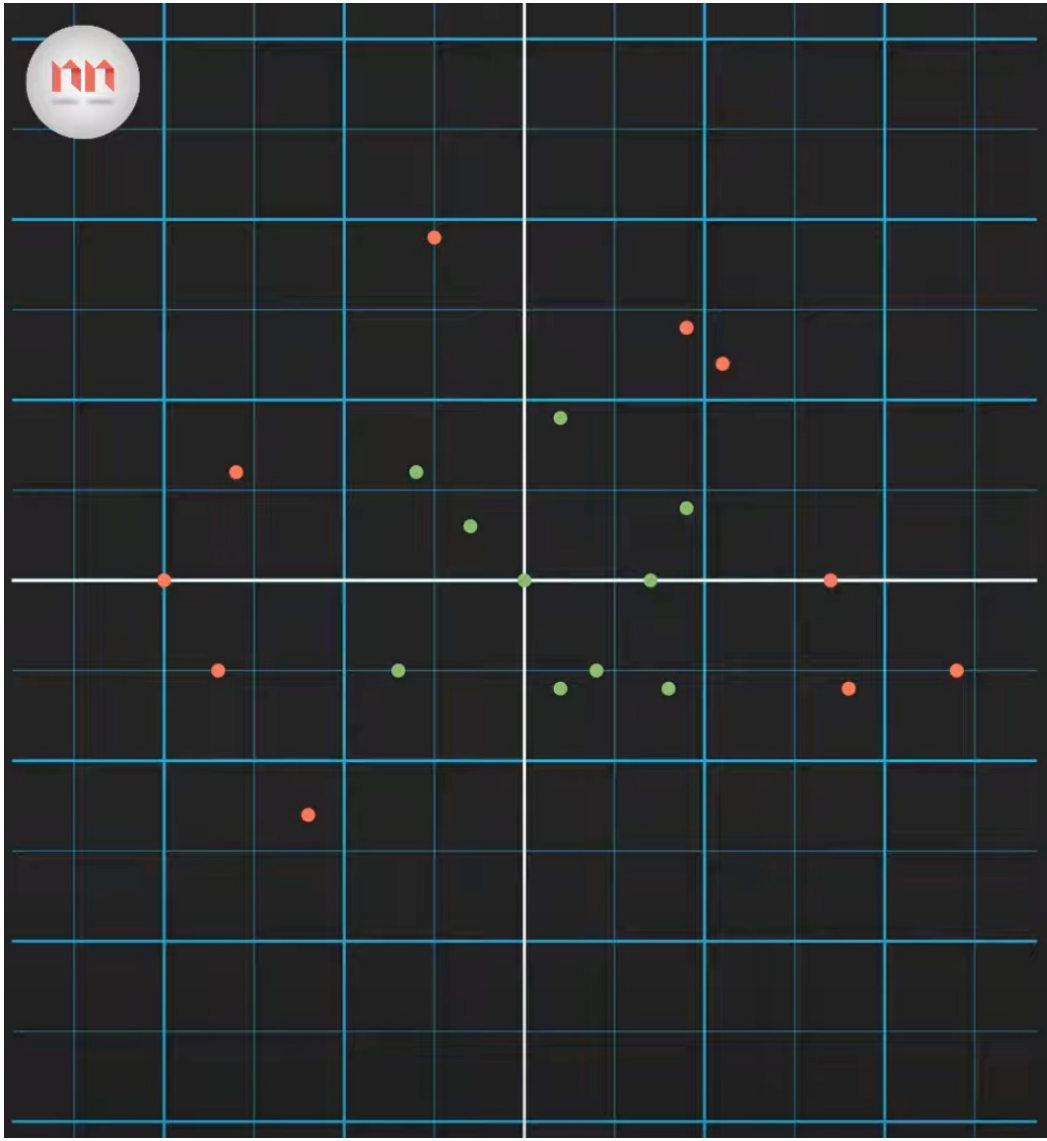
Causal inference from observational data

- Imagine we're interested in studying the effect of attending a private high school (treatment) versus a public high school (control) on college admission rates (outcome).
- Obviously, the problem is confounding factors! E.g., having rich parents.
- We calculate *the propensity score:* the probability of a student attending a private high school given their observed characteristics (covariates). We estimate this probability using statistical or machine learning models
- We match students based on their propensity scores.
- The effect of treatment is then = the average difference between the matched individuals

How do the models actually work?

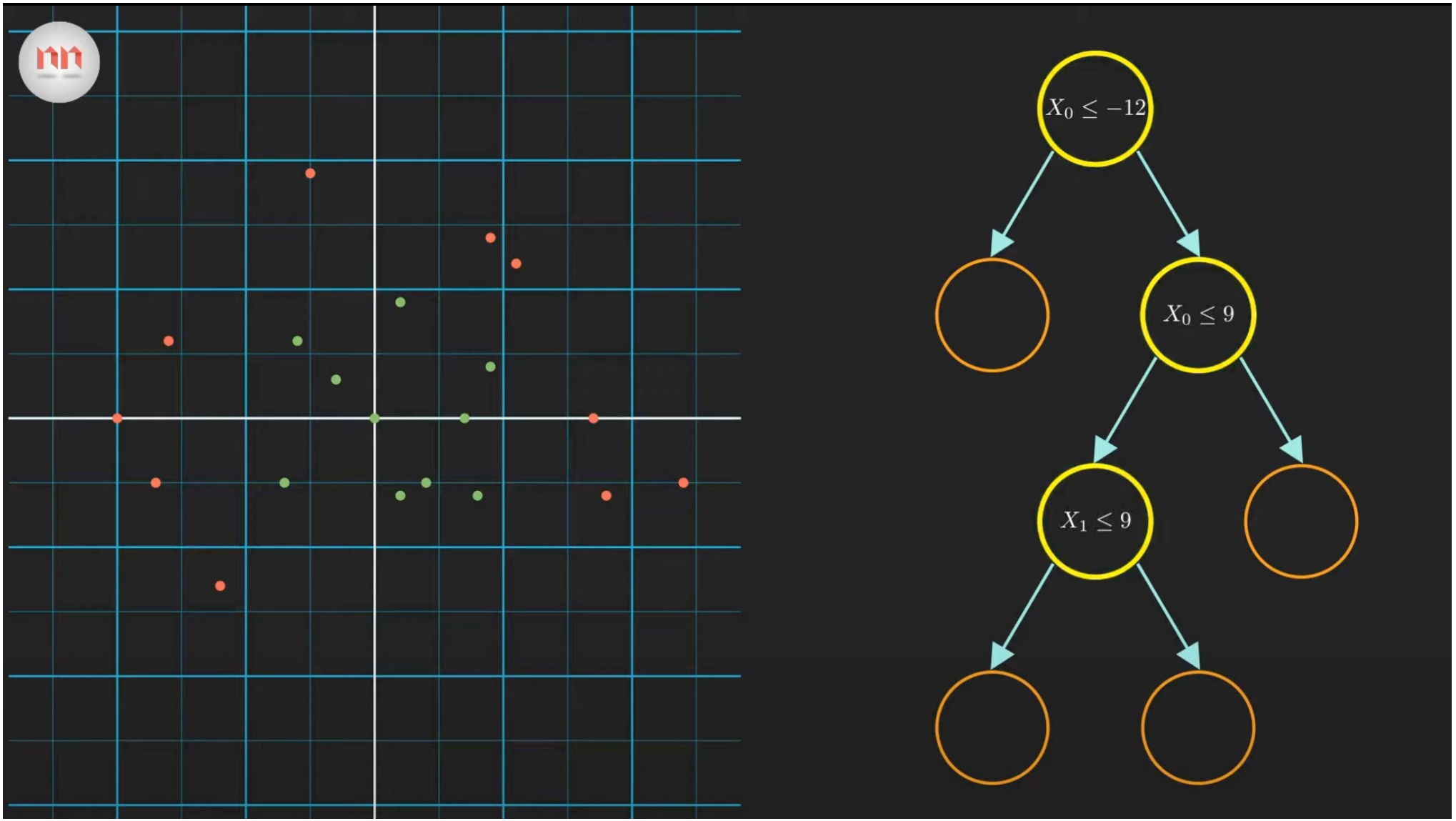
Let's get nerdy!

