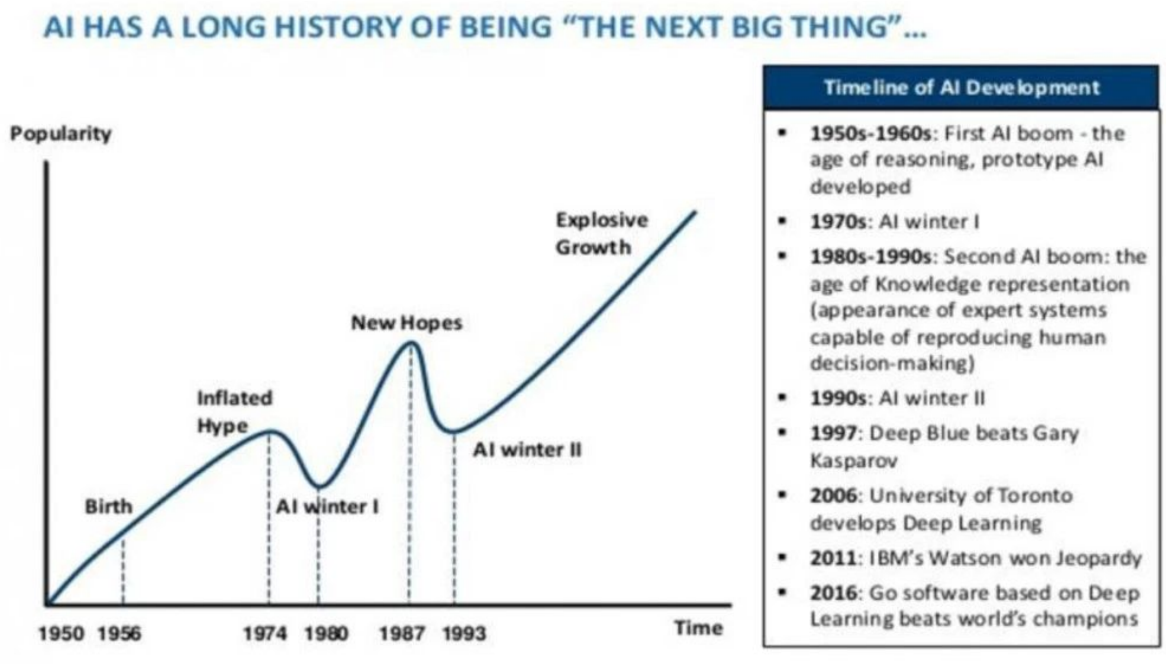
Limitations of Deep Learning

Adversarial Attacks, Data Bias, and Uncertainty

N. Rich Nguyen, PhD

SYS 6016

AI “Hype”: Historical Perspective



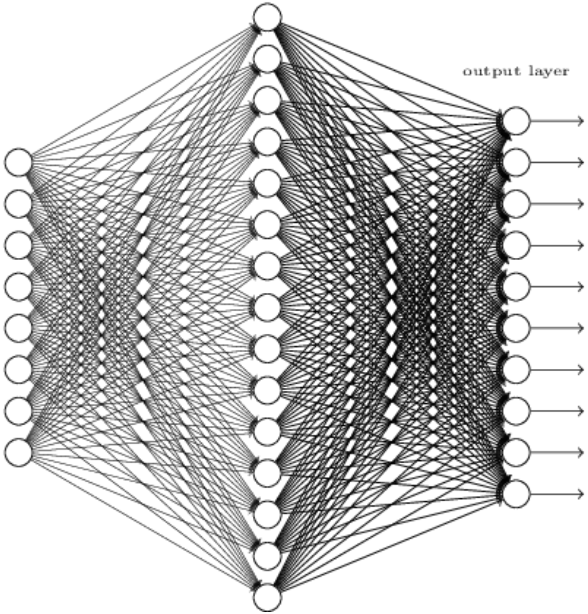
Source: https://venturebeat.com/2019/11/16/is-ai-in-a-golden-age-or-on-the-verge-of-a-new-winter/

Universal Approximation Theorem

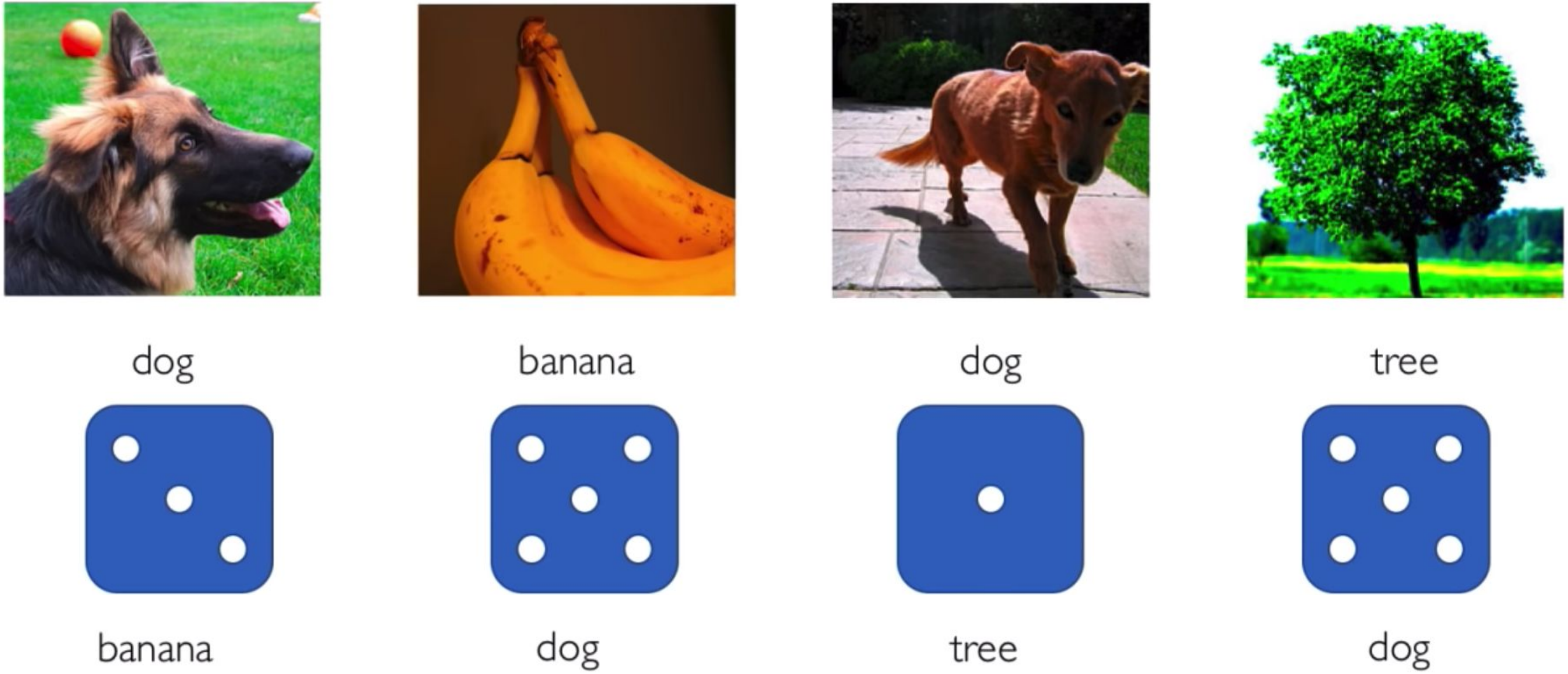
A feedforward neural network with **a single hidden layer** is sufficient to approximate, to an arbitrary precision, any continuous function - George Cybenko, 1989.

