Universal Approximation Theorem (UAT)

The Universality Theorem

The Core Idea

Neural networks with a single hidden layer can approximate any continuous function to any desired precision.

This presentation provides a *visual*, *intuitive* proof.

The Universality Theorem

The Core Idea

Neural networks with a single hidden layer can approximate any continuous function to any desired precision.

This presentation provides a *visual*, *intuitive* proof.

Two Key Caveats

1. The theorem guarantees that such a network *exists*, not that we can *successfully train* a network to find it.
2. The network required might be very large (or even impossibly large) to achieve the desired accuracy.

The Goal: Approximate $f(x)$

We will build a network that can approximate any continuous function $f(x)$.

Our proof will be constructive. We will build the target function using three steps:

1. Create a "step function" from one hidden neuron.
2. Create a "bump function" from two hidden neurons.
3. Create the final function by adding many "bumps".

Step 1: The Hidden Neuron as a Step Function

A single hidden neuron computes $\sigma(wx + b)$, where $\sigma(z) = 1/(1 + e^{-z})$.

- By making the weight w very large ($w \to \infty$), the sigmoid becomes a near-perfect step function.
- The "step" location is controlled by the bias, b.
- Specifically, the step occurs at $x = -b/w$.

▸ Colab

Step 2: Creating a "Bump" Function

We can create a "bump" by combining two step functions.

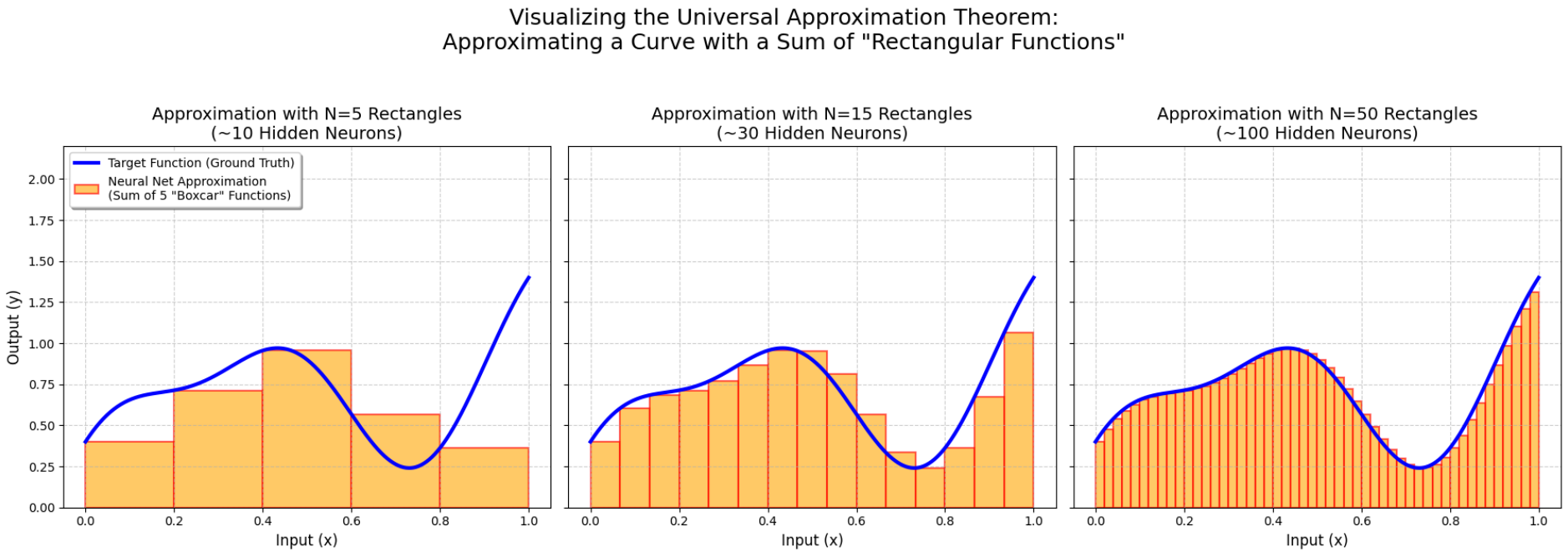
- Use two hidden neurons.
- Set their weights in the output layer to h and $-h$.
- This adds one positive step and one negative step.
- The result is a "bump" of height h in a small region.
- We can control the *location* and *height* of the bump.