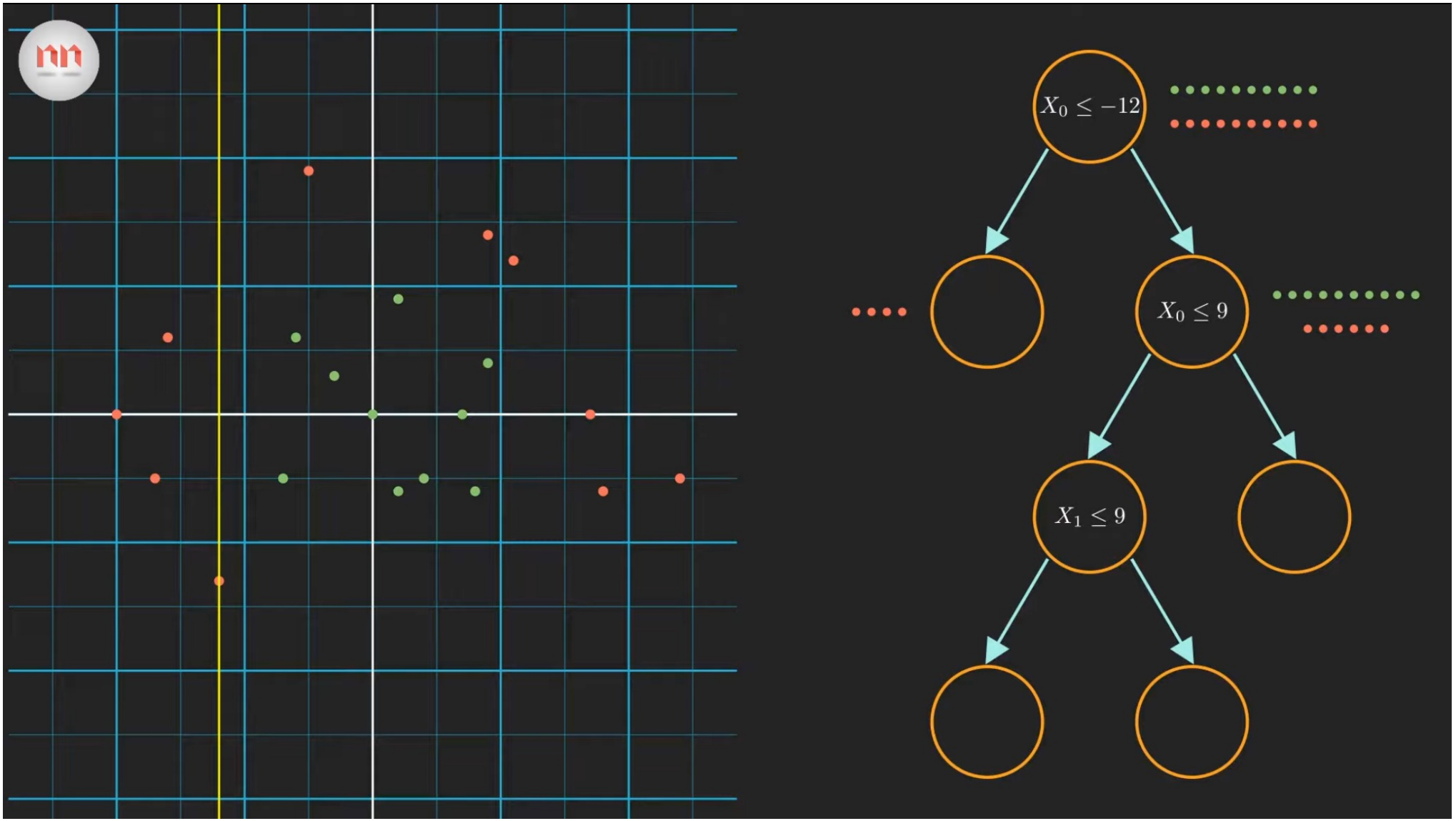

Input
Set of rules of obtain the expected output from the given input.
Output
Algorithm

