NVIDIA

AI FOR MEDIA AND ENTERTAINMENT

August 2018

NVIDIA

DEEP LEARNING – AN OVERVIEW

Rick Grandy & Gary Burnett

Solutions Architects – ProViz Media & Entertainment

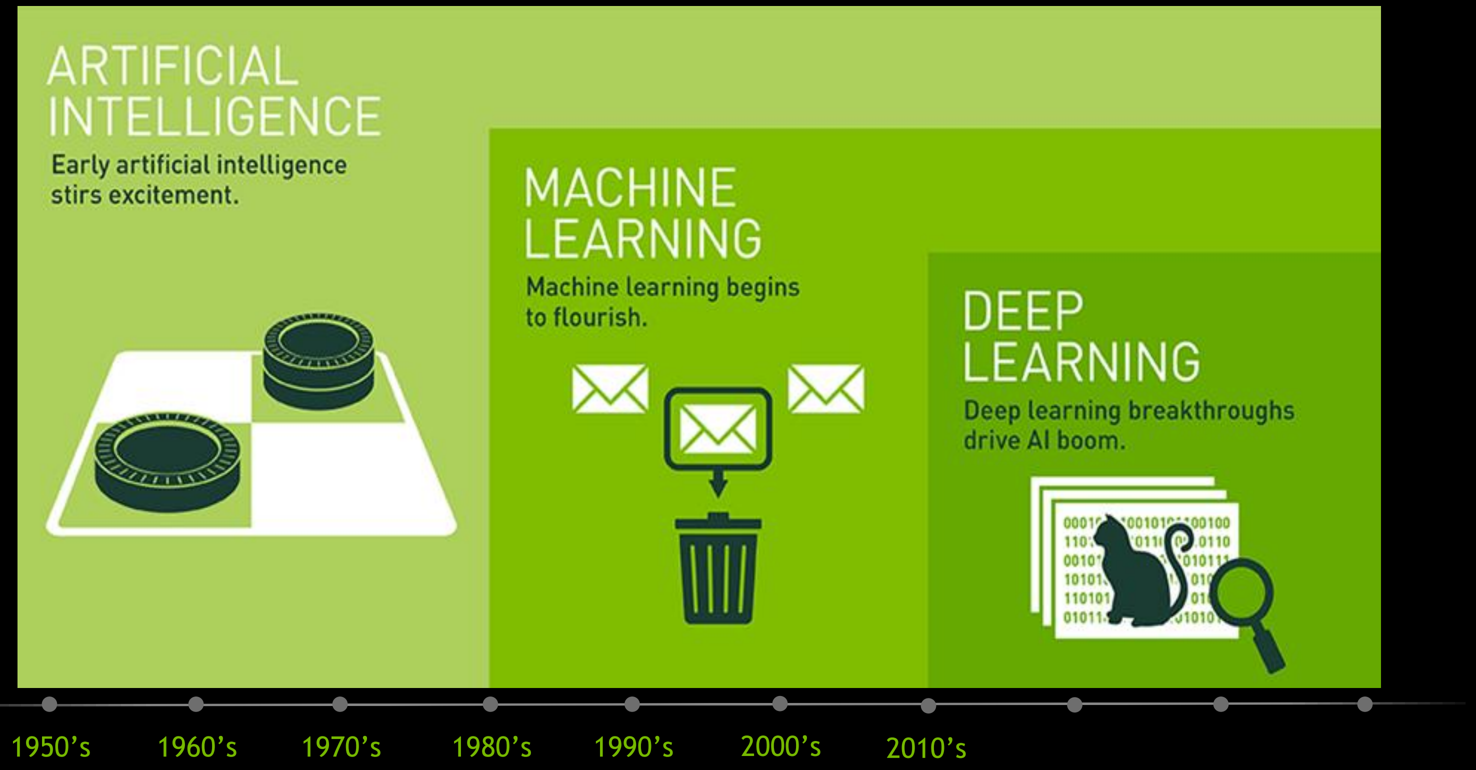
EVOLUTION OF ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI):
general coverall for machines doing interesting things

Machine Learning (ML):
computers complete tasks without explicit programming

Neural Networks (NN):
one technique to achieve ML

Deep Learning (DL):
adds "hidden layers" to Neural Networks to solve complex problems



Great explanation:
https://goo.gl/hkayWG

A NEW COMPUTING MODEL

Algorithms that learn from examples

MACHINE LEARNING

TRADITIONAL APPROACH

Requires domain experts
Time-consuming experimentation
Custom algorithms
Not scalable to new problems

Car

Vehicle

Coupe

A NEW COMPUTING MODEL

Algorithms that learn from examples

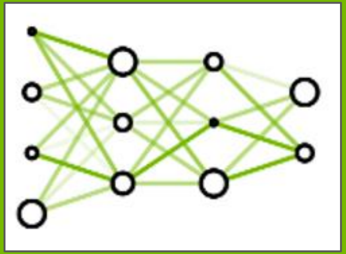
MACHINE LEARNING

TRADITIONAL APPROACH

Requires domain experts
Time-consuming experimentation
Custom algorithms
Not scalable to new problems

