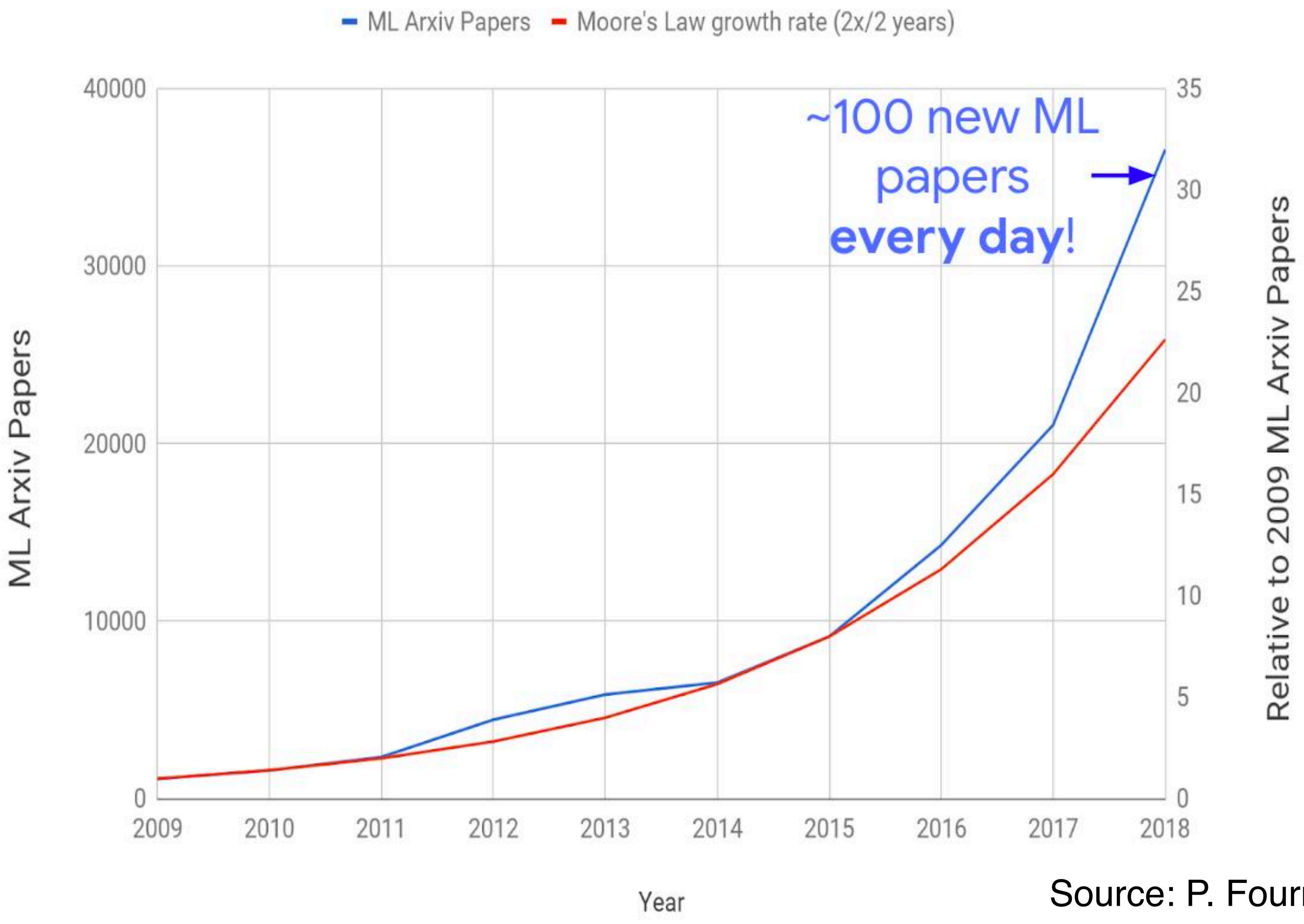
An introduction to machine learning

Zachary E. Ross
California Institute of Technology

What is machine learning?