A few caveats:

- Number of hidden units may be unfeasibly **large**
- How to **train** such a large network
- The network may not **generalize**



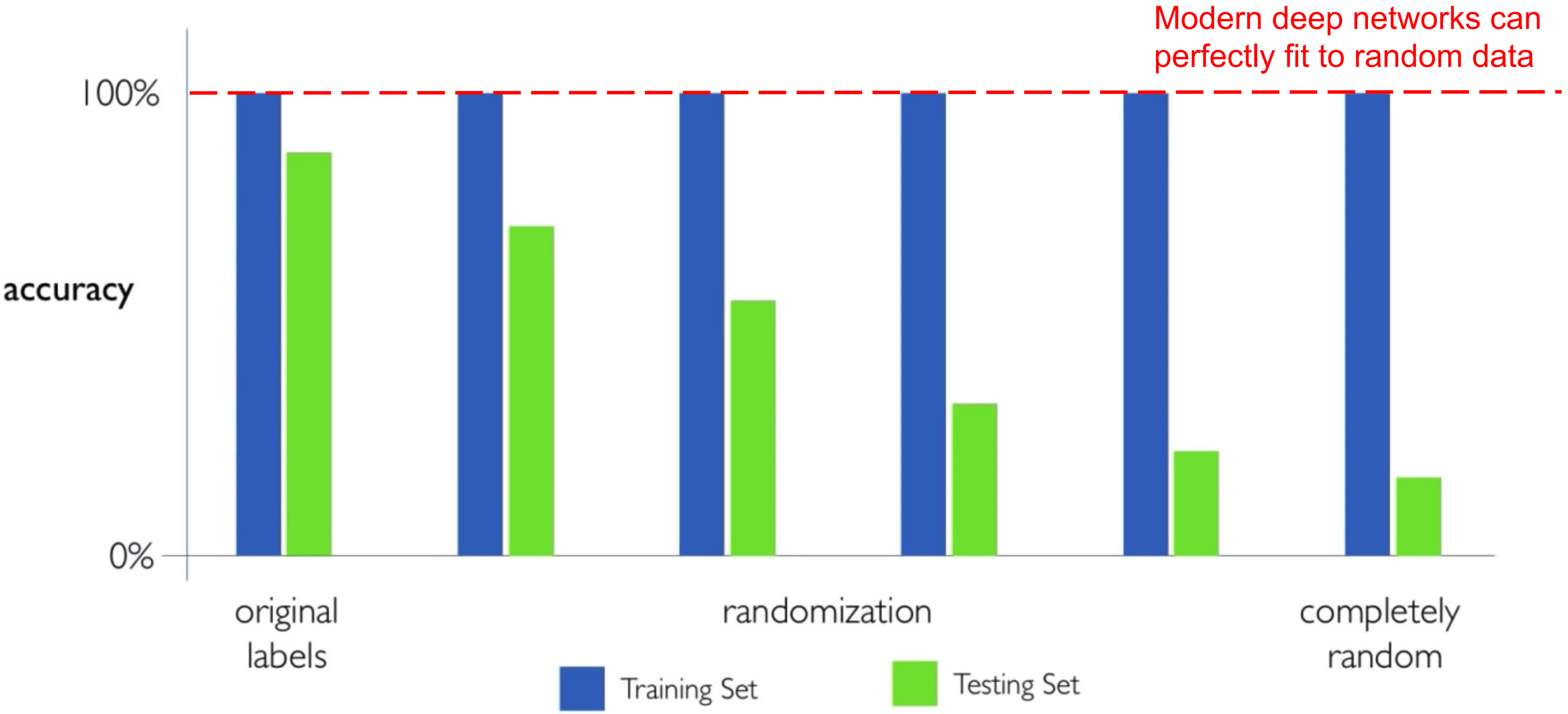
Rethink Generalization



Source: Zhang+ “Understanding Deep Neural Networks Requires Rethinking Generalization”, ICLR 2017

Capacity of Deep Neural Networks

Capacity of Deep Neural Networks



Recognition != Understanding

- A model generates captions to accurately describe pictures (mis)leads to a belief that the model “understand” the content of the pictures and the captions it generates.
- It doesn’t understand **either** the image nor the caption in the sense of humans can!
- Any slight departure from the sort of images present in the training data causes the model to generate completely absurd caption.

The boy is holding a baseball bat.

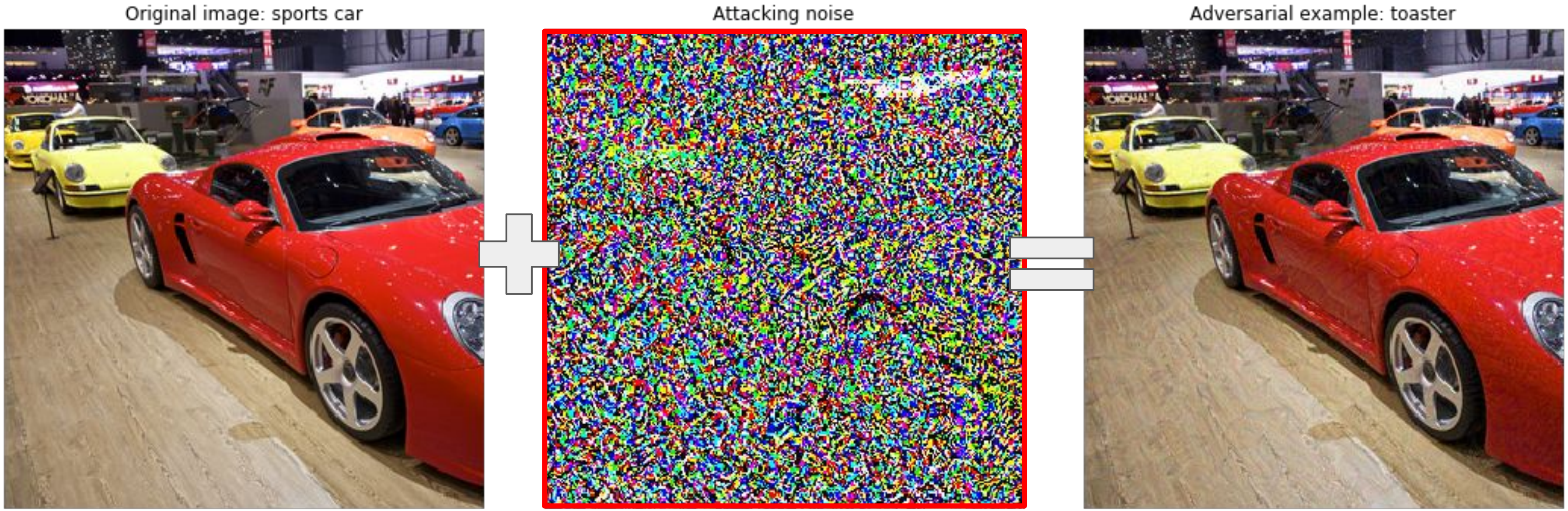
Current Limitations

- Very data hungry
- Computational intensive to train and deploy
- Uninterpretable **black boxes**, difficult to trust
- **Finicky to optimize**: non-convex, choice of architecture, hyperparameters