Car
Vehicle
Coupe

DEEP LEARNING



DEEP NEURAL NETWORKS

Learn from data
Easily to extend
Accelerated with GPUs

Car
Vehicle
Coupe

DEEP LEARNING 101

DEEP LEARNING 101

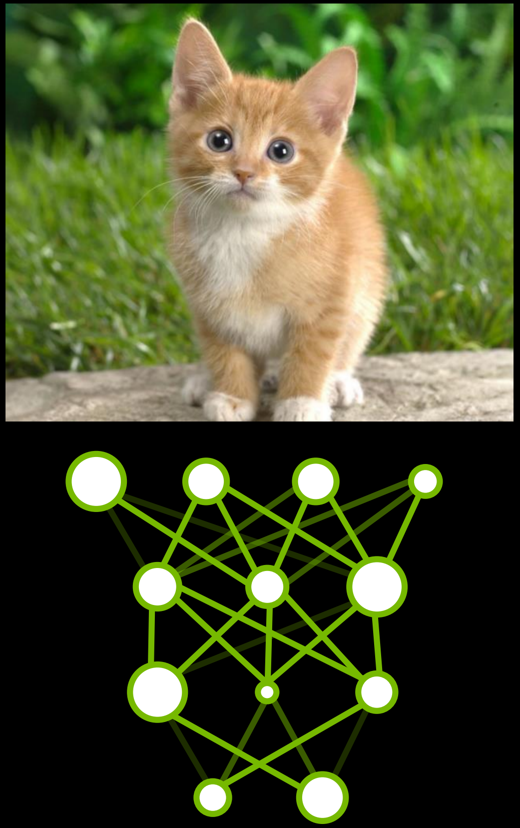
its all about the data

WHAT PROBLEM ARE YOU SOLVING?

Defining the AI/DL Task

INPUTS	QUESTION	AI/DL TASK	EXAMPLE OUTPUTS
Text Data Images Video Audio	Is "it" present or not?	Detection	Object Detection
	What type of thing is "it"?	Classification	Object Identification (Labeling)
	To what extent is "it" present?	Segmentation	Feature Tracking
	What is the likely outcome?	Prediction	Denoised Pixel Values
	What will likely satisfy the objective?	Recommendation	Animation Pose Selection
	What would be a new variant?	Generation	Texture Creation

CLASSIFICATION

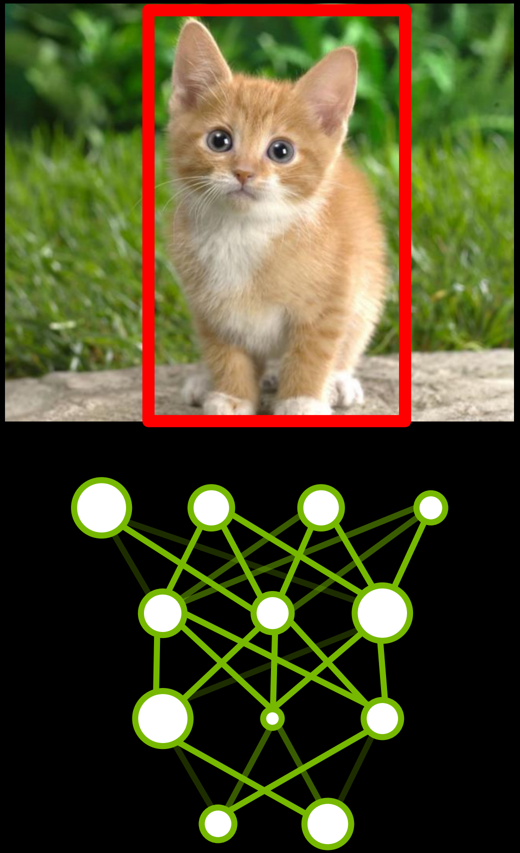


0.3 0.7

P_{DOG} P_{CAT}

Probabilities for each class

OBJECT DETECTION

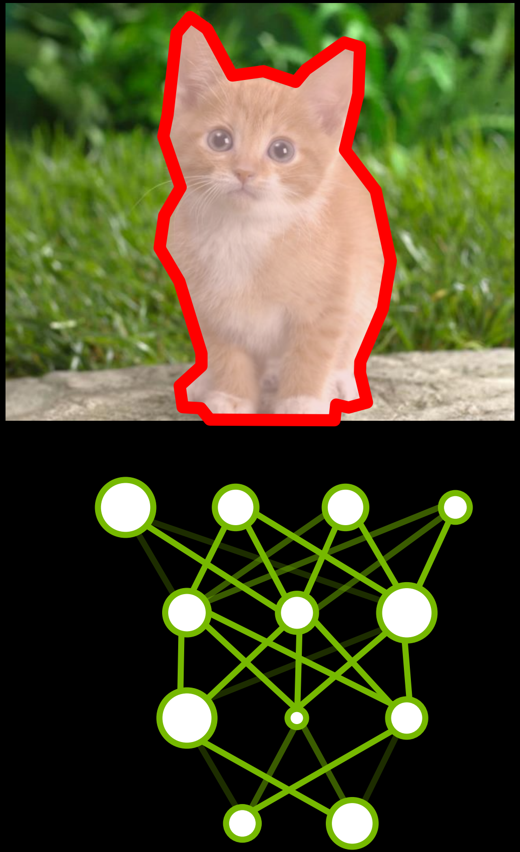


(10, 100)