- Branch of artificial intelligence focused on learning structure **from data**
- Has been around for a while; overlaps with statistics, data mining, others
- 2012: the year of the deep learning revolution
- Today, deep learning is state of the art in many domains of AI

Machine Learning Arxiv Papers per Year



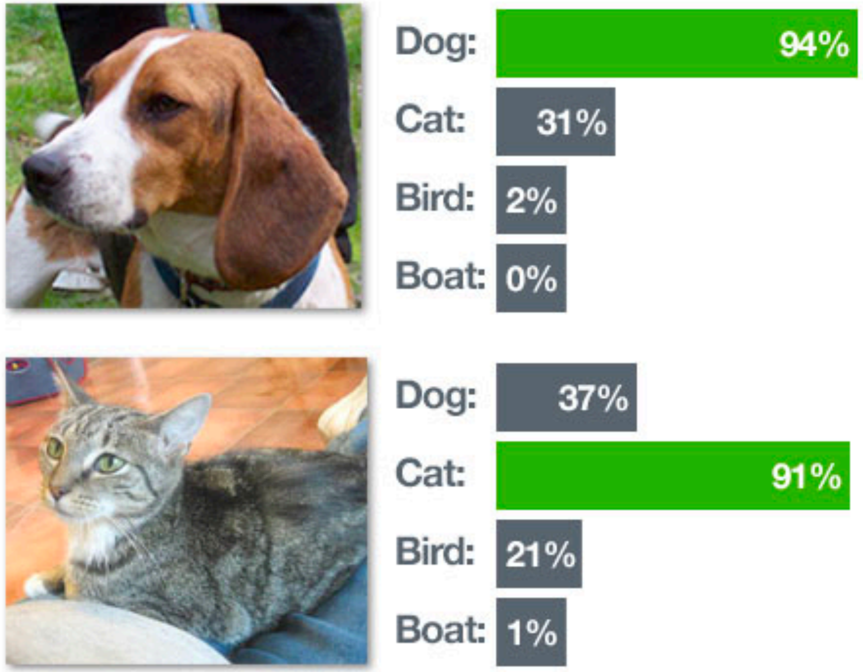
Source: P. Fournier-Viger

Supervised learning

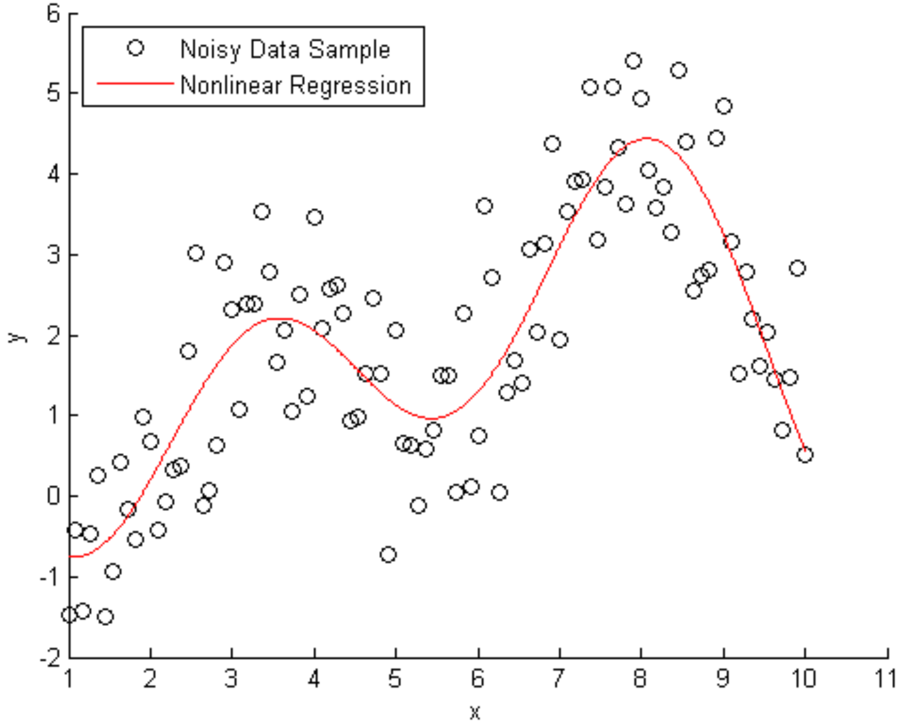
- Supervised learning focuses on learning mapping function, f