- Easily fooled by **adversarial** examples
- Can be subject to algorithmic/data **bias**
- Is poorly structured to represent **uncertainty, and** difficult to **encode structure** and prior knowledge

Adversarial Attacks

Adversarial Attacks on Neural Networks



Adversarial Attacks

We train our networks with gradient descent:

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha \frac{\partial \mathbf{J}(\mathbf{W}, \mathbf{x}, y)}{\partial \mathbf{W}}$$

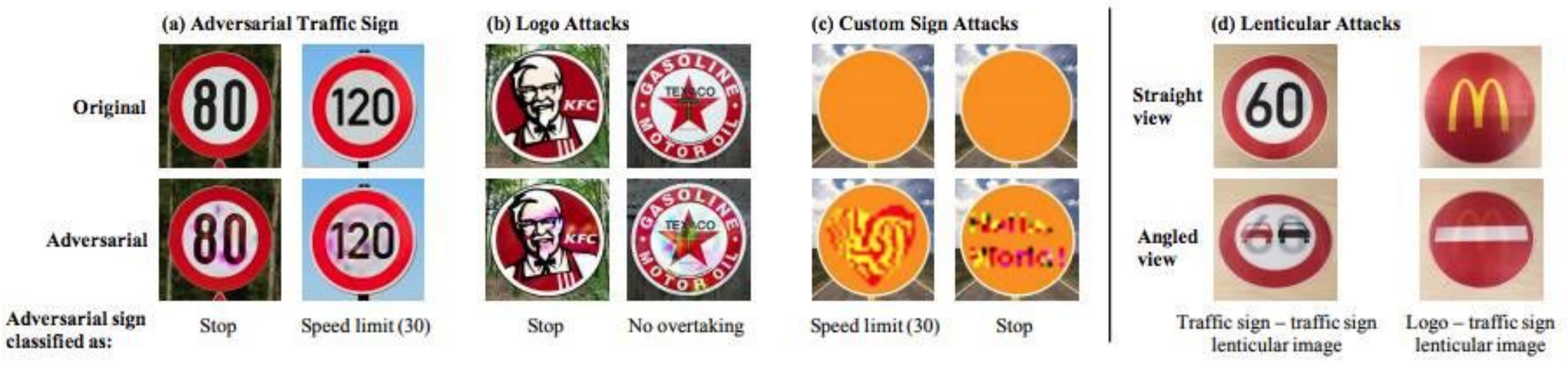
"How does a small change in weights ***decrease*** *our loss?"*

In Adversarial attack, an attacking image was modified to increase error:

$$\mathbf{x} \leftarrow \mathbf{x} + \alpha \frac{\partial \mathbf{J}(\mathbf{W}, \mathbf{x}, y)}{\partial \mathbf{x}}$$

"How does a small change in the input ***increase*** *our loss"*

Adversarial Examples



Data / Algorithmic Bias

ImageNet

The ImageNet contain large number of images (14.2 million) of over 1000 classes, some of which are really subtle (try distinguishing 120 dog breeds)

Over the years, variants of the CNN architectures have been developed, leading to amazing advances in the field and top-5 error has been as low as 1.3%

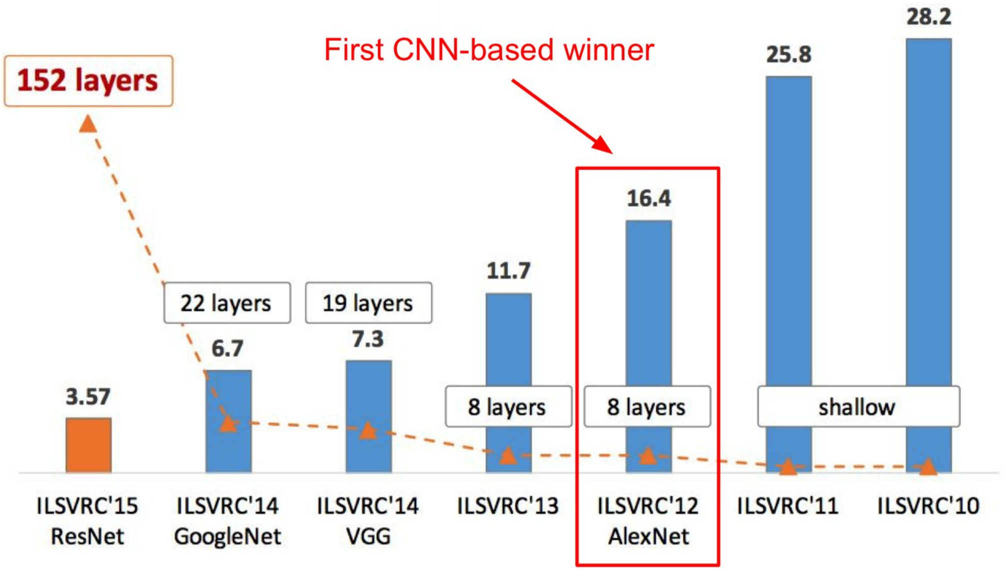


Figure copyright Kaiming He, 2016.

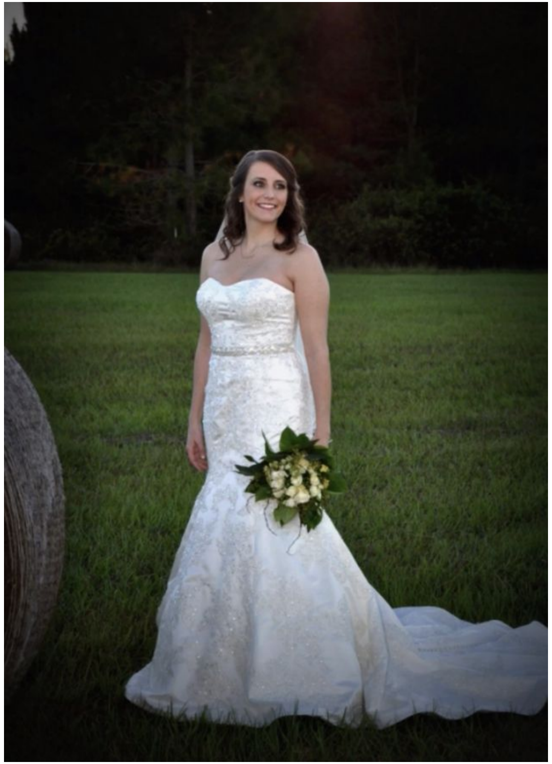
OpenImages Dataset

~9,000,000 images

~6,000 labels (multi-label)