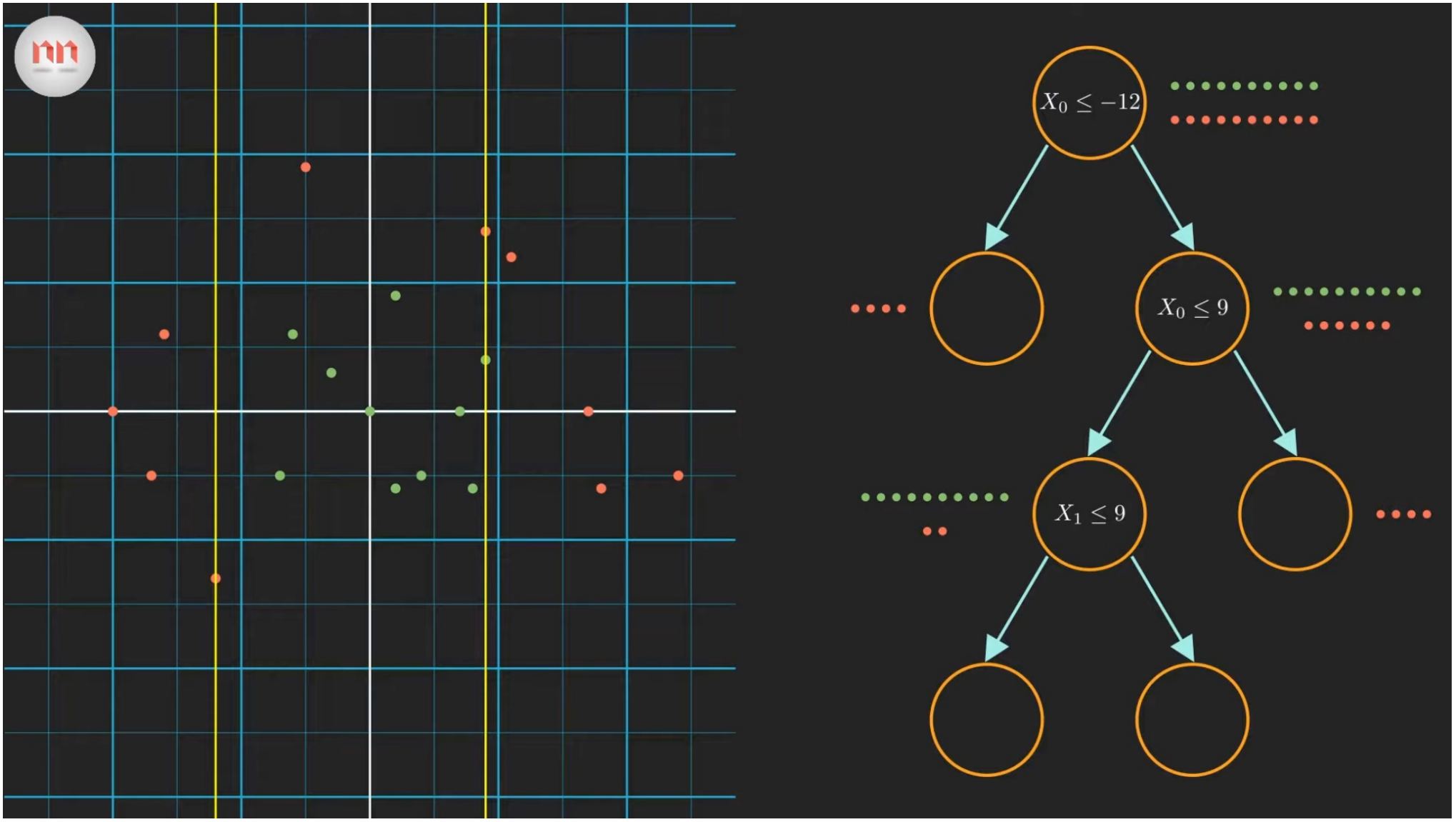


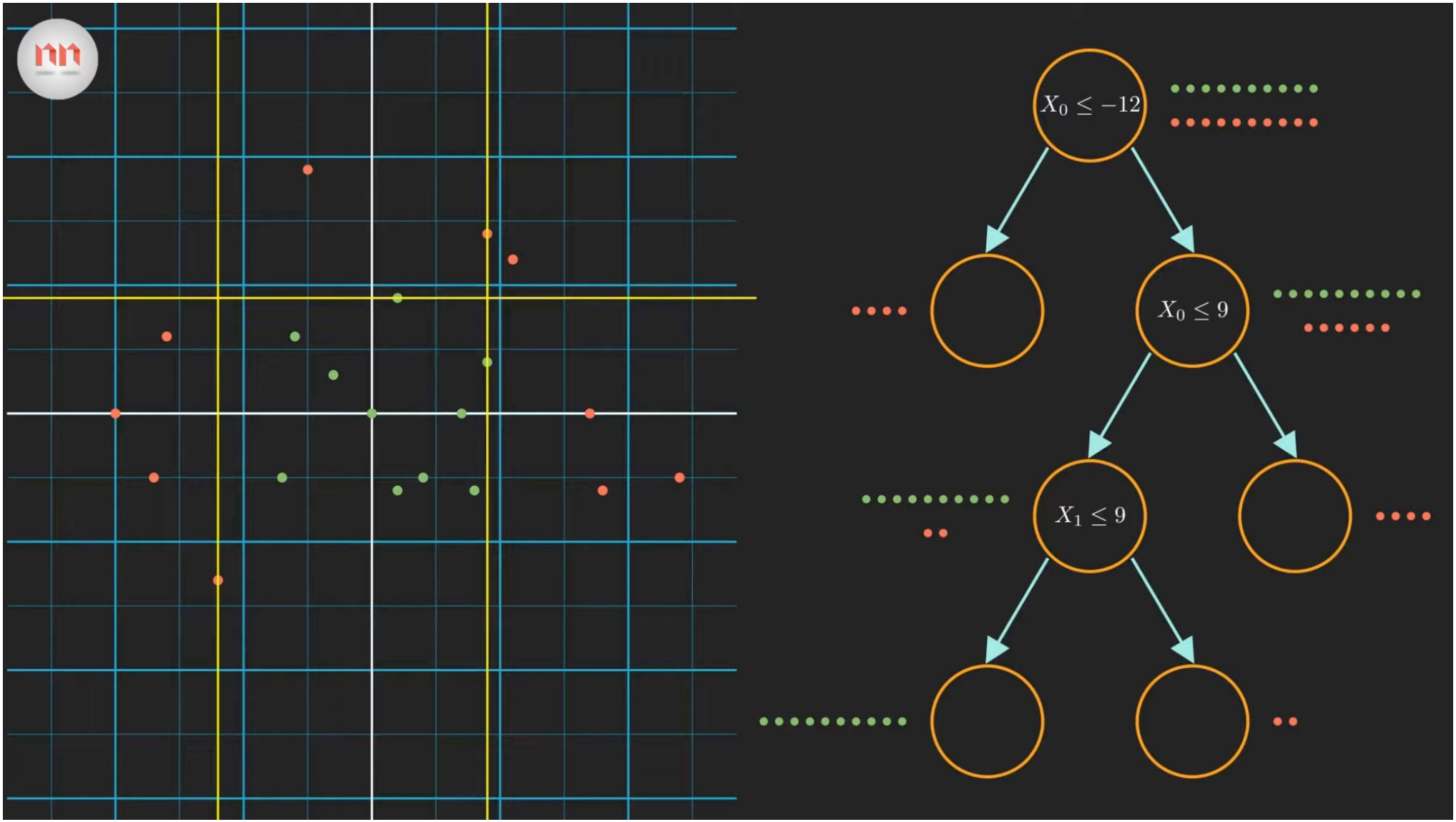
Decision trees

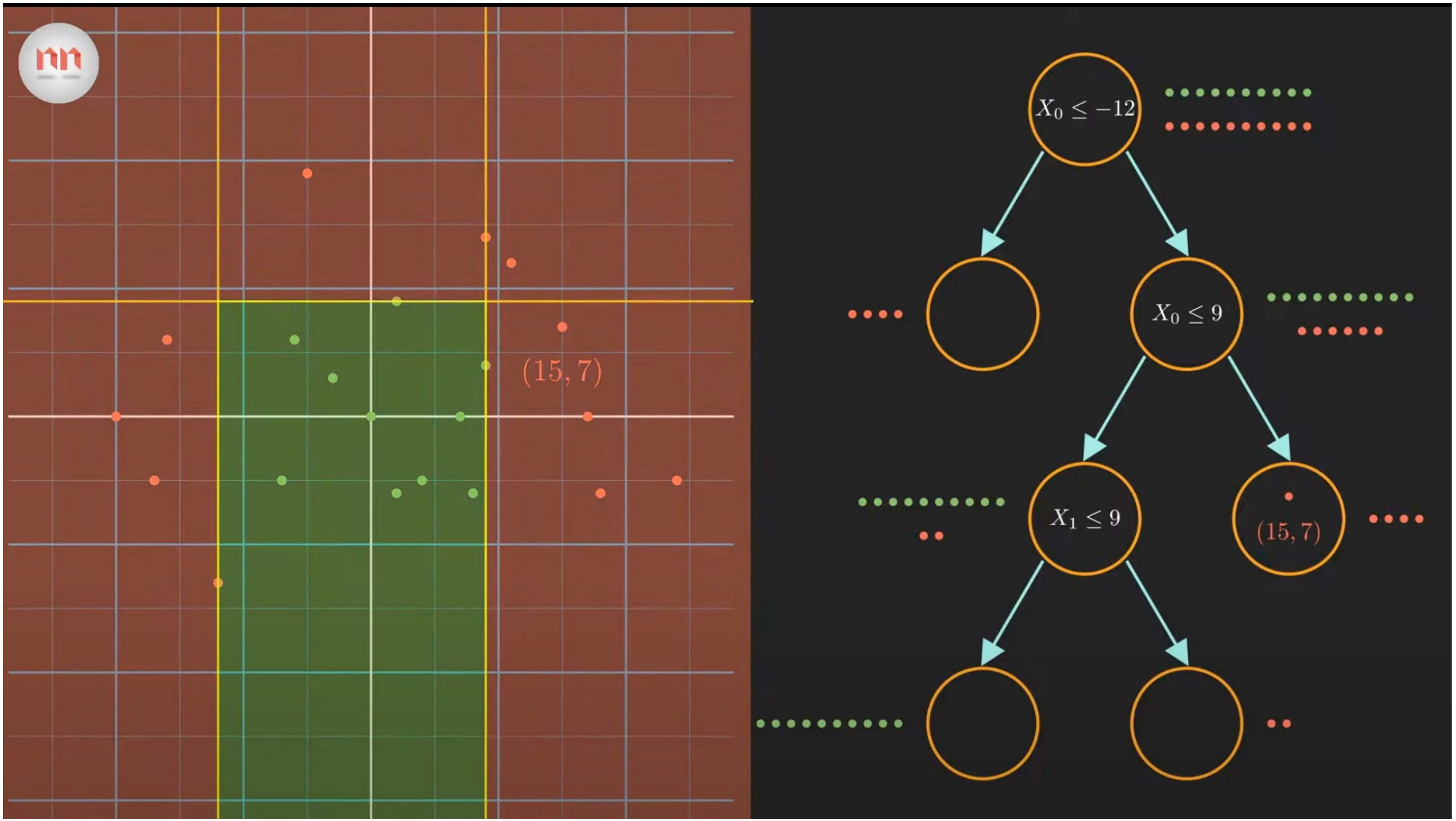
- We want to separate the two classes of dots
- Not possible with a straight line


$X_0 \leq -12$
$X_0 \leq 9$
$X_1 \leq 9$


$X_0 \leq -12$
$X_0 \leq 9$
$X_1 \leq 9$


$X_0 \leq -12$
$X_0 \leq 9$
$X_1 \leq 9$


$X_0 \leq -12$
$X_0 \leq 9$
$X_1 \leq 9$


(15, 7)
$X_0 \leq -12$
$X_0 \leq 9$
$X_1 \leq 9$
(15, 7)

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

id
2
1
3
1
4
4

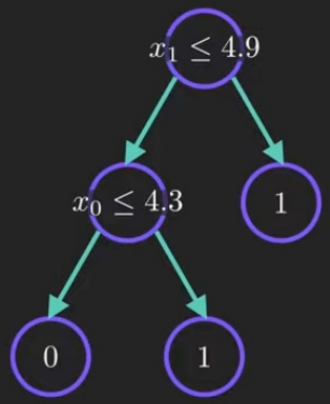
id
4
1
3
0
0
2

id
3
3
2
5
1
2

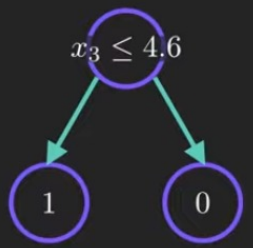
Random Forest

- Decision trees are highly sensitive to training data. It can fail to generalize.
- A forest is several trees.
- Each tree is trained on a random subset of the training data and makes predictions independently. The final prediction is obtained by averaging the predictions of all the trees (for regression tasks) or by using voting (for classification tasks).

