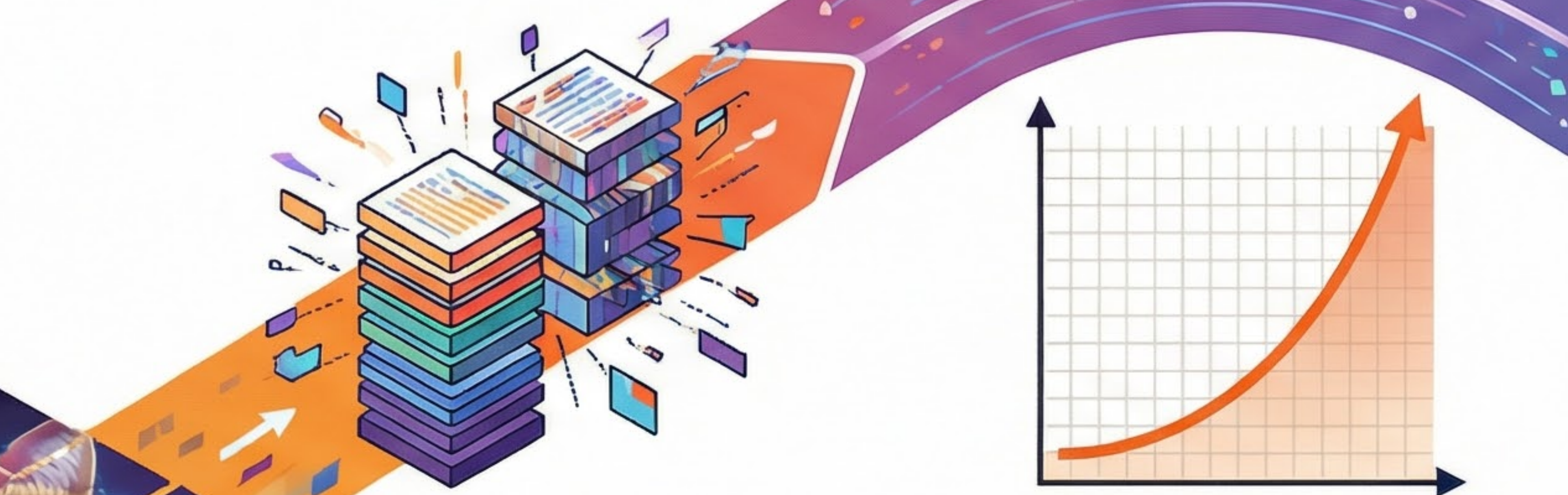
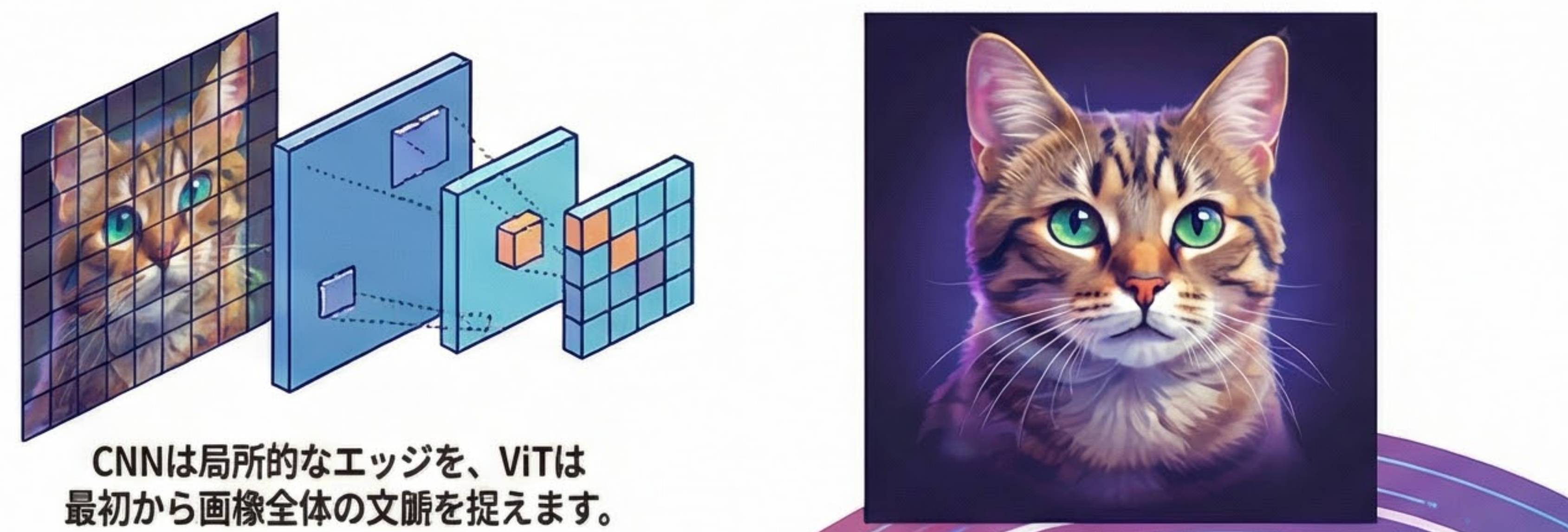