https://github.com/openimages/dataset

Classification Bias

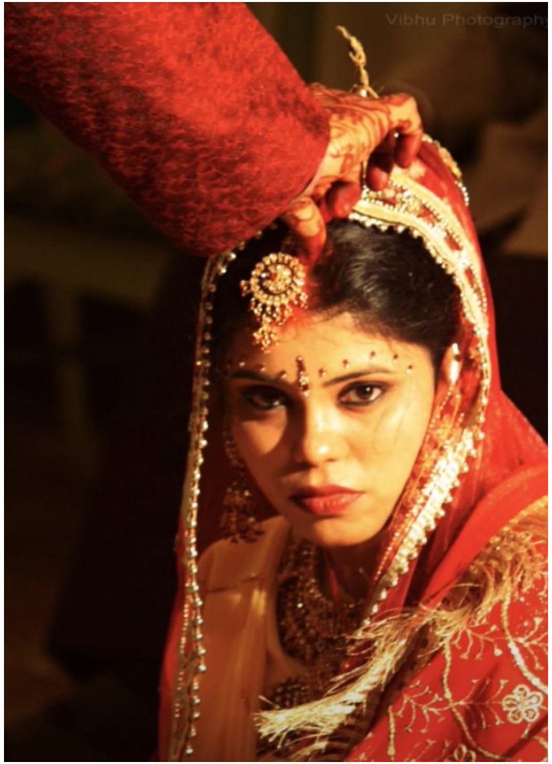


A classification model trained on **OpenImages**

Top 5 classes

- Bride
- Dress
- Ceremony
- Women
- Wedding

Classification Bias

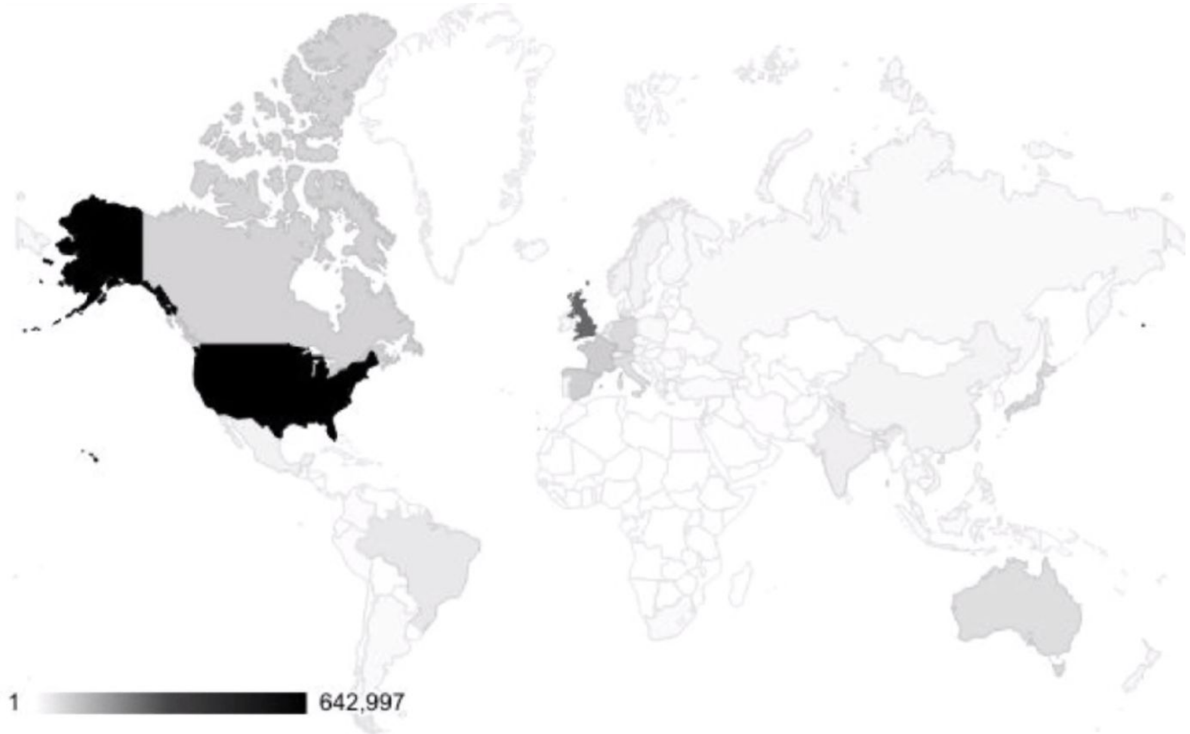


A classification model trained on **OpenImages**

Top 5 classes

- Clothing
- Event
- Costume
- Red
- Performance art

Geographical bias in training



Maps of location of contributors to training data based on ~22% of images with identifiable geolocation

Stereotype

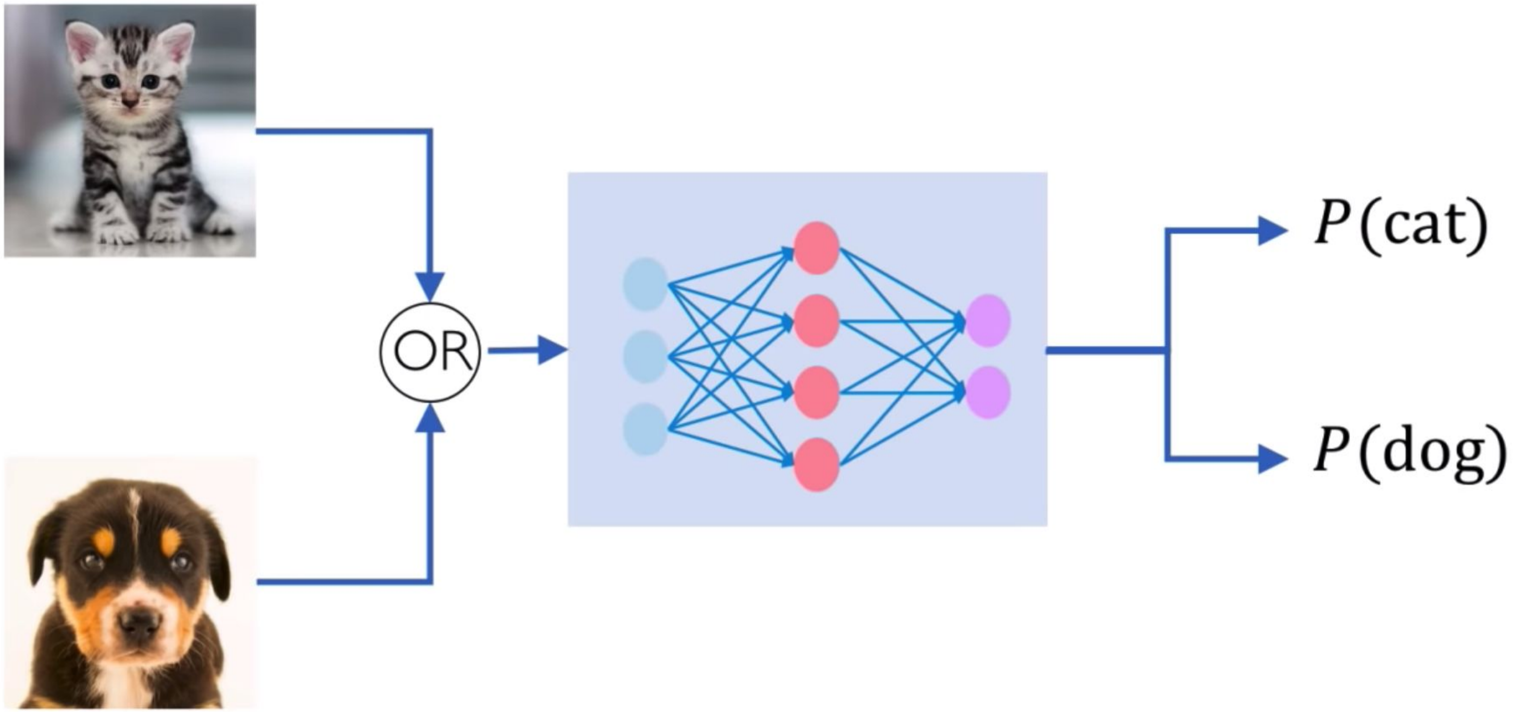
A stereotype is a statistical confounder with **societal basis**