- $y = f(x)$
  - x: features (D dimensional)
  - y: targets; labels (C dimensional)
  - y discrete (classification), or continuous (regression)

Classification

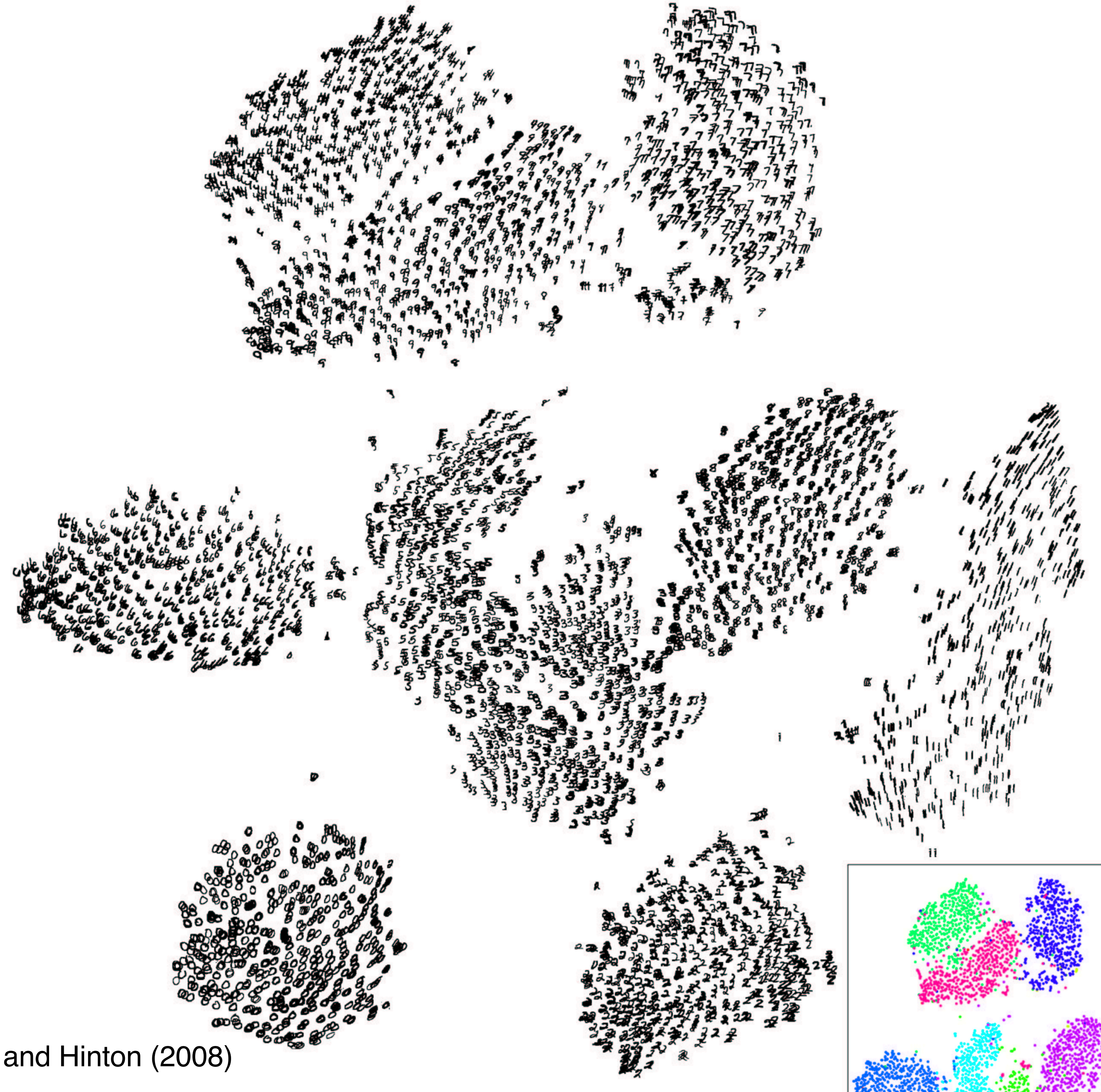


Regression



Unsupervised learning

- Unsupervised learning: learn a useful representation of a dataset



Van Der Maaten and Hinton (2008)