視覚AIの進化系譜：CNNからTransformer、そして次世代アーキテクチャへ（2012-2026）

2010年代を支配した「局所的な特徴」を捉えるCNNから、2020年のViT登場により「大域的な文脈」を重視するTransformerへと主役が交代しました。現在は生成AI（DIT）や、計算効率を極めた次世代モデル（Mamba）が融合・競合する新時代に突入しています。

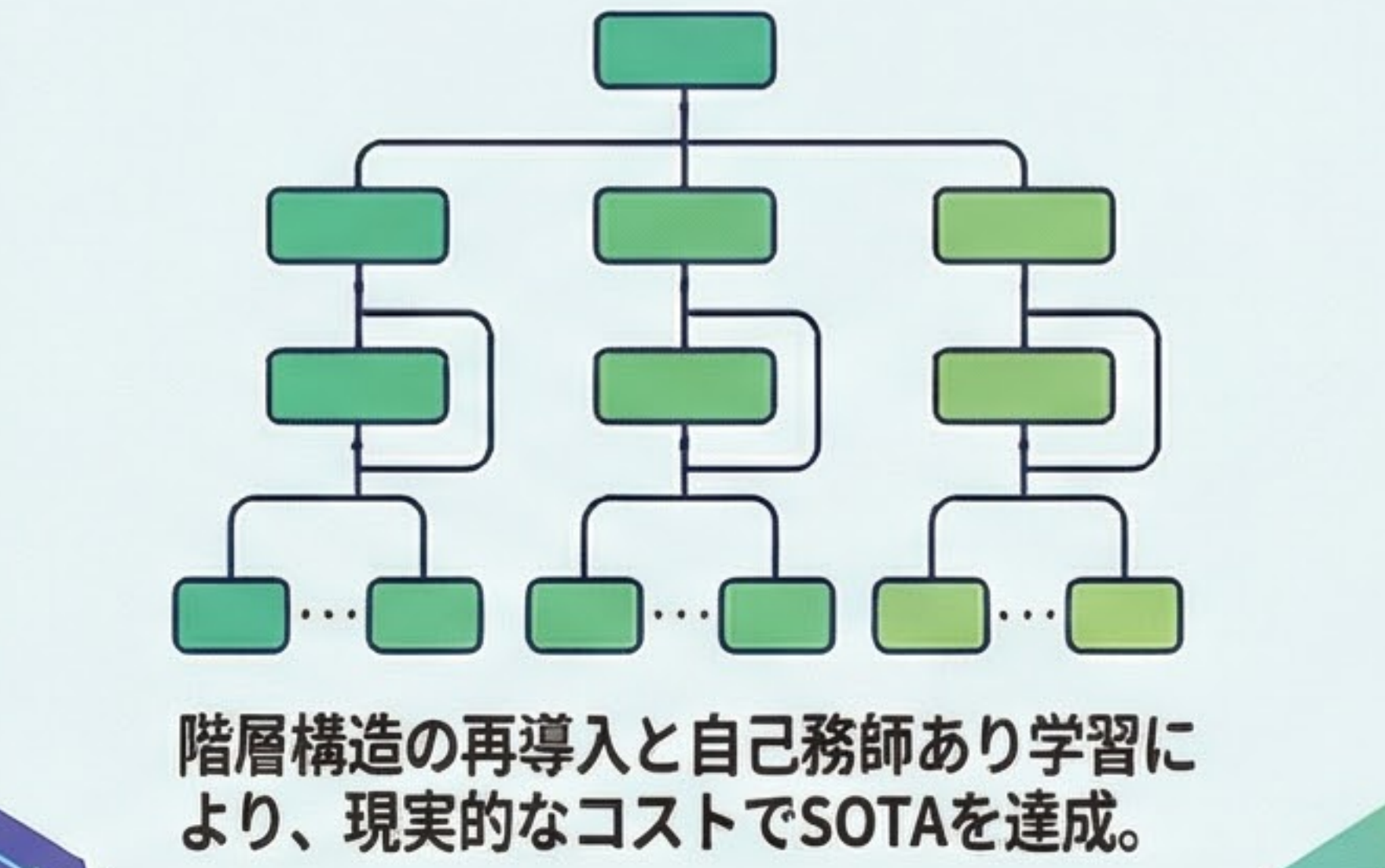
2012-2020：CNNの覇権からTransformerの衝撃へ

CNNの「帰納的バイアス」 vs ViTの「大域的受容野」



2021-2026：実用化・生成AI・そして次世代効率化

Swin & MAEによる実用性とデータ効率の劇的向上



Mambaの台頭：線形計算量 $O(N)$ への挑戦

Transformerの弱点である計算量を克服し、超高解像度や長尺動画の処理を可能に。



DiTと「時空パッチ」が起こした動画生成革命

動画を3次元の塊（パッチ）として処理し、物理法則の近似すら可能に。

主要アーキテクチャにおける計算効率と得意領域の比較

CNN (ResNet等)	Transformer (ViT)	Mamba (Vim)
$O(N)$	$O(N^2)$	$O(N)$
エッジデバイス、リアルタイム操出	高精度な図案、基盤モデル、生成AI	超高解像度画像、長尺動画、3D医療画像