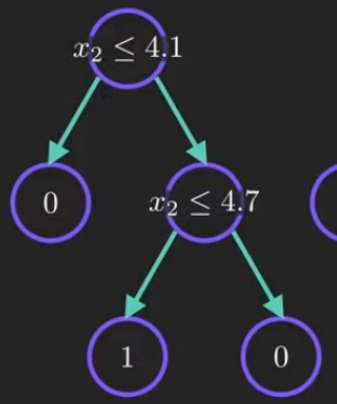
x_0, x_1



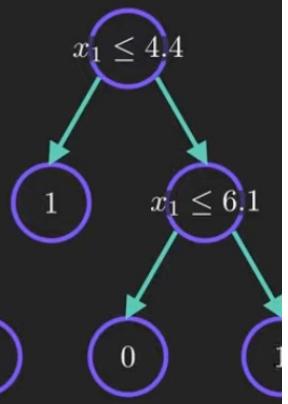
x_2, x_3



x_2, x_4



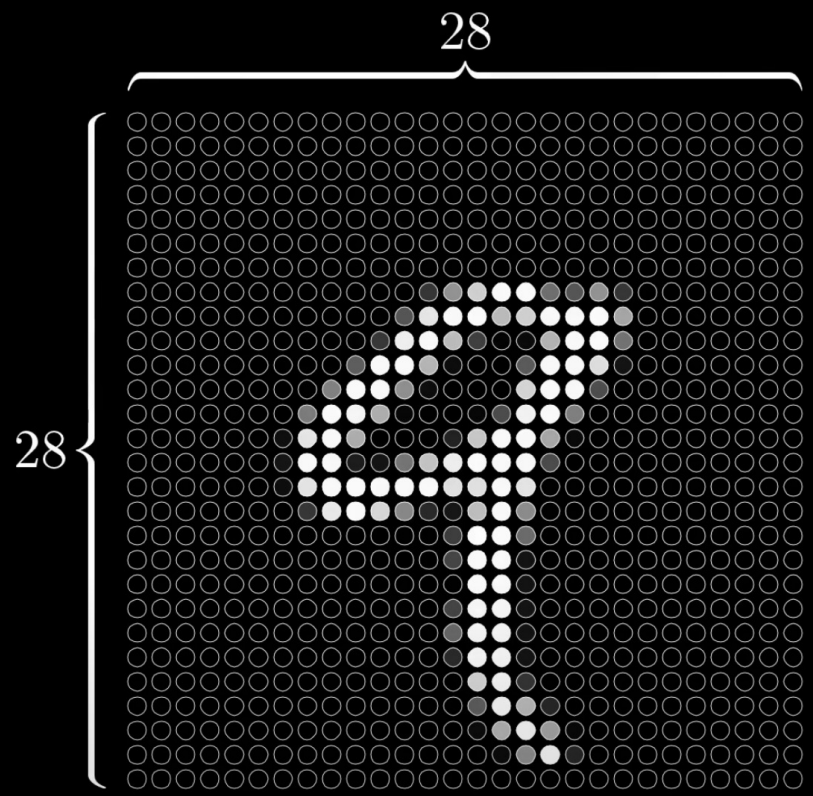
x_1, x_3



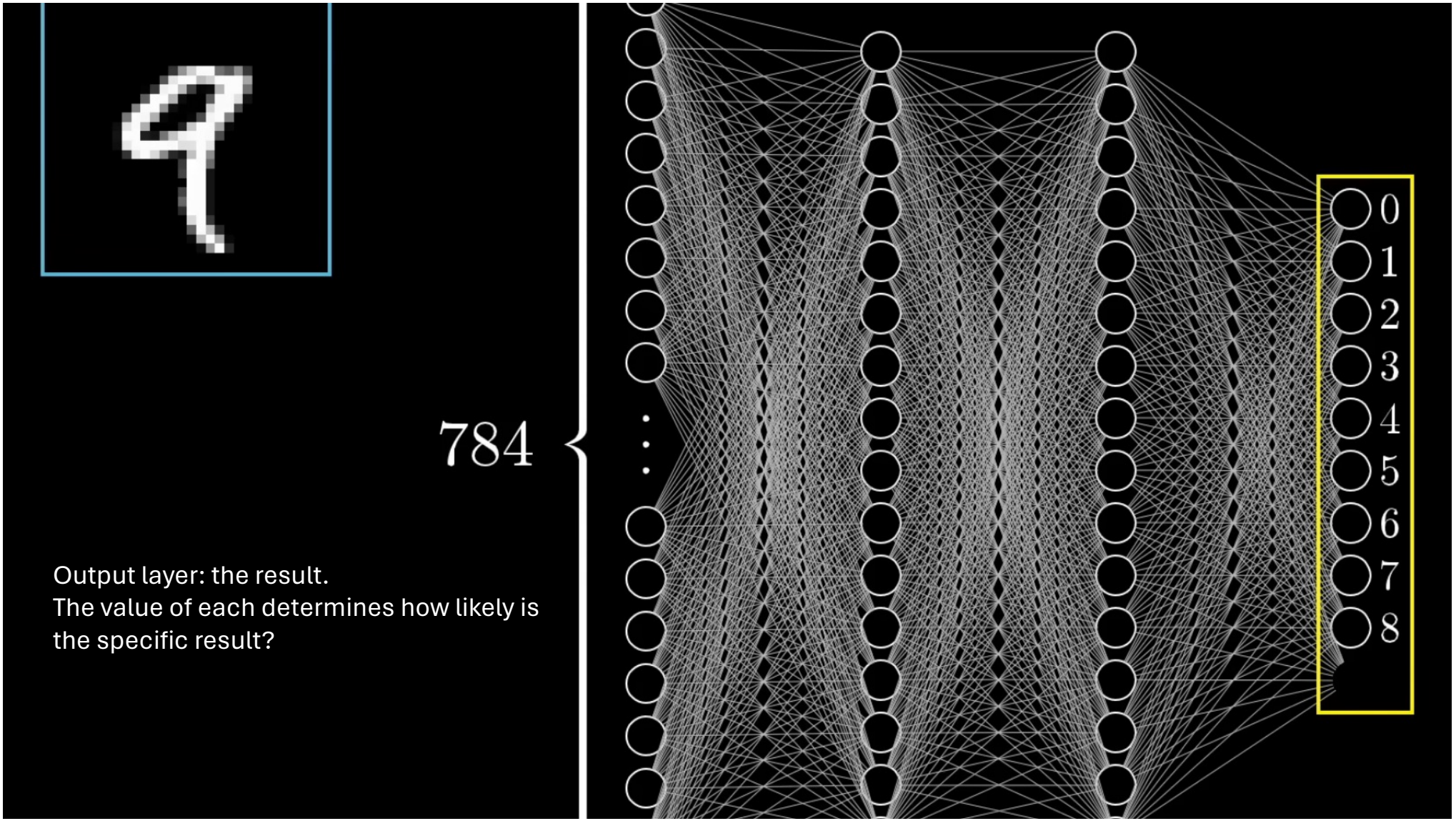
Artificial Neural Networks

Task: identify the handwritten number

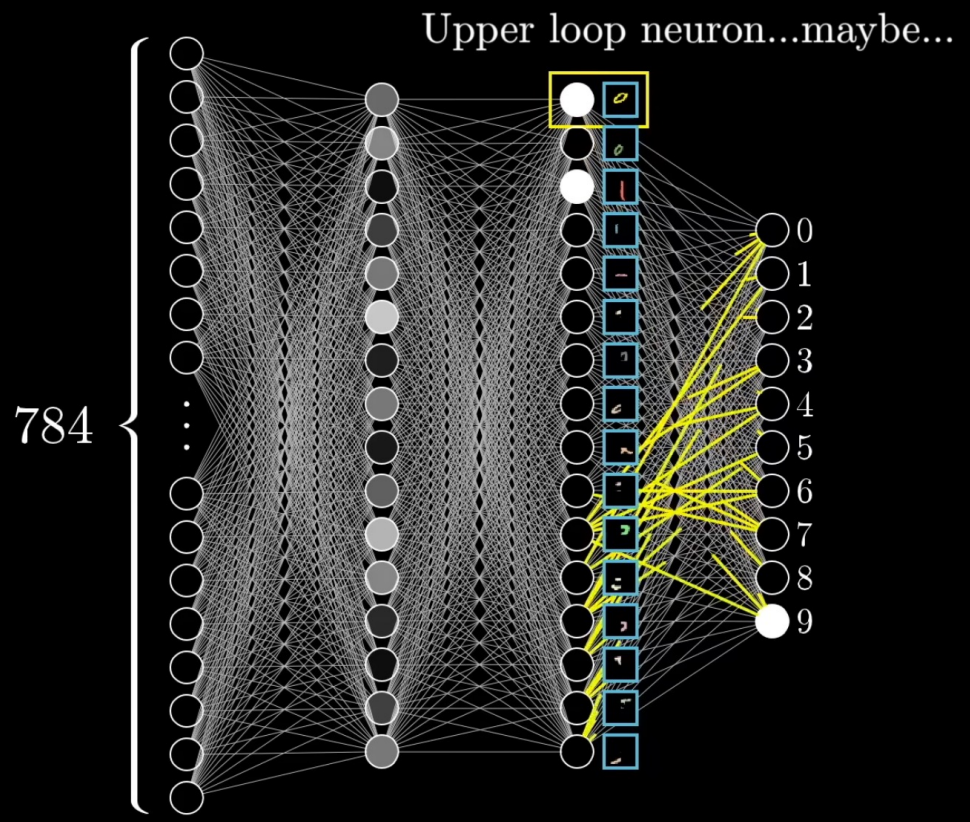
Greyness is activation: from 0 to 1.