▸ Colab

Step 3: Approximating any Function $f(x)$

We can approximate any function by "gluing" many bumps together.

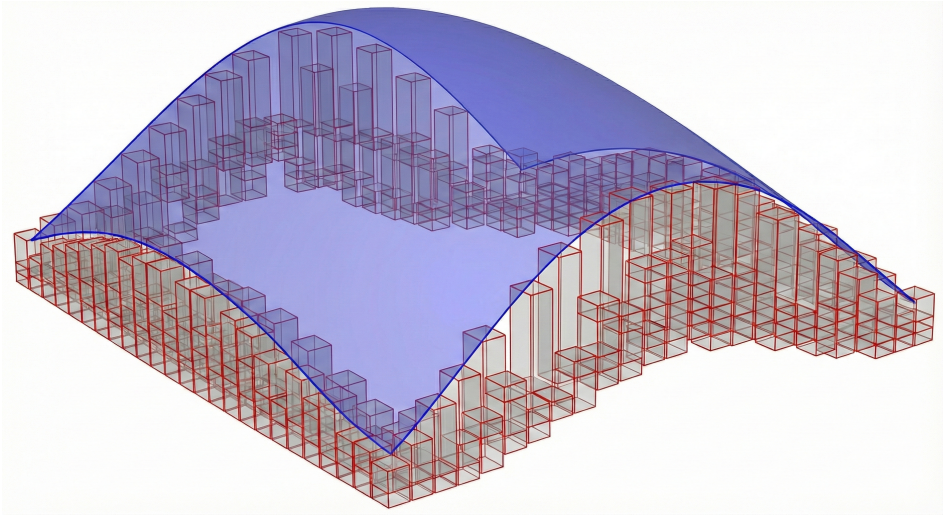
- This is just like a Riemann sum in calculus, where we use rectangles to approximate a curve.
- We use N pairs of hidden neurons to create N bumps.
- By making the bumps narrow and numerous, we can approximate $f(x)$ to any desired accuracy ϵ.



Extending to Many Inputs: $f(x, y, \dots)$

The same logic extends to multiple dimensions.

- For 2D, $f(x, y)$, we design a neuron that steps in one direction (e.g., the x direction).
- We can combine two such neurons to make a "ridge."
- By combining two ridges (one in x, one in y), we can build a 2D "tower."
- This tower is the 2D equivalent of our 1D "bump."



By adding many of these "towers" of different heights and locations, we can approximate any 2D surface $f(x, y)$. This generalizes to n dimensions.

What about other Activation Functions?

The proof does not just apply to the sigmoid function.

General "Sigmoids"

Universality holds for *any* non-linear activation function that is "S-shaped" (i.e., has a lower and upper plateau).

What about other Activation Functions?

The proof does not just apply to the sigmoid function.