(X_1, Y_1)

Corners of a bounding box

SEGMENTATION

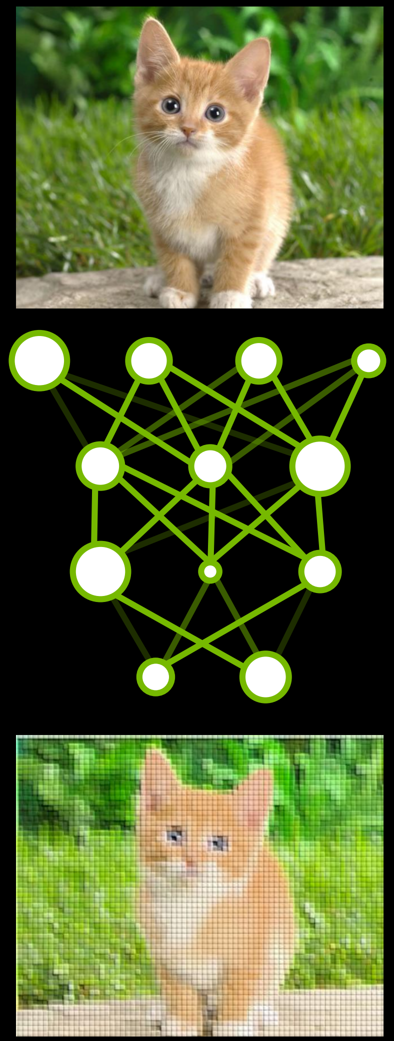


(5, 120)

(X_1, Y_1)

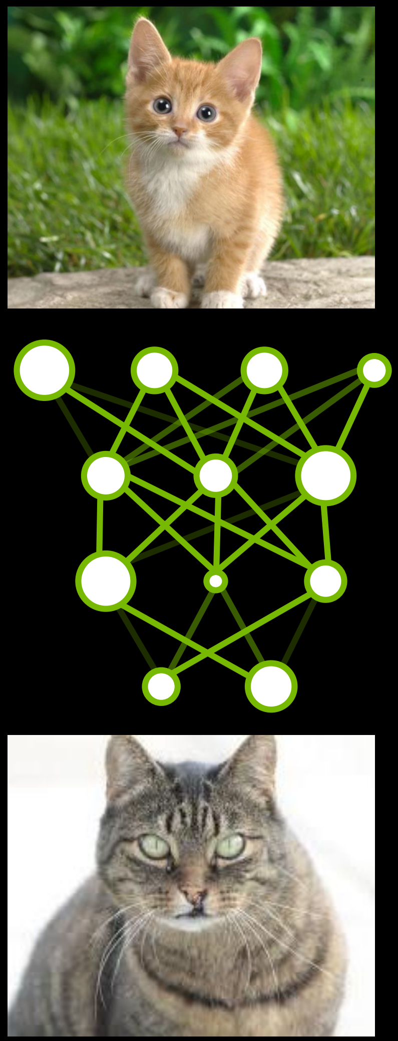
Pixels that belong to the cat

PREDICTION



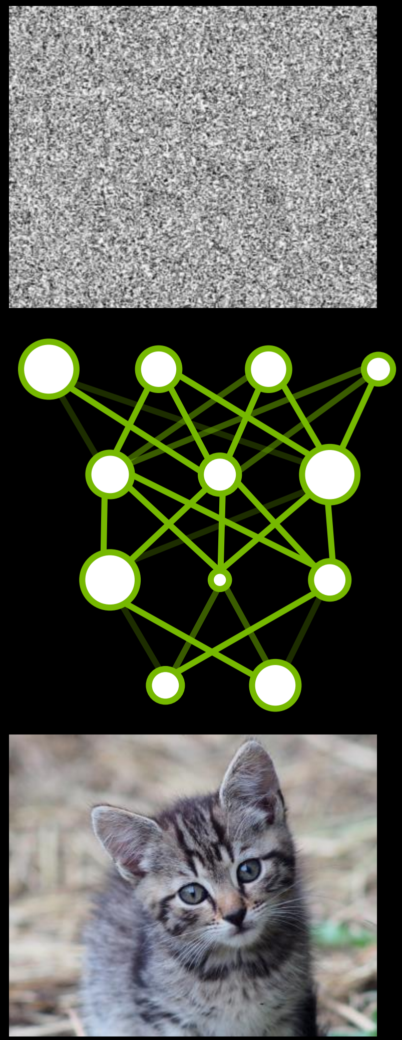
Predict how quilted cat looks

RECOMMENDATION



Recommend other cats

GENERATION



Generate cats from nothing

DEEP LEARNING 101

Application Development

DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING

Learning a new capability
from existing data

INFERENCE

Applying this capability
to new data

DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING

Learning a new capability
from existing data

INFERENCE

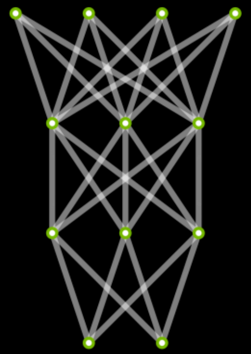
Applying this capability
to new data

DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING
Learning a new capability
from existing data