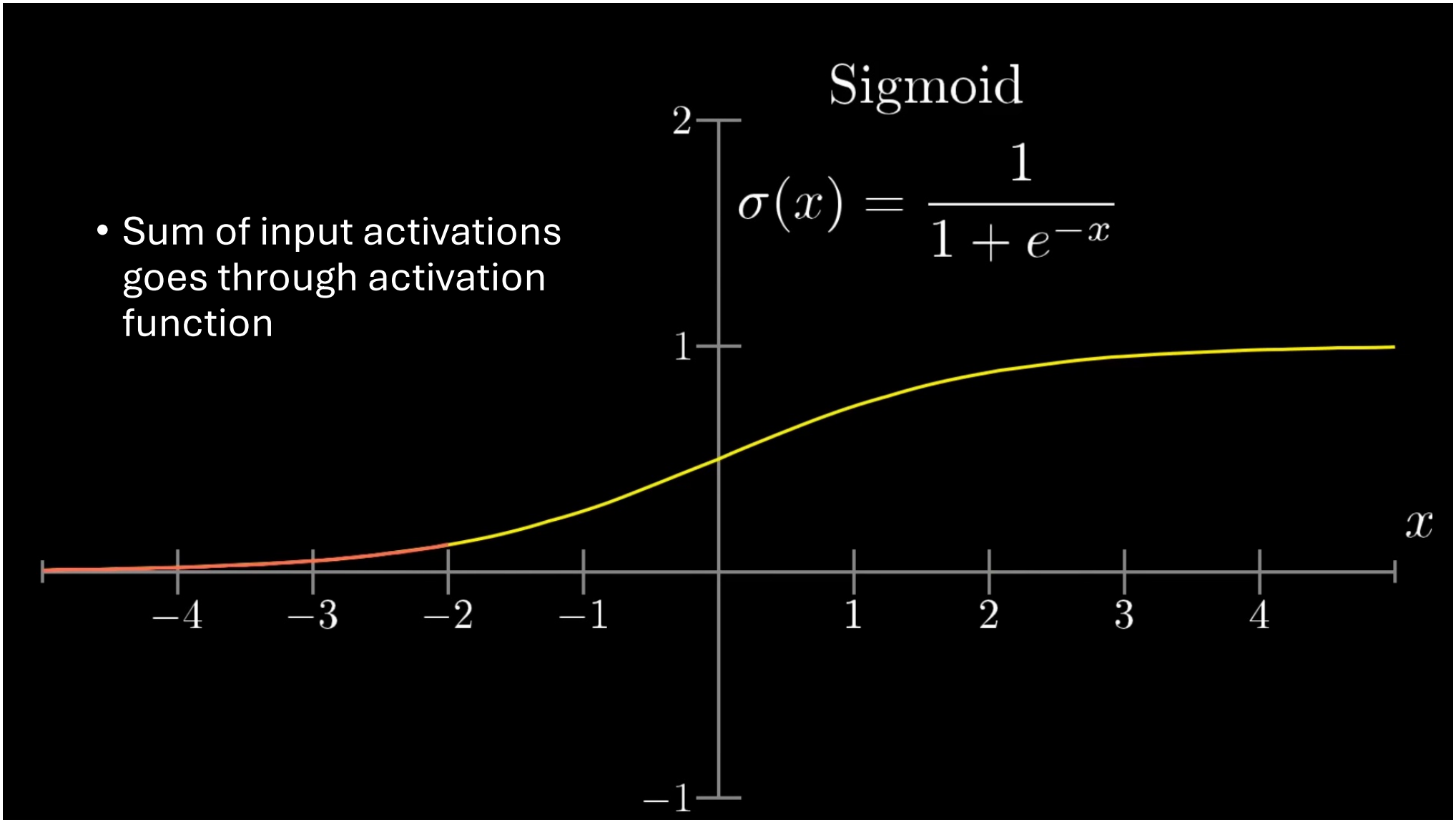
$$28 \times 28 = 784$$

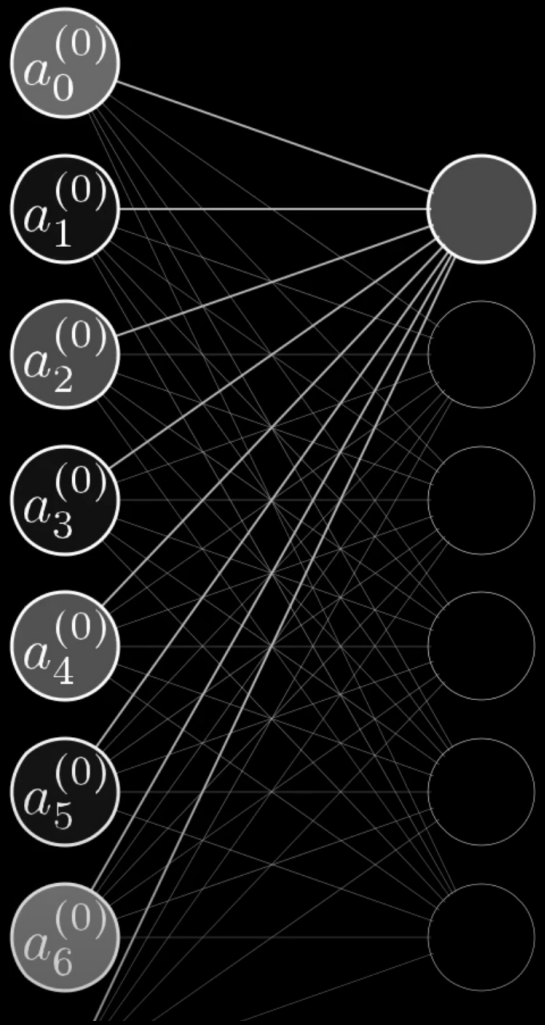

784
0
1
2
3
4
5
6
7
8
Output layer: the result.
The value of each determines how likely is the specific result?

- Activation of each layer depends on the previous layer
- The neurons are connected with *weights*
- Each layer learns ‘features’ - subcomponents



- Sum of input activations goes through activation function





Sigmoid

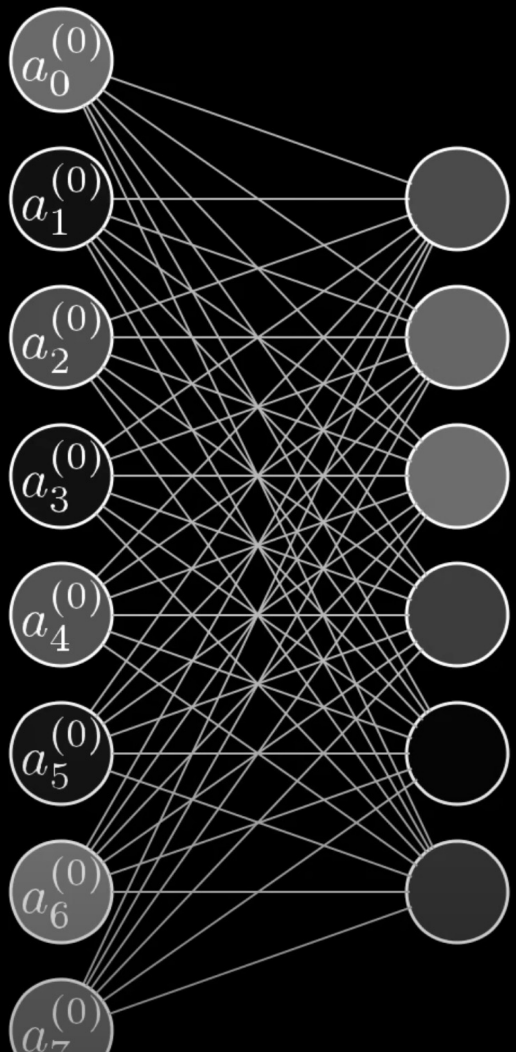
$$a_0^{(1)} = \sigma\left(w_{0,0}\ a_0^{(0)} + w_{0,1}\ a_1^{(0)} + \cdots + w_{0,n}\ a_n^{(0)} + b_0 \right)$$

Bias

Network can be understood as a single function with parameters

This small network has *13,002* parameters!

$$\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}$$



It's just repeated matrix vector products between weights and activation + bias: extremely fast and parallelizable!