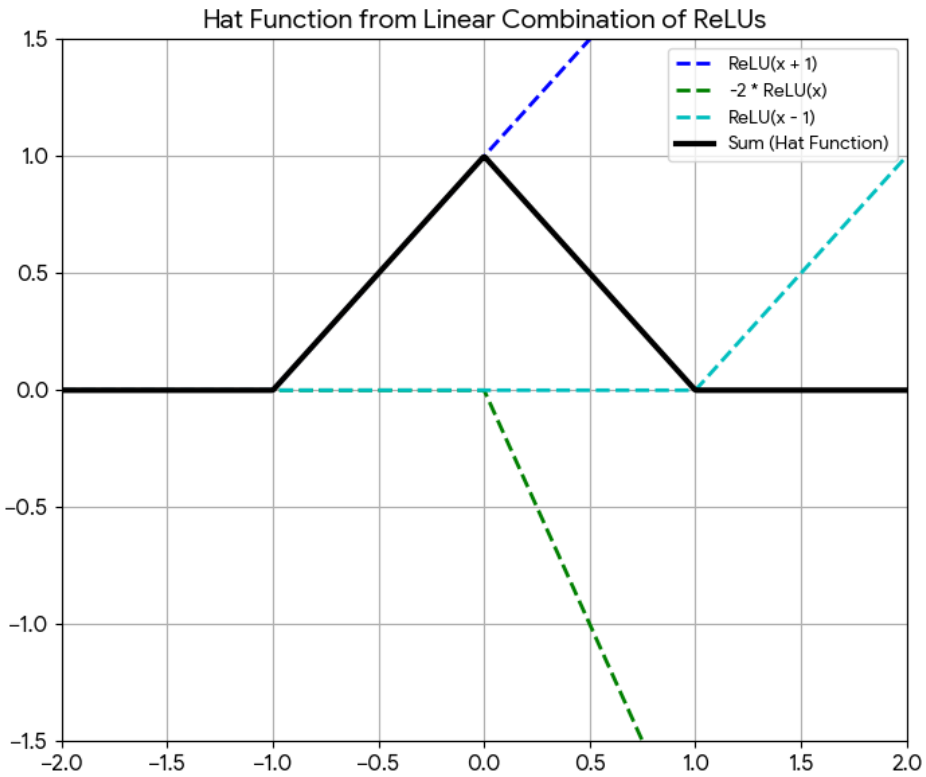
General "Sigmoids"

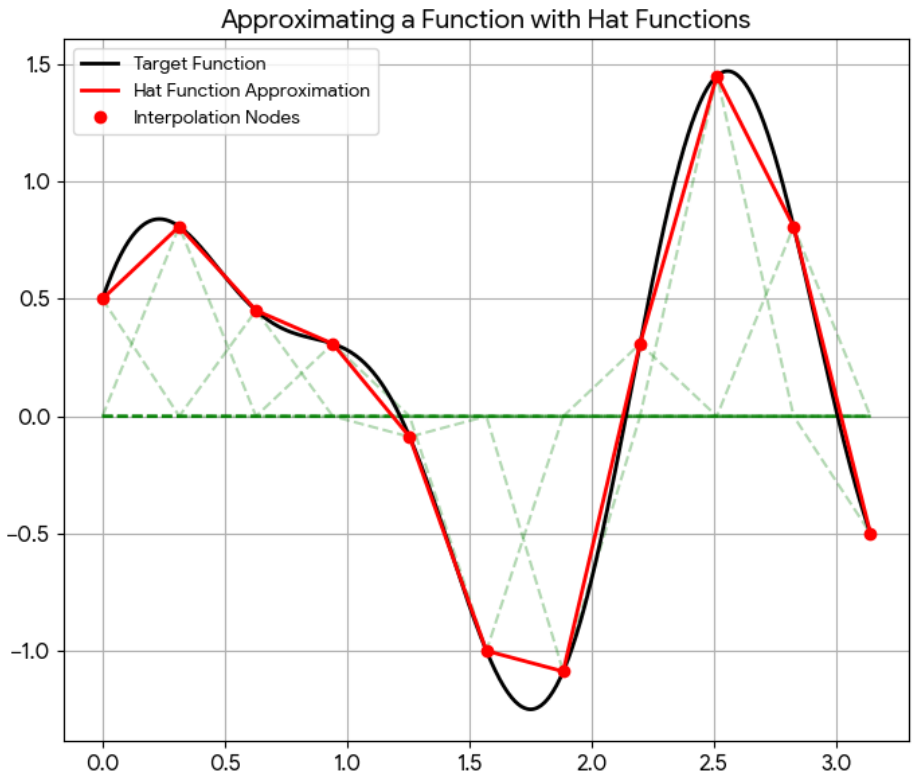
Universality holds for *any* non-linear activation function that is "S-shaped" (i.e., has a lower and upper plateau).

What about ReLU (Rectified Linear Unit)?