Be aware of differences between **training** and **inference** (testing) distributions

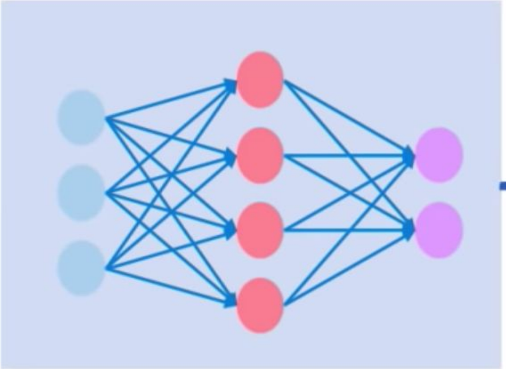
Don’t take any dataset without questioning it!

Uncertainty Estimation

Why care about uncertainty?



Why care about uncertainty?



$P(\text{cat}) = 0.2$

$P(\text{dog}) = 0.8$

Remember: $P(\text{cat}) + P(\text{dog}) = 1$

We need **uncertainty** metrics to assess the network’s confidence in its predictions.

Bayesian Deep Learning for Uncertainty

Network tries to learn output, **`Y`**, directly from raw data, **`X`**

Find mapping function, f, parameterized by weights **`W`** such that

$$\min \mathcal{L}(\mathbf{Y}, f(\mathbf{X}, \mathbf{W}))$$

Bayesian neural networks aim to learn a posterior distribution over network weights:

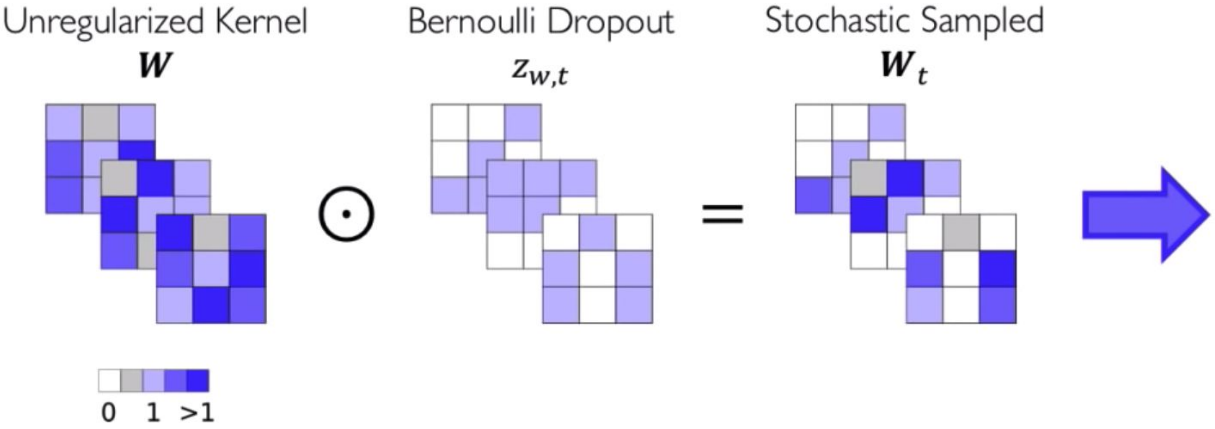
Intractable! $\longrightarrow$ $$P(\mathbf{W}|\mathbf{X}, \mathbf{Y}) = \frac{P(\mathbf{Y}|\mathbf{X},\mathbf{W})P(\mathbf{W})}{P(\mathbf{Y}|\mathbf{X})}$$

Approximate the posterior using **sampling**

Stochastic Sampling for Uncertainty

Evaluate `T` stochastic forward passes through the network $\{\mathbf{W}_t\}_{t=1}^{T}$

Dropout as a form of stochastic sampling $z_{w,t} \sim \text{Bernoulli}(p)\ \forall w \in \mathbf{W}$



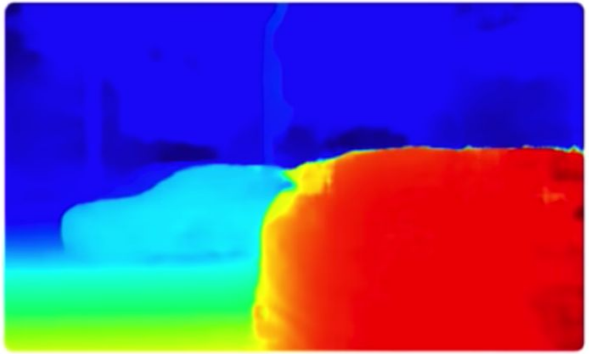
$$\mathbb{E}(\hat{Y}|X) = \frac{1}{T}\sum_{t=1}^{T} f(X|W_t)$$