INFERENCE
Applying this capability
to new data

Untrained
Neural Network
Model



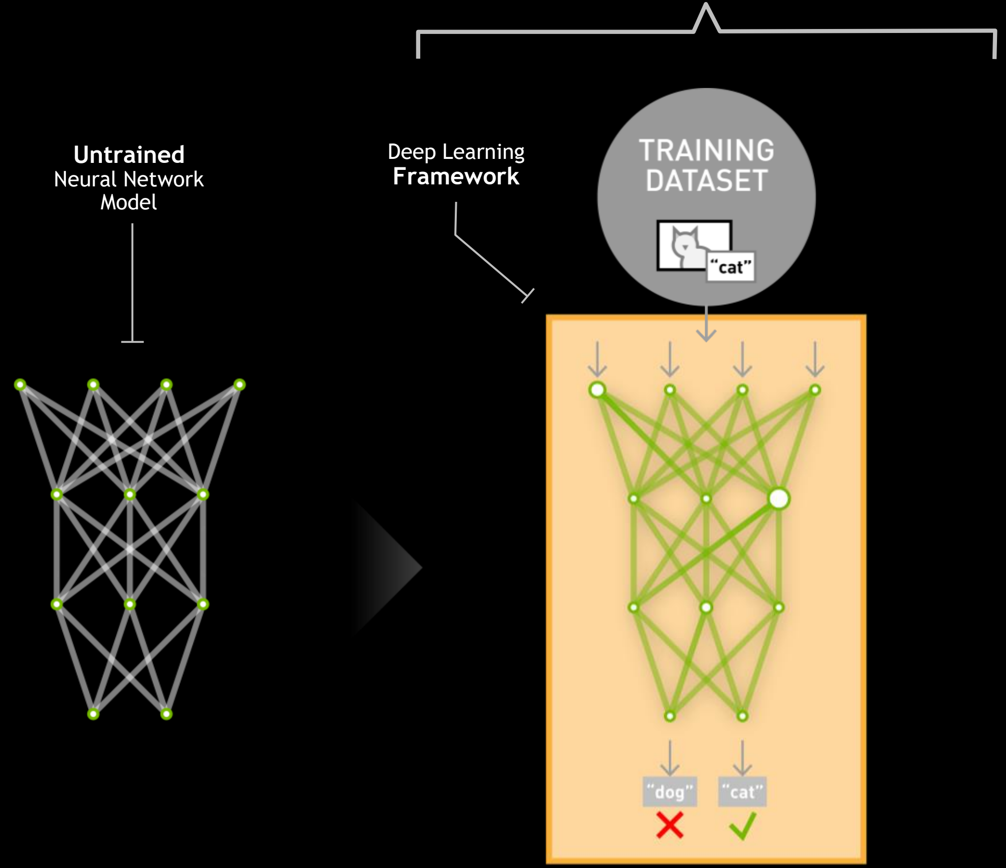
DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING

Learning a new capability from existing data

INFERENCE

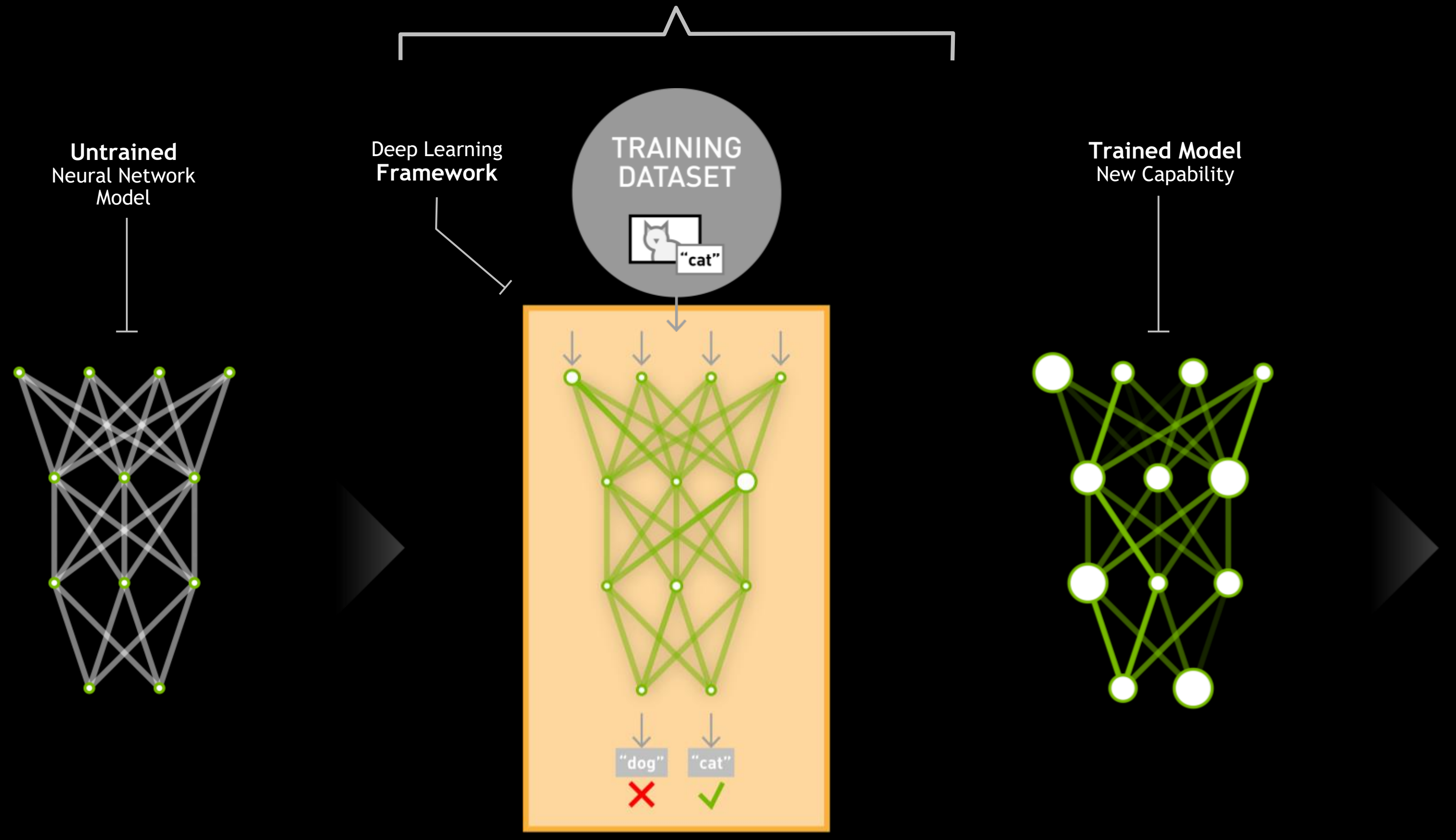
Applying this capability to new data



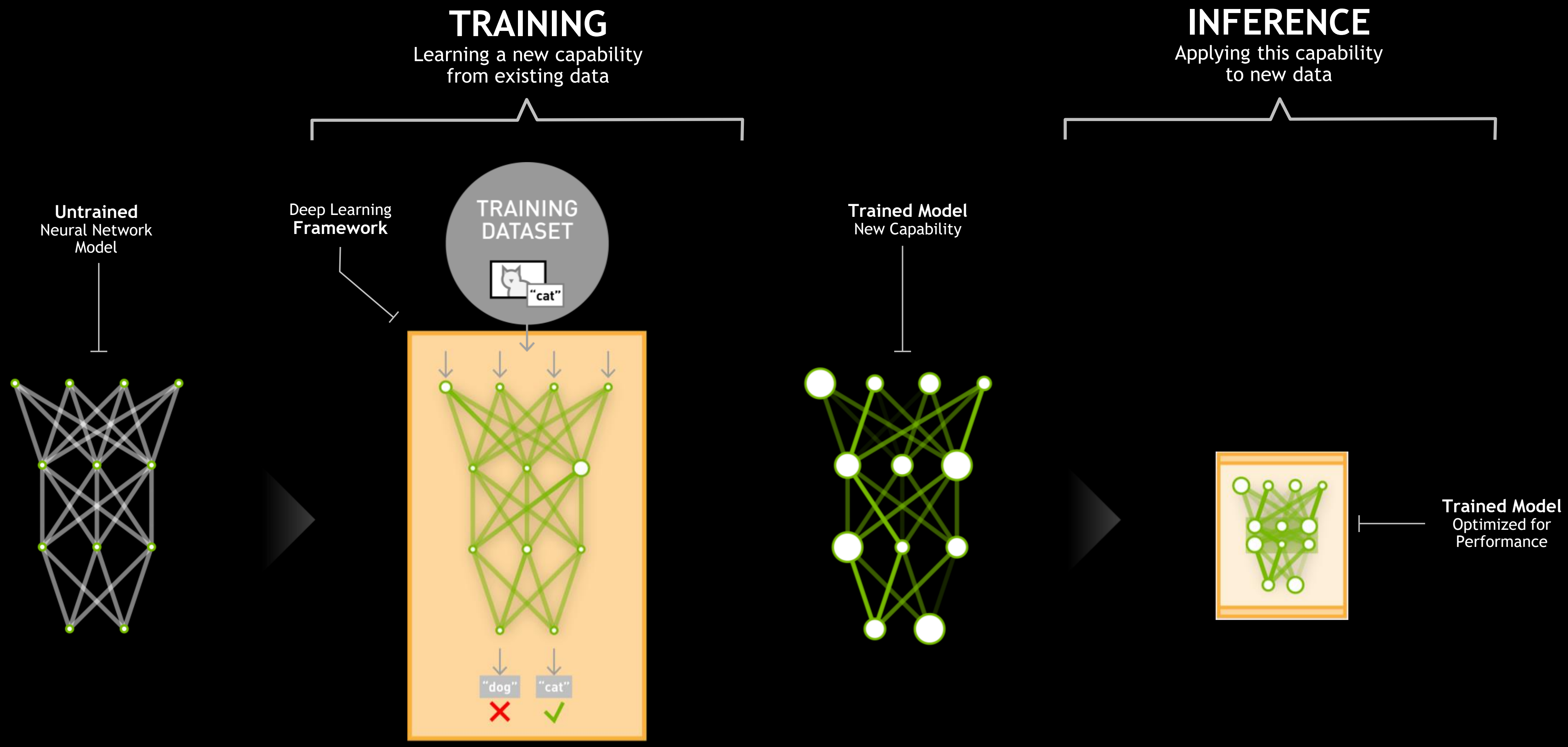
DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING
Learning a new capability
from existing data