- ReLU is $f(z) = \max(0, z)$.
- *Universality still holds.*
- The proof is similar: you can use ReLUs to create "hat" functions, which can then be summed to approximate any function (connection with FEM).
- In practice, ReLUs are now more common than sigmoid neurons in hidden layers.

ReLU case





▸ Colab

Formal proof of the UAT. The Setup

Let's define the framework for approximation.

- Let x be the input variable and z be the target.
- The target function is denoted by $z = f(x)$ belonging to a certain function space S with a distance d.
- We define the unit hypercube input space: $I_n = [0, 1]^n$.

Definition (Universal Approximator)

We say that a neural network is a ***universal approximator*** *for the space (S, d) if the space of possible network outcomes U is d-dense in S. That is:*

$$\forall f \in S, \forall \epsilon > 0, \exists g \in U : d(f - g) < \epsilon$$

Activation Functions: Sigmoidal

The core of a neural network is its activation function.

Definition (Sigmoidal Function)

A function $\sigma : \mathbb{R} \to [0, 1]$ *is called* ***sigmoidal*** *if:*

$$lim_{r \to -\infty} \sigma(x) = 0$$

and

$$lim_{x \to \infty} \sigma(x) = 1$$

Typical examples include the logistic sigmoid or tanh functions used traditionally in neural networks.

Activation Functions: Discriminatory Property

Definition (n-discriminatory)

Let n be a natural number. An activation function $f : \mathbb{R} \to \mathbb{R}$ is ***n-discriminatory*** *if the only signed Borel measure μ such that*

$$\int f(y \cdot x + \theta) d\mu(x) = 0, \quad \forall y \in \mathbb{R}^n, \theta \in \mathbb{R},$$

is the zero measure ($\mu = 0$).

Definition (Discriminatory)

An activation function f is ***discriminatory*** *if it is n-discriminatory for any n.*

Remark

A discriminatory function is "volumetrically non-destructive" when acting on linear transformations of input.

Relevant Spaces

To prove the theorem, we need tools from Functional Analysis.

- Let K denote a compact set in $\mathbb{R}^n$ (e.g., I_n).
- $C(K)$ denotes the set of real-valued continuous functions on K.
- $M(I_n)$ is the space of finite signed regular [1] (Borel) measures on I_n.

Theorem (Riesz Representation Theorem)

Let F be a bounded linear functional on $C(K)$. Then there exists a unique finite signed Borel measure μ on K such that:

$$F(f) = \int_K f(x) d\mu(x), \quad \forall f \in C(K).$$

[1] *Regular* means that while different measures may define different sizes for a single set, they all ideally convey some idea of how much space that set takes up relative to the larger space in which it resides

The Hahn-Banach Theorem and its consequnces I

Theorem

Le X be a linear real vector space, X_0 a linear subspace, p a linear convex functional on X, and $f : X_0 \to \mathbb{R}$ a linear functional such that $f(x) \leq p(x)$ for all $x \in X_0$.
Then there is a linear functional $F : X \to \mathbb{R}$ such that:

1. $F_{X_0} = f$ *(the restriction of F to X_0 is f).*
2. $F(x) \leq p(x)$ *for all $x \in X$.*

The Hahn-Banach theorem allows us to separate spaces using linear functionals.

Lemma (1. Reformulation of Hahn-Banach separation)

*Let U be a linear, **non-dense** subspace of a normed linear space X. Then there is a bounded linear functional L on X such that $L \neq 0$ on X and $L_{|U} = 0$ (L vanishes on U).*

The Hahn-Banach Theorem and its consequnces II

We apply this specifically to our space $C(I_n)$:

Lemma (2. The Crucial Link)

Let U be a linear, non-dense subspace of $C(I_n)$. Then there is a measure $\mu \in M(I_n)$ (where $\mu \neq 0$) such that

$$\int_{I_n} h d\mu = 0, \quad \forall h \in U.$$

Proof Idea: Use Lemma 1 to find functional L vanishing on U, then use Riesz Representation to represent L as an integral against a measure μ.

The Universal Approximation Theorem

We are now ready to state the main result regarding standard feedforward neural networks.