Supervised learning

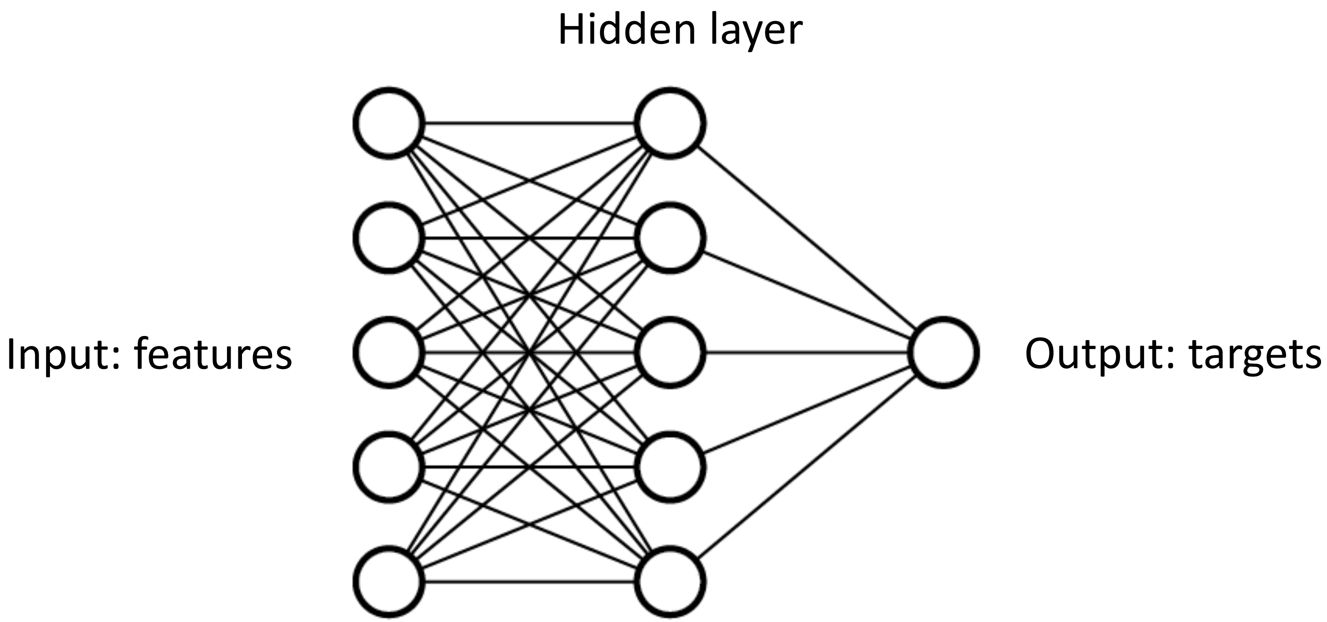
- f: a (potentially) non-linear parametric model
- f is learned from data (x, y) to minimize prediction error on **independent** test data
  - The hallmark of supervised learning!
- Requires ground-truth labeled samples
- Which f is learned? Depends on:
  - Model architecture (parameterization of f)
  - Loss function (a definition of error)
  - Optimization algorithm for training (many ML models are highly non-convex)
  - Many other factors

Many types of supervised models:

- Linear models
- Neural networks (really broad category)
- Support vector machines
- Ensemble models (e.g. random forest)
- k-nearest neighbors
- …