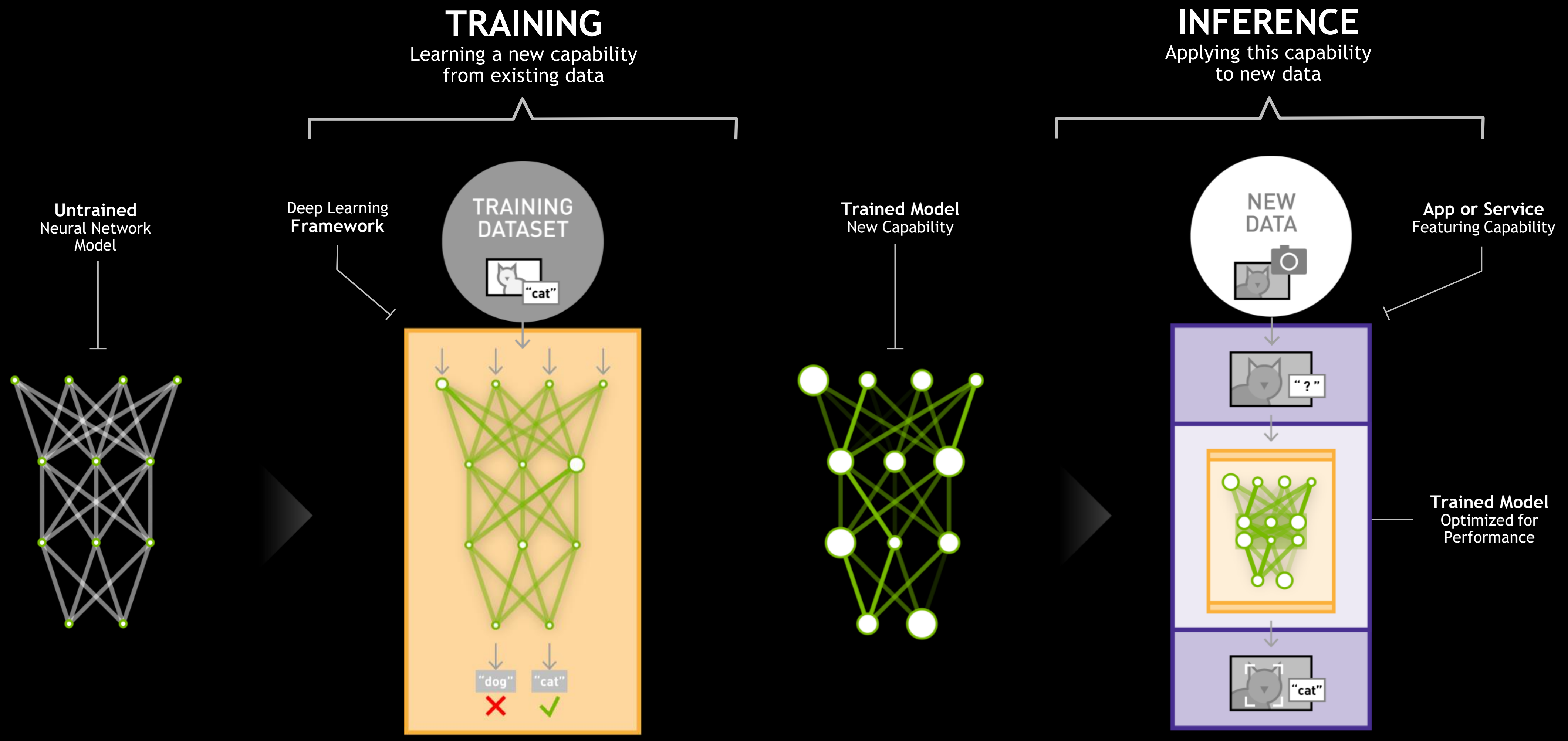
INFERENCE
Applying this capability
to new data