Proposition (Density of Neural Network Outputs)

Let σ be any continuous ***discriminatory*** *function. Then the finite sums of the form:*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + \theta_j)$$

are dense in $C(I_n)$. Here $w_j \in \mathbb{R}^n, \theta_j \in \mathbb{R}, \alpha_j \in \mathbb{R}$.

Note: $G(x)$ represents the output of a neural network with a single hidden layer containing N neurons, weights w_j, biases θ_j, and output weights α_j.

Proof Strategy: Contradiction

We proceed by contradiction.

- Let U be the set of all possible network outputs: $U = \{G; G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + \theta_j)\}$ U is a linear subspace of $C(I_n)$.
- **Assumption:** Assume U is **not** dense in $C(I_n)$.
- By Lemma 2, there must exist a non-zero measure $\mu \in M(I_n)$ such that it is orthogonal to all functions in U:

$$\int_{I_n} h d\mu = 0, \quad \forall h \in U$$

Completing the Proof

The condition $\int h d\mu = 0$ for all $h \in U$ means:

$$\int \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + \theta_j) d\mu = 0$$

for all choices of parameters.
By choosing convenient coefficients (e.g., $N = 1, \alpha_1 = 1$), we obtain:

$$\int_{I_n} \sigma(w^T x + \theta) d\mu = 0, \quad \forall w \in \mathbb{R}^n, \theta \in \mathbb{R}$$

Conclusion: However, the hypothesis of the proposition states that σ is a **discriminatory** function. By the definition of discriminatory, this condition implies that $\mu = 0$. This contradicts the finding from Lemma 2 that $\mu \neq 0$. Therefore, the assumption that U is not dense must be false. U is dense in $C(I_n)$. □

Connecting Sigmoids to the Proof

The previous proof relies entirely on σ being "discriminatory". Do standard activation functions satisfy this?

Proposition

*Any continuous **sigmoidal** function is discriminatory for all measures $\mu \in M(I_n)$.*

To prove this, we need a geometric lemma about measures.

Definition (Geometry)

Define the hyperplane $P_{w,\theta} = \{x; w^T x + \theta = 0\}$ and half-spaces $H^+_{w,\theta} = \{x; w^T x + \theta > 0\}$ and $H^-_{w,\theta} = \{x; w^T x + \theta < 0\}$.

Lemma (3)

Let $\mu \in M(I_n)$. If μ vanishes on all hyperplanes (P) and open half-spaces (H) in $\mathbb{R}^n$, then μ is zero.

Sketch: Sigmoidal $\Longrightarrow$ Discriminatory I

Assume $\int_{I_n} \sigma(w^T x + \theta) d\mu(x) = 0$ for a fixed μ. We want to show $\mu = 0$.

1. Construct scaled functions $\sigma_\lambda(x) = \sigma(\lambda(w^T x + \theta) + \phi)$.
2. As $\lambda \to \infty$, using sigmoidal properties, σ_λ converges pointwise to a step-like function $\gamma(x)$:

$$\gamma(x) = \begin{cases} 1, & \text{if } x \in H^+_{w,\theta} \\ 0, & \text{if } x \in H^-_{w,\theta} \\ \sigma(\phi), & \text{if } x \in P_{w,\theta} \end{cases}$$

3. By Bounded Convergence Theorem, we swap limit and integral:

$$\lim_{\lambda \to \infty} \int \sigma_\lambda d\mu = \int \gamma d\mu = \mu(H^+_{w,\theta}) + \sigma(\phi)\mu(P_{w,\theta}) = 0$$

Hence we have established that: $\mu(H^+_{w,\theta}) + \sigma(\phi)\mu(P_{w,\theta}) = 0$ holds for any ϕ.

Sketch: Sigmoidal $\implies$ Discriminatory II

4. Taking limits on ϕ (using sigmoidal properties $\sigma(\phi) \to 0$ or 1):

$$\phi \to +\infty \implies \mu(H^+_{w,\theta}) + \mu(P_{w,\theta}) = 0$$

$$\phi \to -\infty \implies \mu(H^+_{w,\theta}) = 0$$

5. These imply μ vanishes on all half-spaces H^+ (and H^- by symmetry) and hyperplanes P.
6. By Lemma 3, a measure vanishing on all such sets must be the zero measure ($\mu = 0$). Therefore, σ is discriminatory.