Neural networks form the basis for modern ML

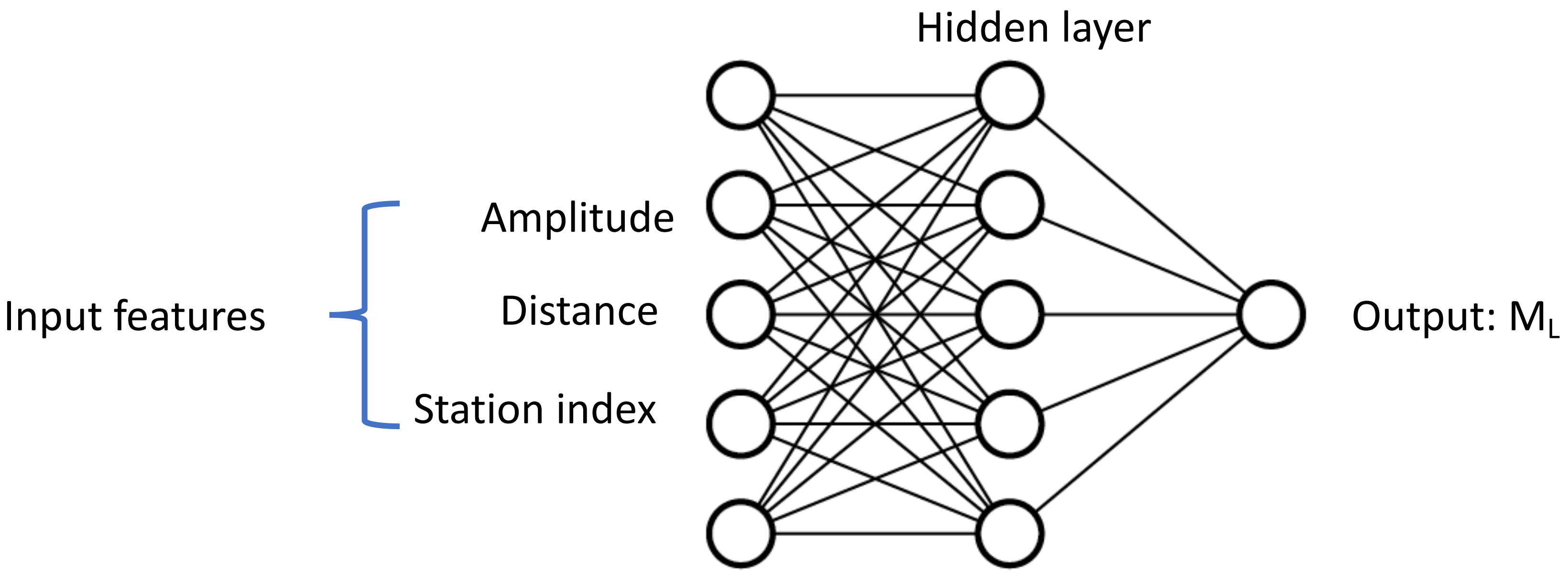
- Neural networks are layered systems designed to perform non-linear mappings
- Mapping function to be learned from lots of (*x*, *y*) data
- Can approximate **any*** non-linear function
- f : Not required to be known
- Excellent at classification and regression but generally require well-engineered features