DEEP LEARNING APPLICATION DEVELOPMENT



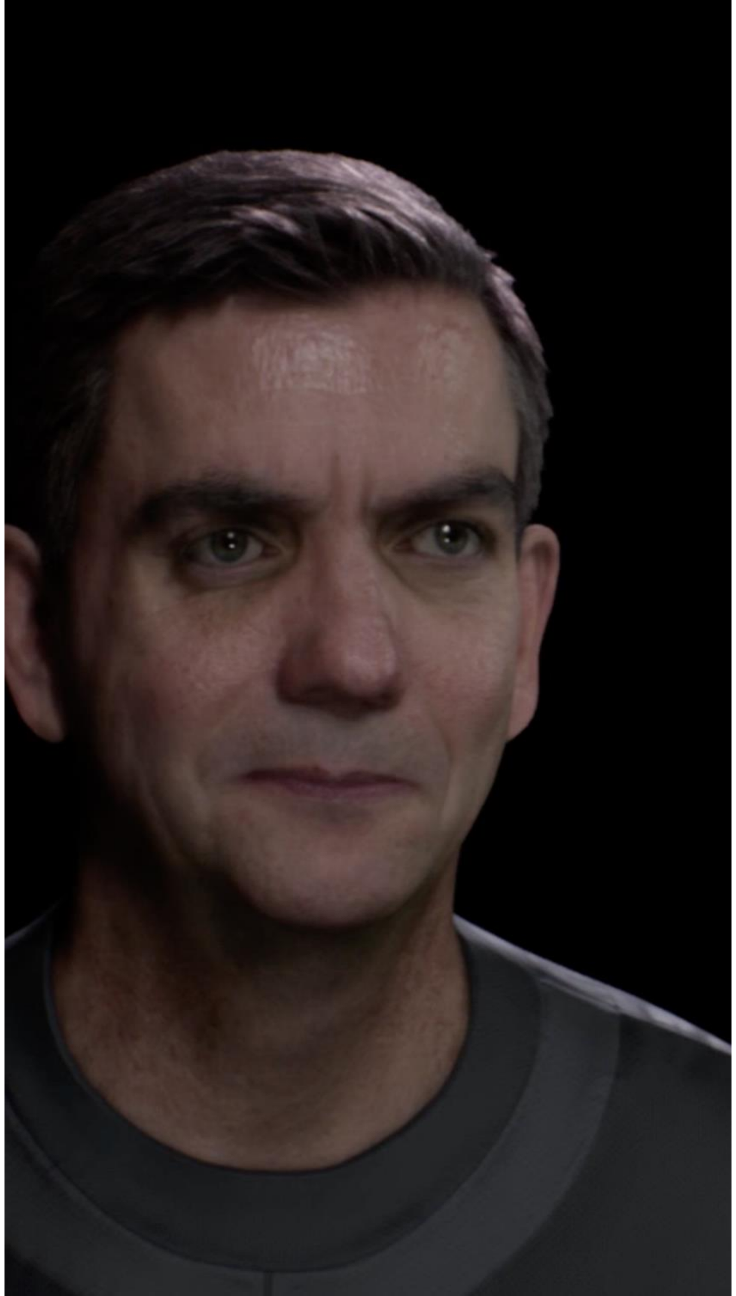
DEEP LEARNING APPLICATION DEVELOPMENT



AI FOR MEDIA AND ENTERTAINMENT

Digital Domain
NVIDIA
2018

DEEP LEARNING CHANGES EVERYTHING



VFX + MACHINE = LEARNING

- Image Processing
  - (Nearly) Automatic Rotoscoping
  - Noise Removal
- Character Animation
  - Facial Animation
  - Muscle Simulation
  - Hair Simulation
- Effects
  - Fluid Simulation
  - FEM Simulation

CAN YOU INTERPOLATE AN EFFECT?

EXAMPLE: FACIAL ANIMATION

TRADITIONAL APPROACHES

- Bones and skin deformation
- Blendshapes and FACS poses

GET A MACHINE TO LEARN THE FACE

Input

Great Big Non-Linear Interpolator

Or

A Great Big Mapping from one probability distribution to another

Output

“Torture the data, and it will confess to anything.”

-Ronald Coase

“My face is tired.”

-Doug Roble

TEMPORALLY CONSISTENT HIGH-RESOLUTION MOVING MESHES

WE HAVE DATA.

NOW WHAT?

POINTS TO HIGH-RESOLUTION MESH

- Built on work by Bermano et al (2014) and Bickel et al (2008)
- Lucio Moser created Masquerade (2017)
- A data-driven method to take tracked points to a high-resolution mesh.