The ReLU I

Modern networks often use Rectified Linear Units (ReLU), which are not sigmoidal. Do they still work?

Proposition

The ReLU function is 1-discriminatory.

Proof. *Idea:*Let us assume that $y \in \mathbb{R}$ and $\theta \in \mathbb{R}$:

$$\int ReLU(yx + \theta)d\mu(x) = 0.$$

We want to show that $\mu = 0$.

The ReLU II

1. We construct a sigmoid bounded, continuous (and therefore measurable) function from subtracting two ReLU functions with different parameters. In particular, consider the function

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \in [0, 1] \\ 1 & \text{if } x > 1 \end{cases}$$

2. Then any function of the form $g(x) = f(yx + \theta)$ with $y \neq 0$ can be described as

$$g(x) = ReLU\left(yx + \theta_1\right) - ReLU\left(yx + \theta_2\right)$$

by setting $\theta_1 = -\theta/y$ and $\theta_2 = (1 - \theta)/y$. If $y = 0$, then instead set

$$g(x) = f(\theta).$$

The ReLU III

3. Which means that for any $y \in \mathbb{R}, \theta \in \mathbb{R}$

$$\begin{aligned}\int f(yx+\theta)d\mu(x) &= \int \left(ReLU\left(yx+\theta_1\right) - ReLU\left(yx+\theta_2\right)\right) d\mu(x) \\ &= \int ReLU\left(yx+\theta_1\right) d\mu(x) - \int ReLU\left(yx+\theta_2\right) d\mu(x) \\ &= 0 - 0 = 0. \quad (1)\end{aligned}$$

4. By the previous lemma, f is discriminatory, and therefore, $\mu = 0$.

Scaling Dimensions

The ReLU proof showed it was 1-discriminatory. We need it to be n-discriminatory.

Definition

Let $\Sigma_n(f) = span\{f(y \cdot x + \theta) | y \in \mathbb{R}^n, \theta \in \mathbb{R}\}$.

Proposition (Dimensional Lifting)

If $\Sigma_1(f)$ *is dense in* $C([0,1])$ *then* $\Sigma_n(f)$ *is dense in* $C([0,1]^n)$.

Proof Idea: The span of ridge functions $\{g(a \cdot x) | g \in C([0,1])\}$ is known to be dense in $C([0,1]^n)$. Since $\Sigma_1(f)$ is dense in $C([0,1])$, we can approximate any $g(a \cdot x)$ arbitrarily well using sums of $f(y_i \cdot x + \theta_i)$. Applying the triangle inequality shows we can approximate any function in $C([0,1]^n)$.

Conclusion

- The Universal Approximation Theorem guarantees that a neural network with a single hidden layer can approximate any continuous function on a compact set to arbitrary precision.
- The proof relies heavily on functional analysis, specifically the relationship between continuous functions and measures (Riesz Representation) and separation theorems (Hahn-Banach).
- The crucial condition for the activation function is that it must be **discriminatory**—it cannot "annihilate" any non-zero measure.
- Both classical Sigmoidal functions and modern ReLU functions (via construction) satisfy this property, explaining their theoretical success.

Conclusion: Universality is Not Everything

The theorem is profound, but it has practical limits.

- **Existence vs. Training:** Just because a shallow network *can* do the job doesn't mean our training algorithms (like gradient descent) can *find* it.
- **Efficiency:** The theorem doesn't say a shallow network is the *most efficient* way.

Shallow vs. Deep Networks

Deep networks may be far more efficient at learning complex, real-world problems. The hierarchical structure (layers of layers) seems better suited for learning a "hierarchy of concepts" (e.g., pixels $\rightarrow$ edges $\rightarrow$ shapes $\rightarrow$ faces).

Thank You

Questions?

All content and images from *Neural Networks and Deep Learning* by Michael A. Nielsen.