$$Var(\hat{Y}|X) = \frac{1}{T}\sum_{t=1}^{T} f(X)^2 - \mathbb{E}(\hat{Y}|X)^2$$

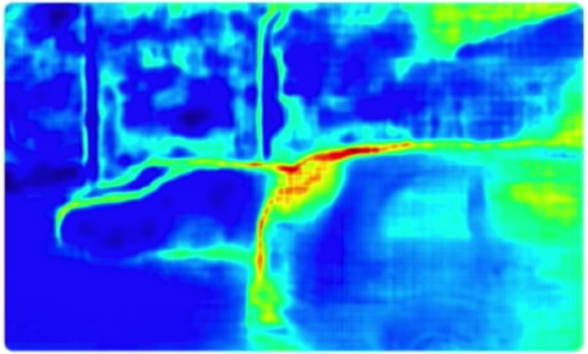
Source: Gal and Ghahramani+, ICML 2016

Model Uncertainty Application

Input Image



Predicted Depth

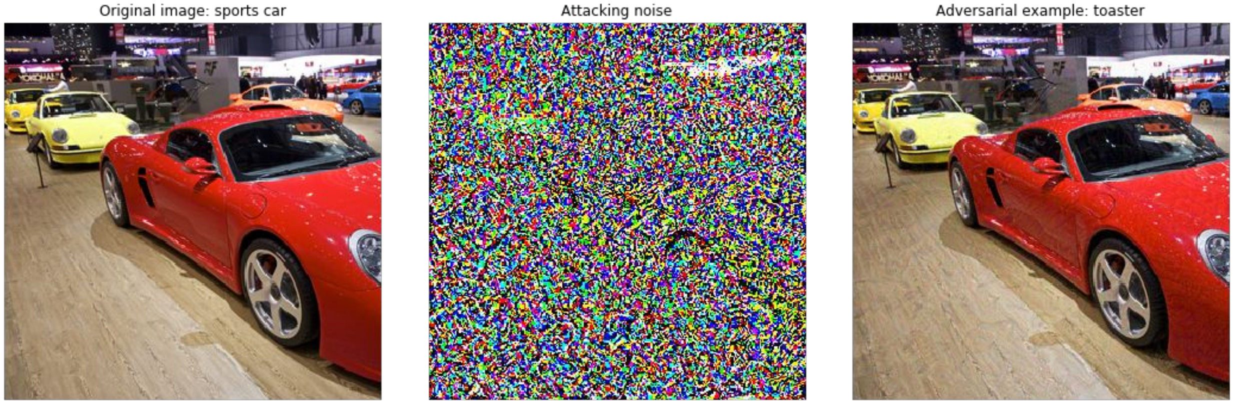


Model Uncertainty

Source: Kendall+, NIPS 2017

Summary: Limitations of Deep Learning

- Easily fooled by **adversarial** examples
- Can be subject to algorithmic/data **bias**
- Is poorly structured at representing **uncertainty, and** difficult to **encode structure** and prior knowledge



Acknowledgements

Slides contain figures from Andrej Karpathy (Stanford), Ava Soleimany (MIT), and various researchers reproduced only for educational purposes.

Bonus Content

In the press

≡ WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY TRANSPORTATION

JASON PONTIN IDEAS 02.02.2018 08:00 AM

Greedy, Brittle, Opaque, and Shallow: The Downsides to Deep Learning

We've been promised a revolution in how and why nearly everything happens. But the limits of modern artificial intelligence are closer than we think.

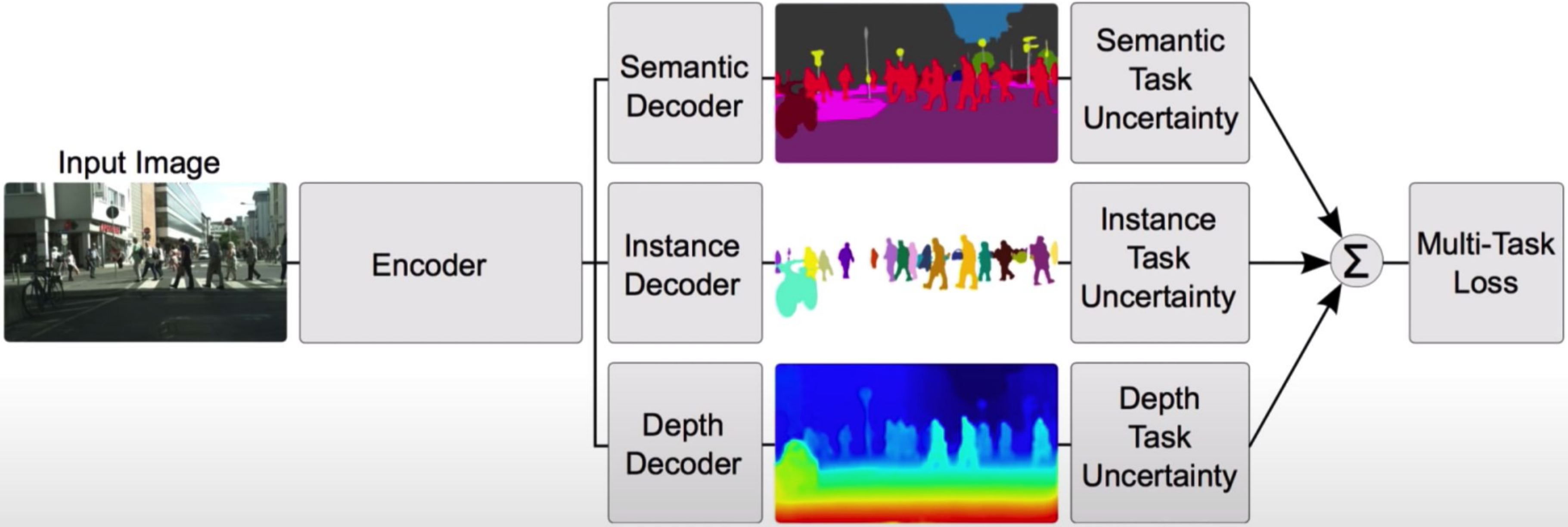
According to skeptics like Marcus, a a professor of cognitive psychology at NYU and briefly director of Uber's AI lab, deep learning is greedy, brittle, opaque, and shallow:

- **Greedy** because they demand huge sets of training data.
- **Brittle** because when a neural net is given a "transfer test"—confronted with scenarios that differ from the examples used in training—it cannot contextualize the situation and frequently breaks.
- **Opaque** because their parameters can only be interpreted in terms of their weights within a mathematical geography. Consequently, they are black boxes, whose outputs cannot be explained, raising doubts about their reliability and biases.
- **Shallow** because they are programmed with little innate knowledge and possess no common sense about the world or human psychology.