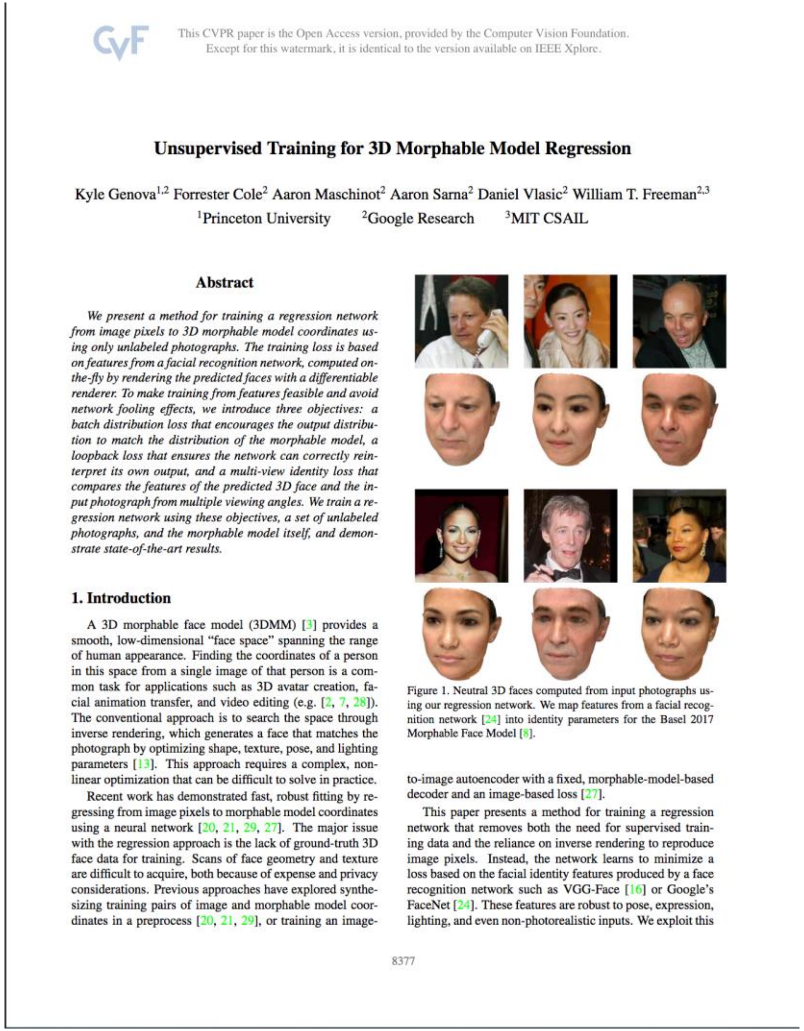
IMAGE TO HIGH RESOLUTION MESH

- Images (no markers) as input
- High resolution mesh as output

- Supervised Learning: Images correspond to meshes.
- Using the RIGHT data is important.

- Convolutional Neural Network
  - Training takes a long time.
  - Inference runs at 60 fps.

FULL PERFORMANCE IN REAL-TIME MOCAP SUIT



This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Unsupervised Training for 3D Morphable Model Regression

Kyle Genova[1,2] Forrester Cole[2] Aaron Maschinot[2] Aaron Sarna[2] Daniel Vlasic[2] William T. Freeman[2,3]
[1]Princeton University [2]Google Research [3]MIT CSAIL

Abstract

We present a method for training a regression network from image pixels to 3D morphable model coordinates using only unlabeled photographs. The training loss is based on features from a facial recognition network, computed on-the-fly by rendering the predicted faces with a differentiable renderer. To make training from features feasible and avoid network fooling effects, we introduce three objectives: a batch distribution loss that encourages the output distribution to match the distribution of the morphable model, a loopback loss that ensures the network can correctly reinterpret its own output, and a multi-view identity loss that compares the features of the predicted 3D face and the input photograph from multiple viewing angles. We train a regression network using these objectives, a set of unlabeled photographs, and the morphable model itself, and demonstrate state-of-the-art results.

1. Introduction

A 3D morphable face model (3DMM) [3] provides a smooth, low-dimensional "face space" spanning the range of human appearance. Finding the coordinates of a person in this space from a single image of that person is a common task for applications such as 3D avatar creation, facial animation transfer, and video editing (e.g. [2, 7, 28]). The conventional approach is to search the space through inverse rendering, which generates a face that matches the photograph by optimizing shape, texture, pose, and lighting parameters [13]. This approach requires a complex, nonlinear optimization that can be difficult to solve in practice.

Recent work has demonstrated fast, robust fitting by regressing from image pixels to morphable model coordinates using a neural network [20, 21, 29, 27]. The major issue with the regression approach is the lack of ground-truth 3D face data for training. Scans of face geometry and texture are difficult to acquire, both because of expense and privacy considerations. Previous approaches have explored synthesizing training pairs of image and morphable model coordinates in a preprocess [20, 21, 29], or training an image-

Figure 1. Neutral 3D faces computed from input photographs using our regression network. We map features from a facial recognition network [24] into identity parameters for the Basel 2017 Morphable Face Model [8].

to-image autoencoder with a fixed, morphable-model-based decoder and an image-based loss [27].

This paper presents a method for training a regression network that removes both the need for supervised training data and the reliance on inverse rendering to reproduce image pixels. Instead, the network learns to minimize a loss based on the facial identity features produced by a face recognition network such as VGG-Face [16] or Google's FaceNet [24]. These features are robust to pose, expression, lighting, and even non-photorealistic inputs. We exploit this

8377

LAYERED MACHINE LEARNING

- Use machine learning to train a machine!

SUPERVISED LEARNING
VS
UNSUPERVISED LEARNING

NVIDIA®