$$\mathbf{a}^{(1)} = \sigma(\mathbf{W}\mathbf{a}^{(0)} + \mathbf{b})$$

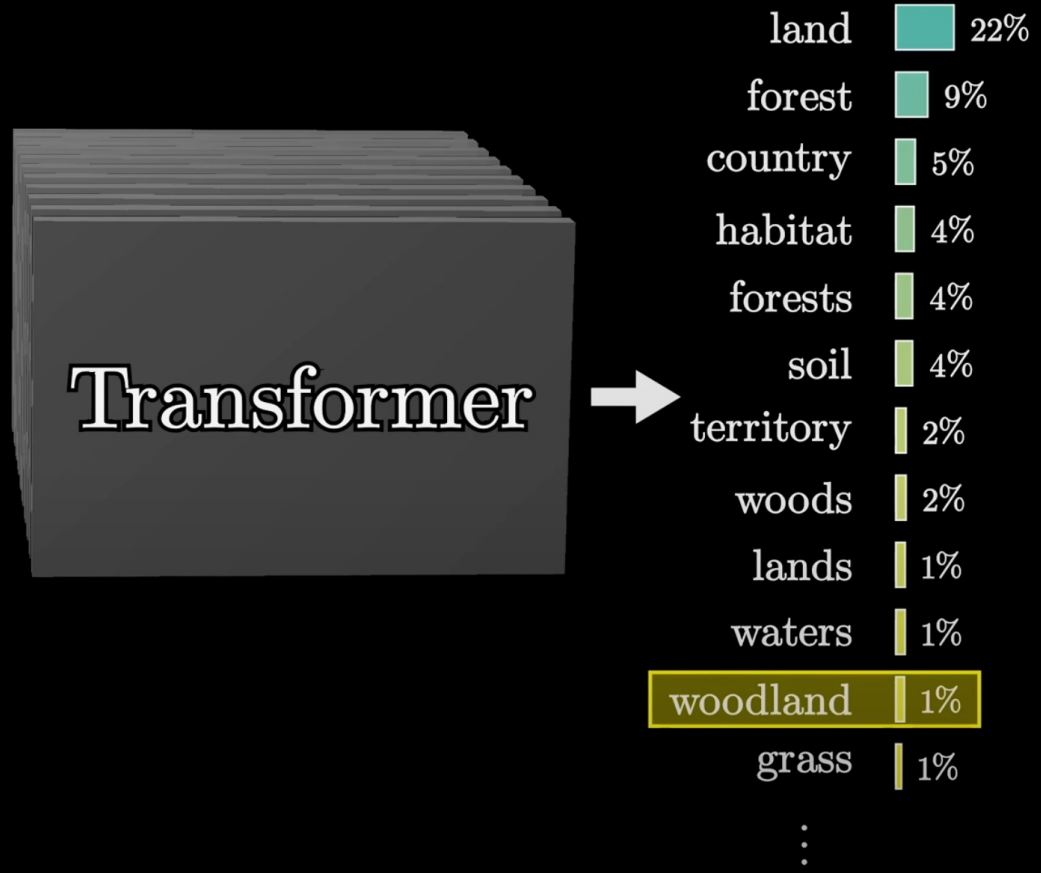
$$\sigma\left(\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}\right)$$

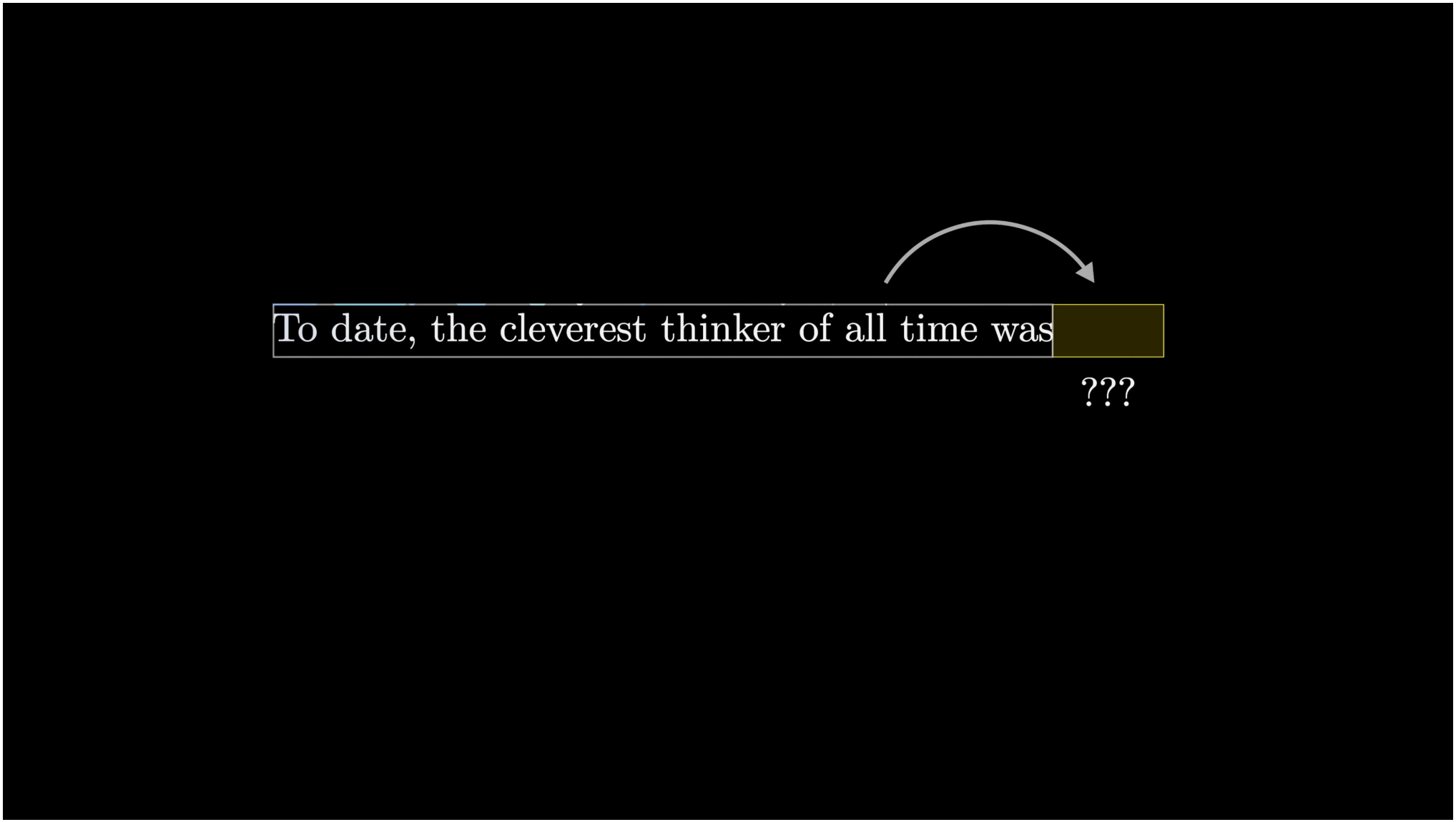
Behold, a wild pi creature, foraging in its native woodland _____

Transformers

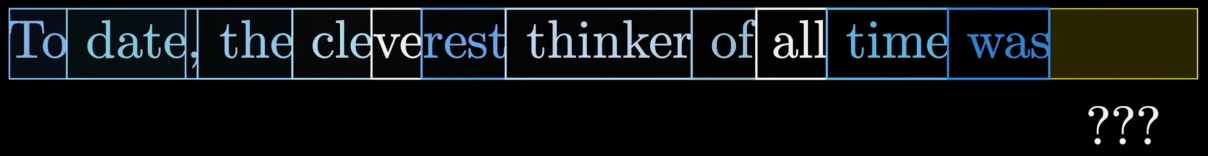
(as in GPT – Generative Pretrained *Transformers*)

- Transformers are a particular neural network structure.
- The task of predicting the next word given a word sequence.
- The easiest way of predicting the next word is to build an understanding of the world and how it works.