https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning/

Multi-Task Learning Using Uncertainty

Model ensembling for estimating uncertainty

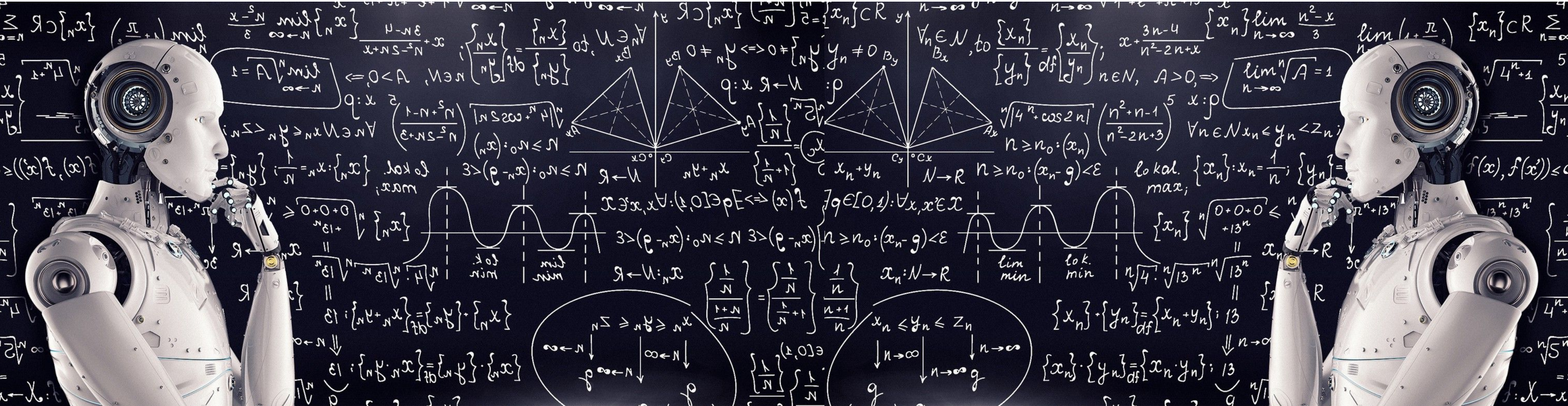


Source: Kendall+, CVPR 2018

Connection to Artificial General Intelligence?

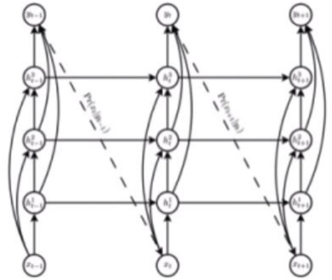
- Design an AI algorithm that can build new models capable of solving a task
- Reduce the need for experienced engineers to design the networks
- Make deep learning more accessible to the public

→ The connection to **Artificial General Intelligence (AGI)**: the ability to intelligently reason about how we learn

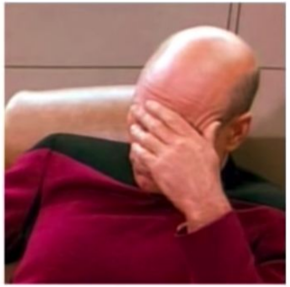


Learning to Learn

Motivation for Automated ML



Complexity of models increases



Greater need for specialized engineers

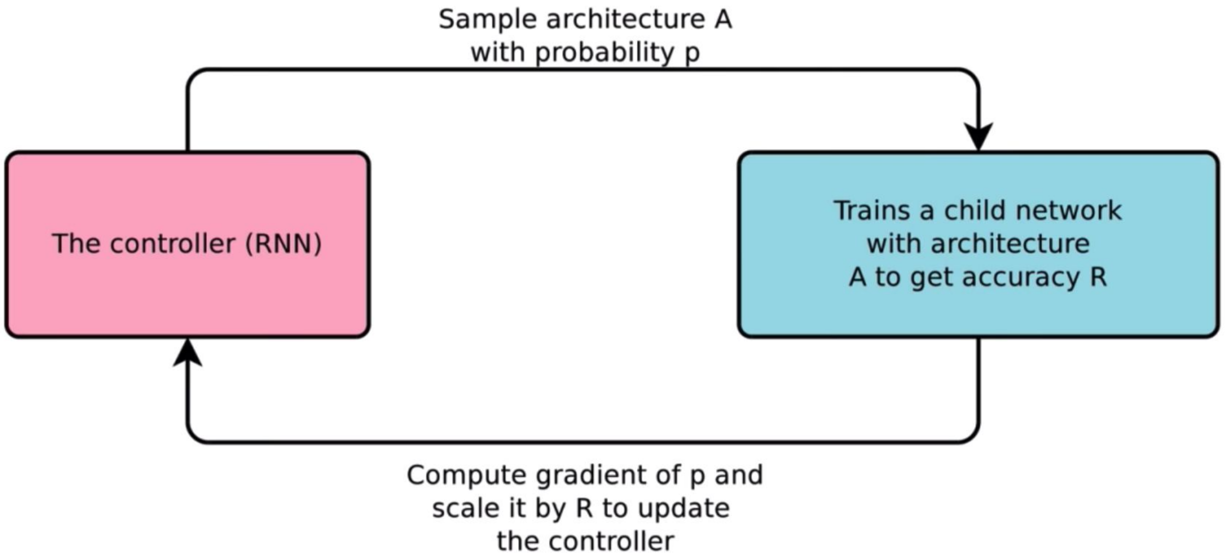
Standard deep neural networks are optimized for a **single task**

Often require expert knowledge to build an architecture for a given task

Build a learning algorithm that **learns which model to use** to solve a given problem → **AutoML**

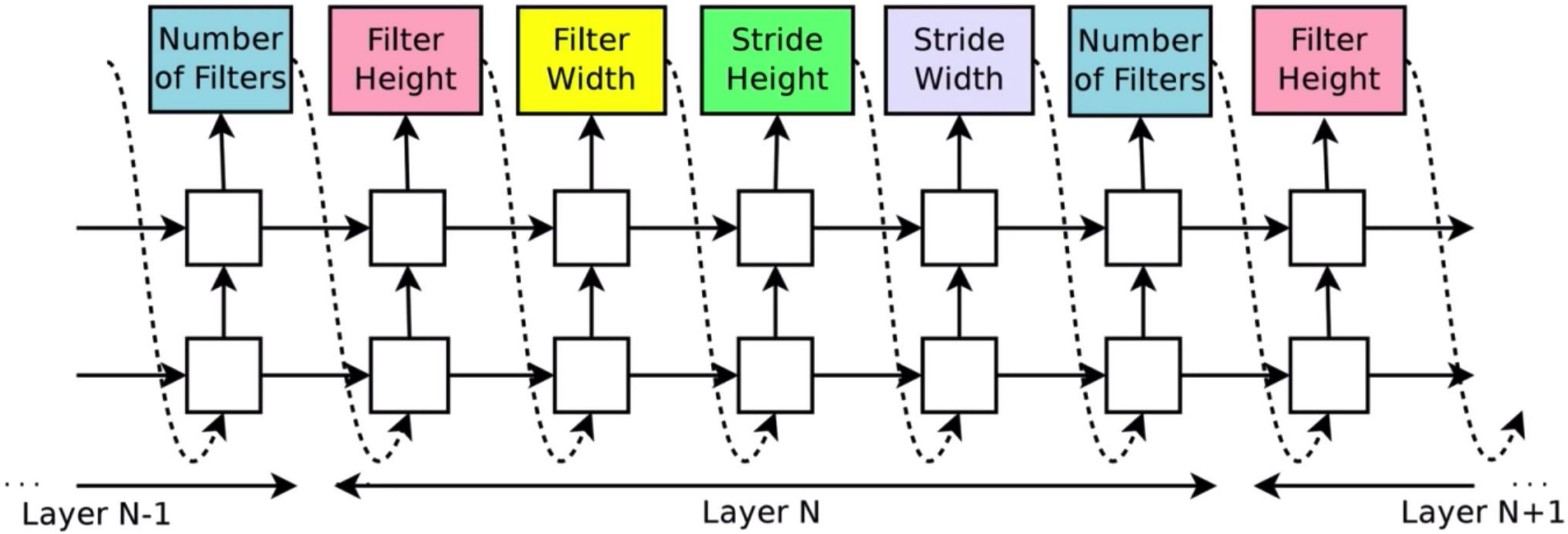
Automated Machine Learning (AutoML)

A reinforcement Learning framework to learn a appropriate network for a given task.