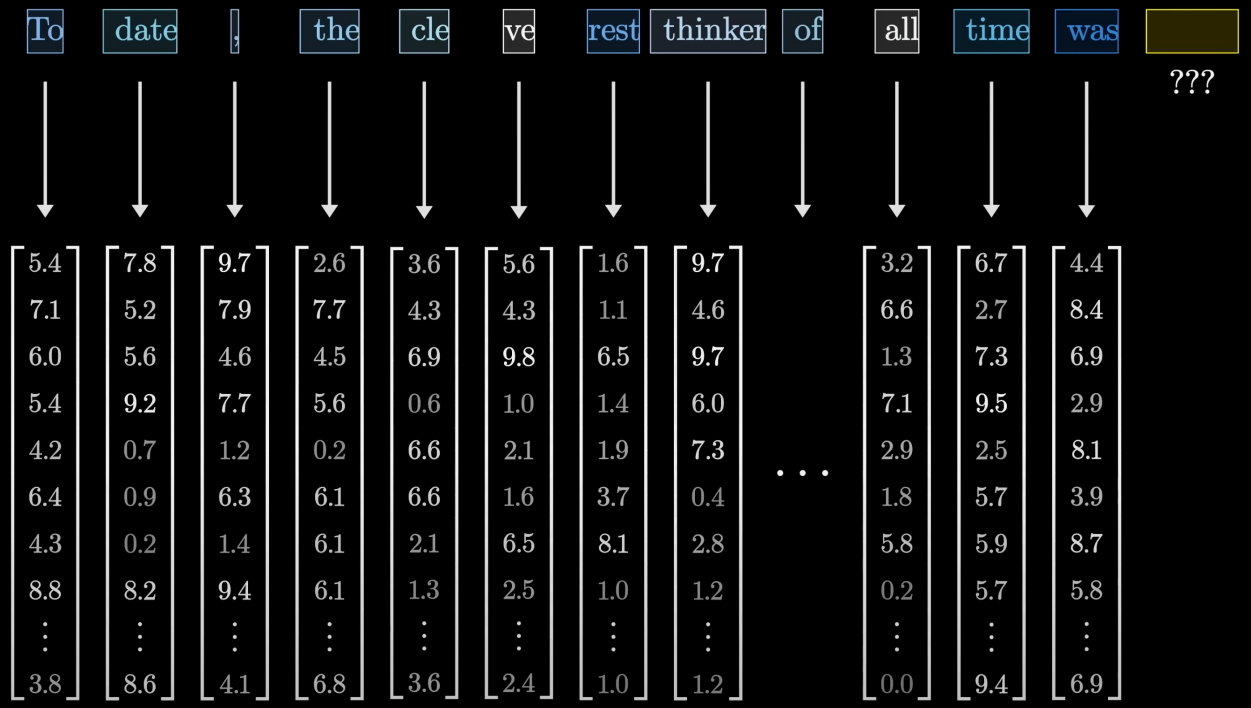



To date, the cleverest thinker of all time was
???

Tokenization

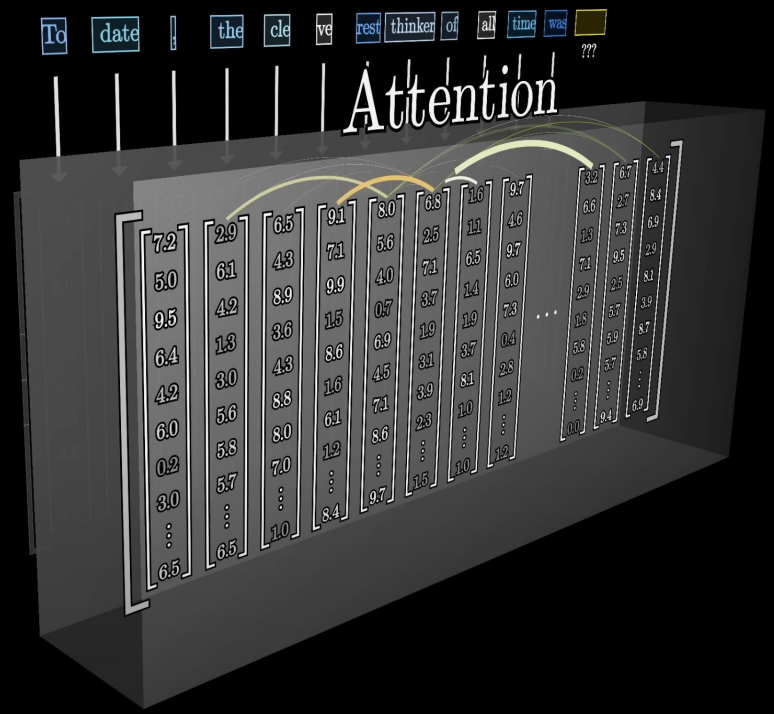


Word embeddings



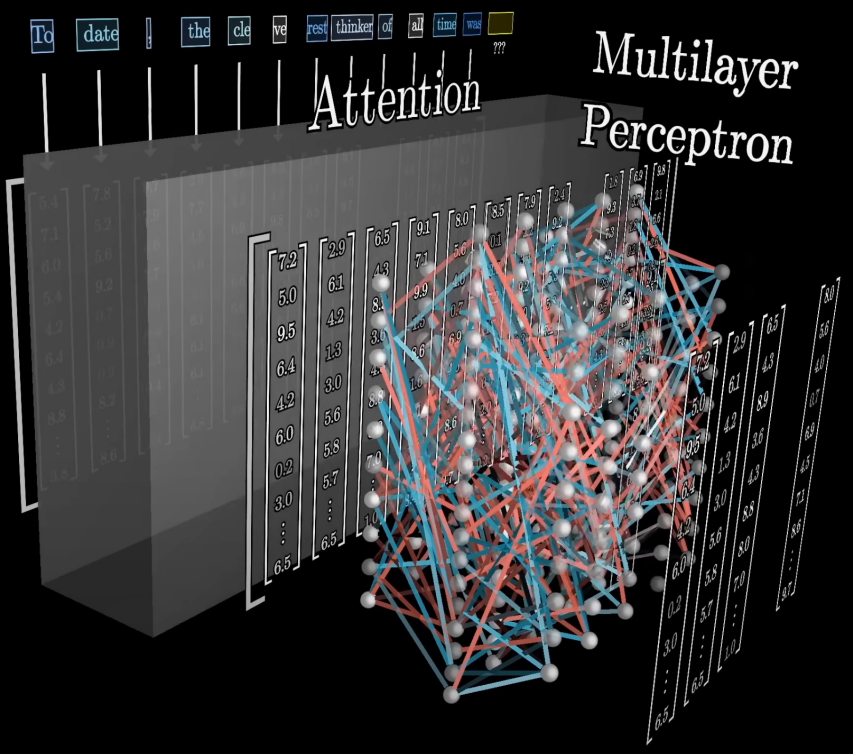
Attention-block

The words talk to each others and give meaning to one another.
Which words change the meaning of each other, and how?