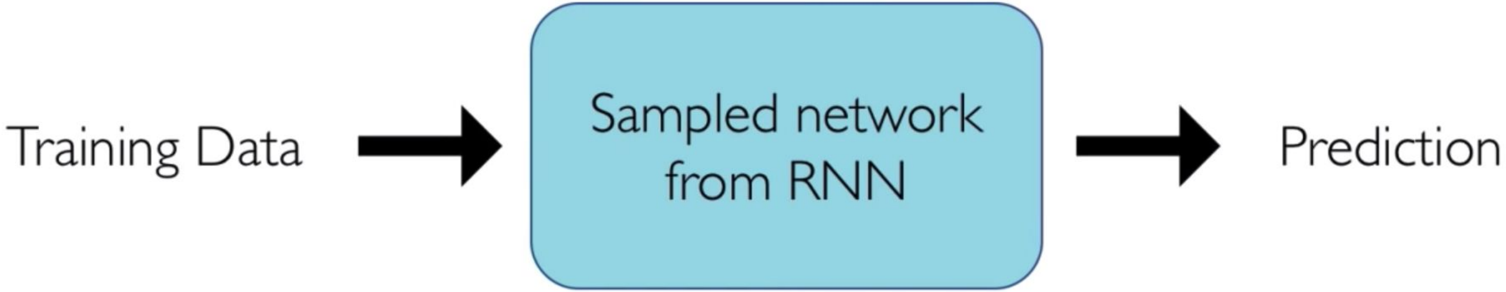


AutoML: Model Controller

At each step, the model samples a brand new network



AutoML: The Child Network

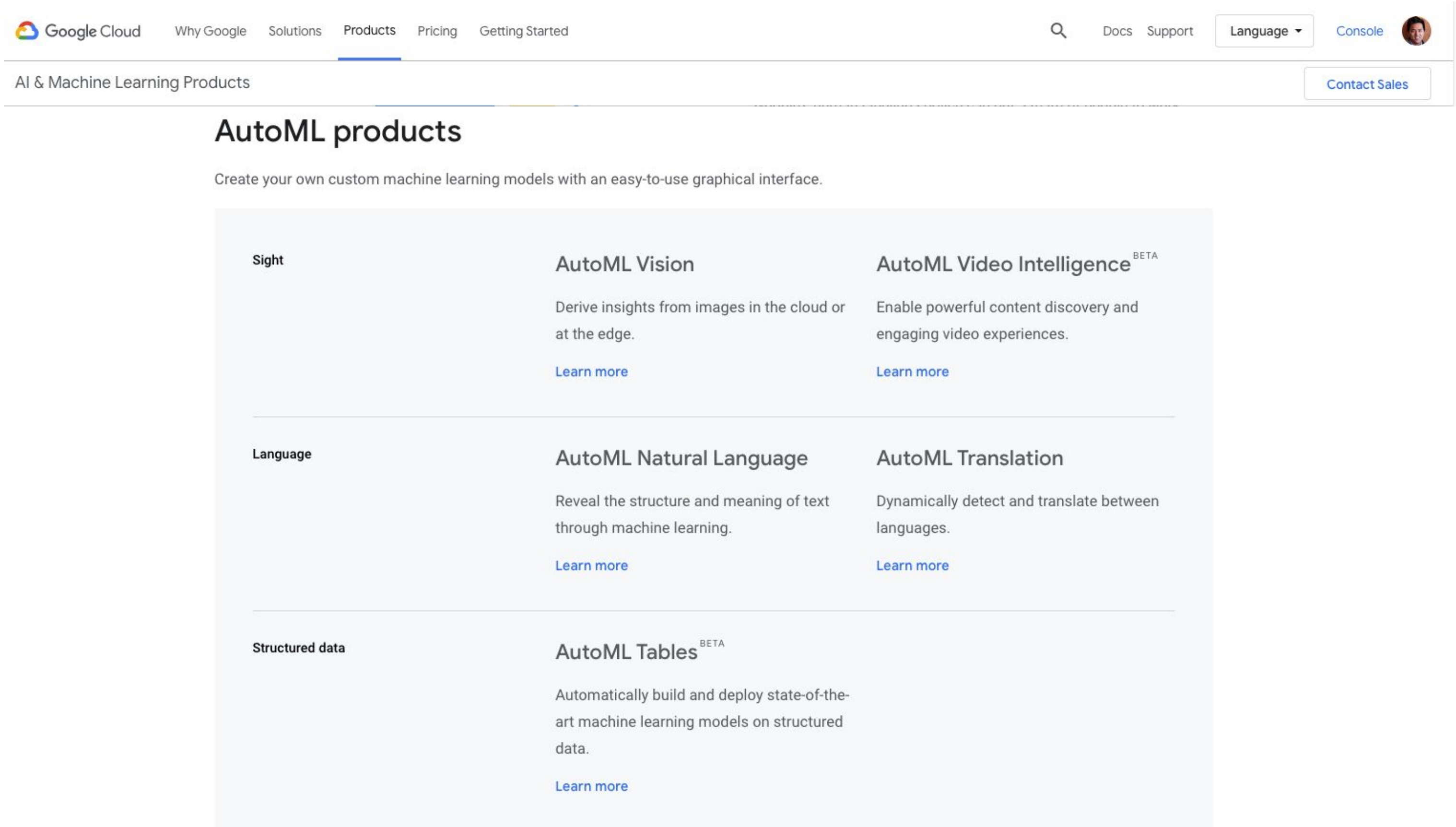


Compute final accuracy on this dataset.

Update RNN controller based on the accuracy of the child network after training

Source: Zoph and Le, ICLR 2017

AutoML on the Cloud


Google Cloud | Why Google | Solutions | Products | Pricing | Getting Started | Docs | Support | Language | Console

AI & Machine Learning Products | Contact Sales

AutoML products

Create your own custom machine learning models with an easy-to-use graphical interface.

Sight

AutoML Vision

Derive insights from images in the cloud or at the edge.

Learn more

AutoML Video Intelligence BETA

Enable powerful content discovery and engaging video experiences.

Learn more

Language

AutoML Natural Language

Reveal the structure and meaning of text through machine learning.

Learn more

AutoML Translation

Dynamically detect and translate between languages.

Learn more

Structured data

AutoML Tables BETA

Automatically build and deploy state-of-the-art machine learning models on structured data.

Learn more

https://cloud.google.com/automl