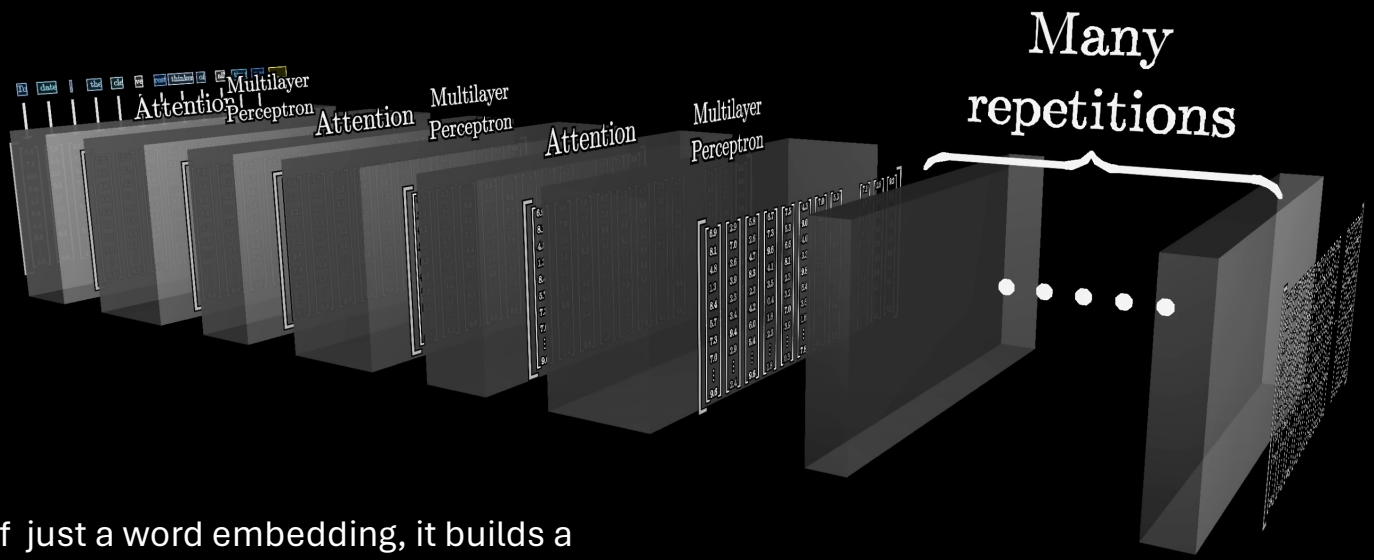


Multi-layer perceptron / Feed-forward layer

A normal neural network: does the actual processing

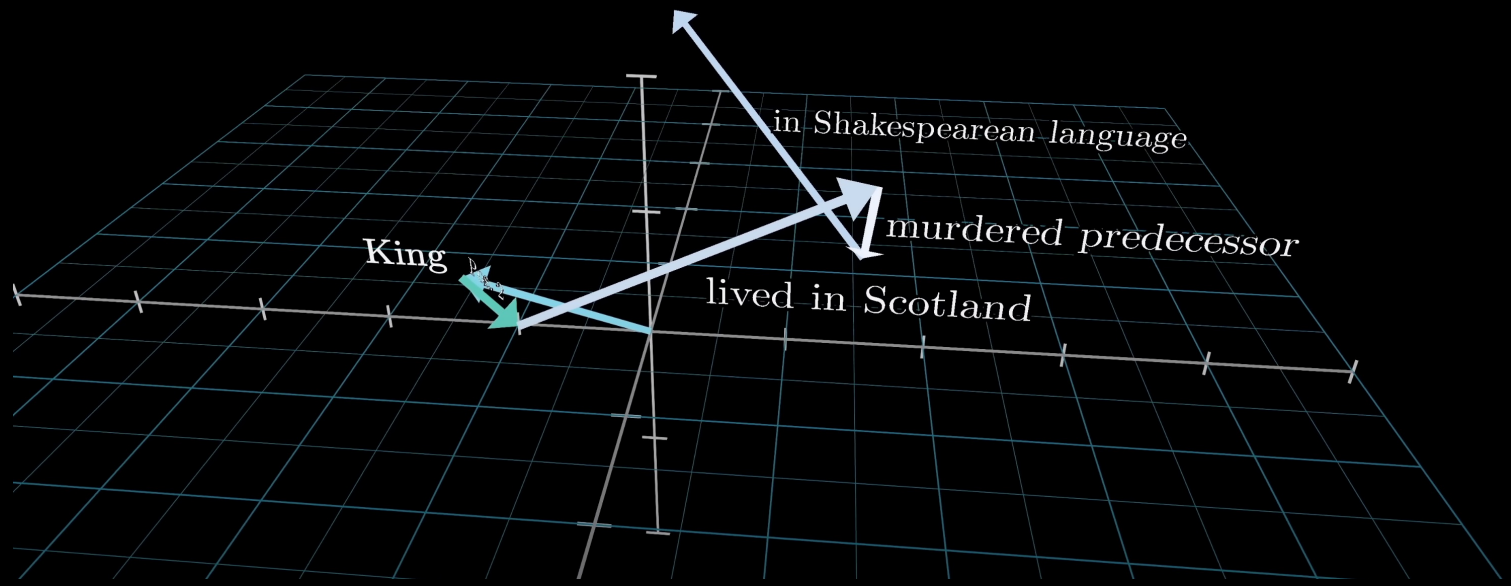


Rinse and repeat: attention x multilayer

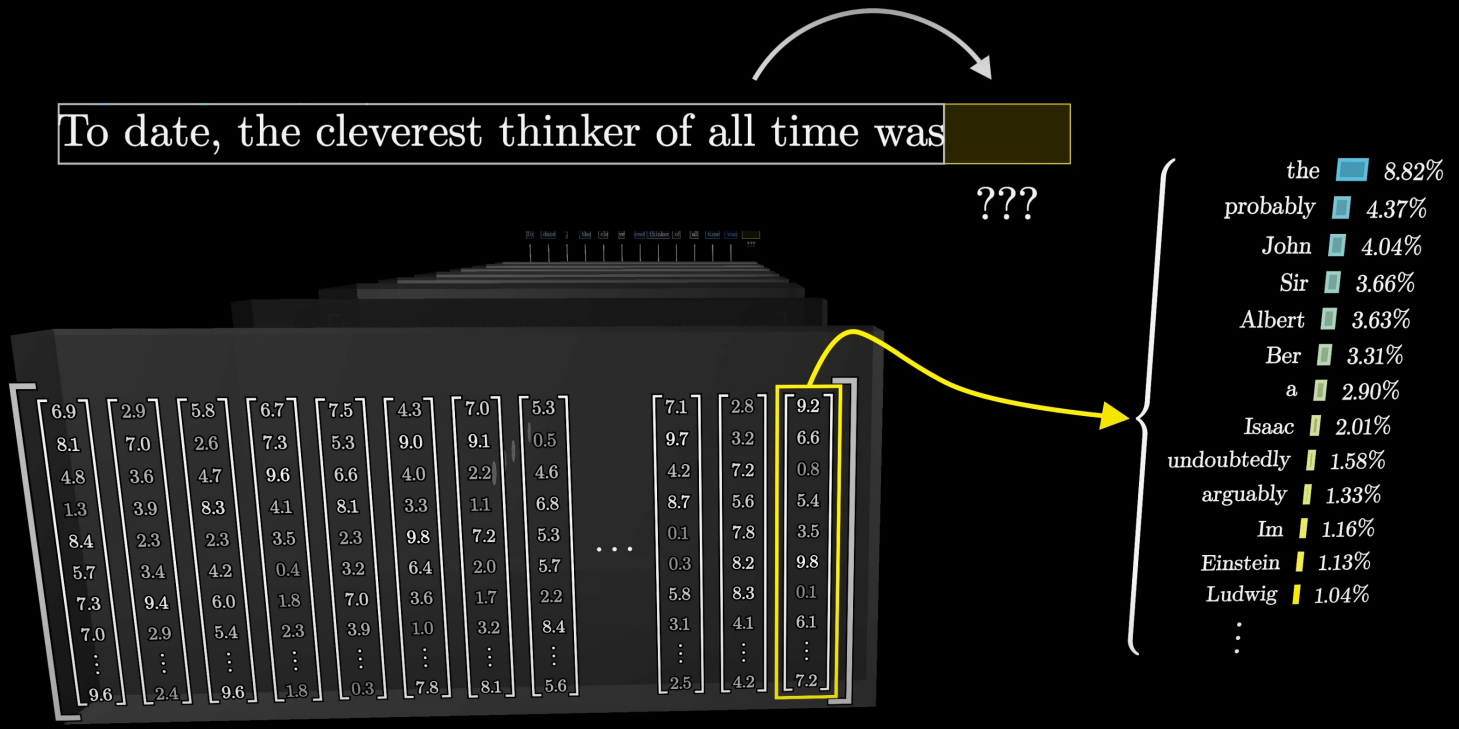


Instead of just a word embedding, it builds a much richer representation of contextual meaning

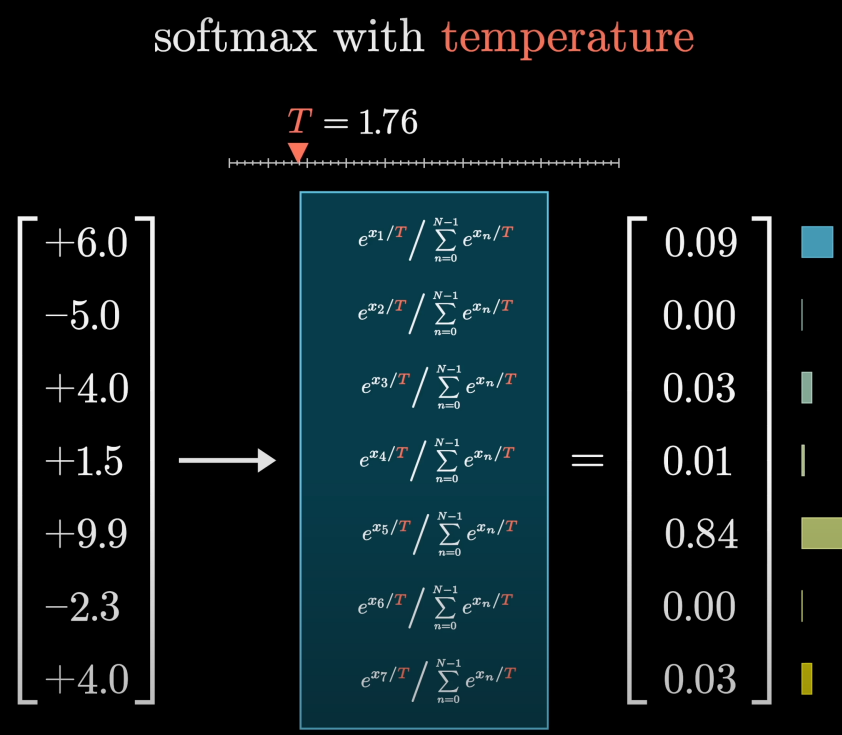
The King doth wake tonight and takes his rouse ...



At the end, the last vector is a probability distribution over words
We sample over this distribution



The sampling is done with softmax



A way of turning numbers into a probability distribution

Temperature controls evenness of the distribution

BERT (Bidirectional Encoder Representations From Transformers)

- Pre-trained to understand language and context: trained to guess next word, and to guess a masked word.
- Gives an understanding of language and context

- We can finetune it to problems we want to solve!
- Additional layers that are learned from scratch

This is the crème of supervised text models!