$$y = \varphi(w_1\varphi(w_2x + b_2) + b_1)$$

$\varphi(x)$: A non-linear activation function

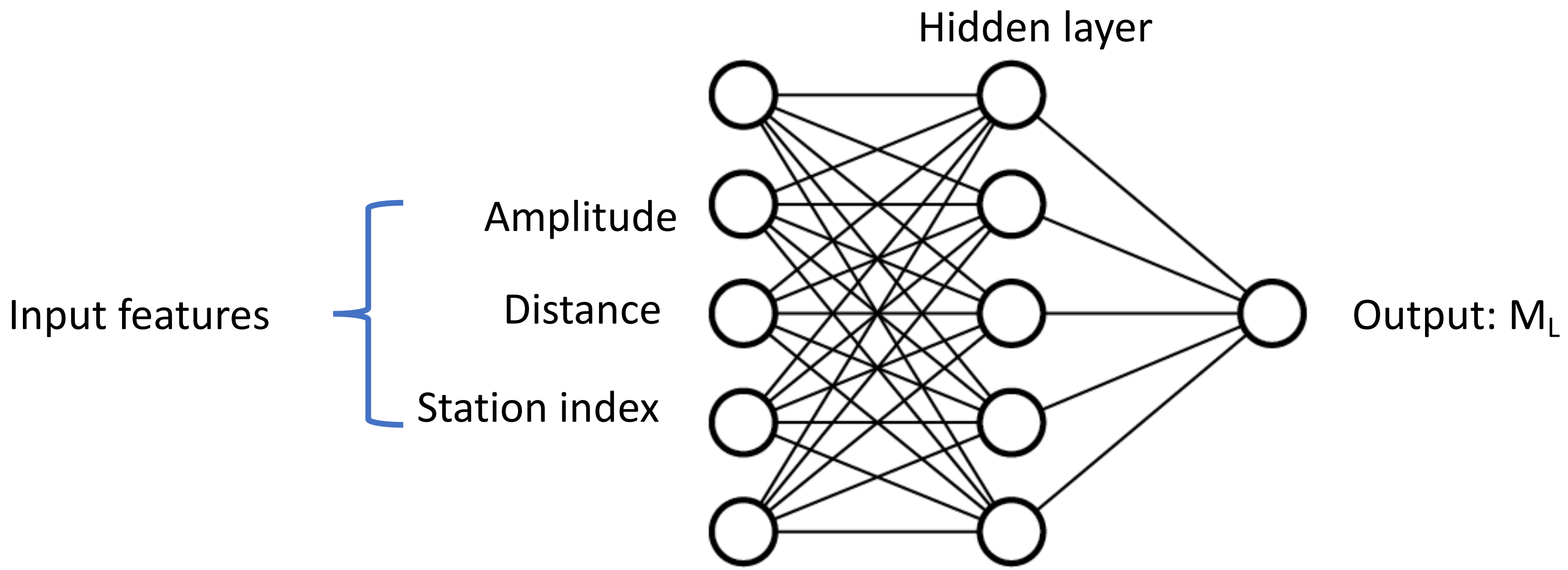
Example: local magnitude



We could train this model to learn ML without knowing the functional form.

The network would learn from a large catalog of earthquakes

Example: local magnitude

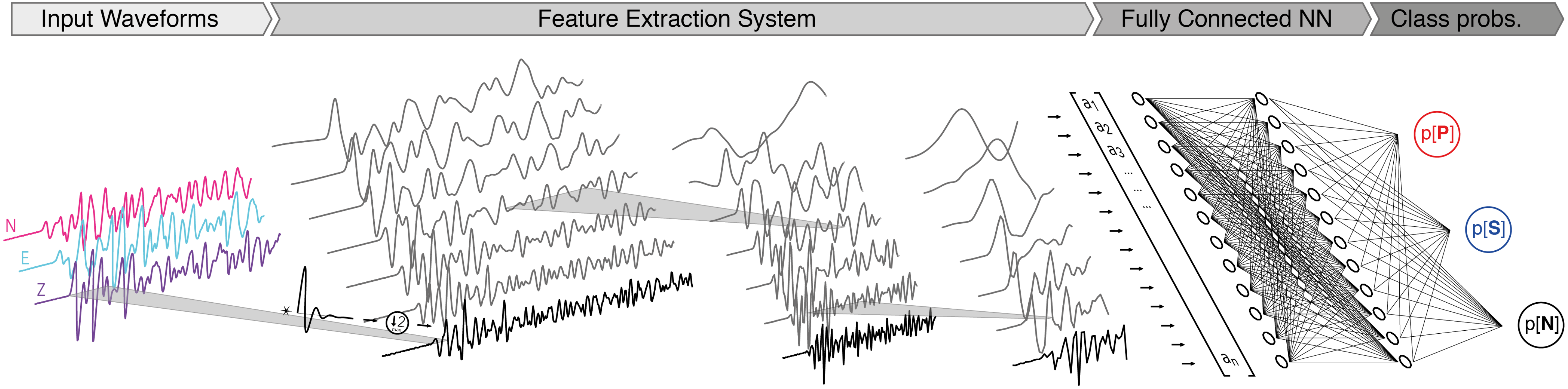


The features have been “engineered” from the waveforms and location

Could we have used a raw waveform as input instead?

Convolutional neural networks

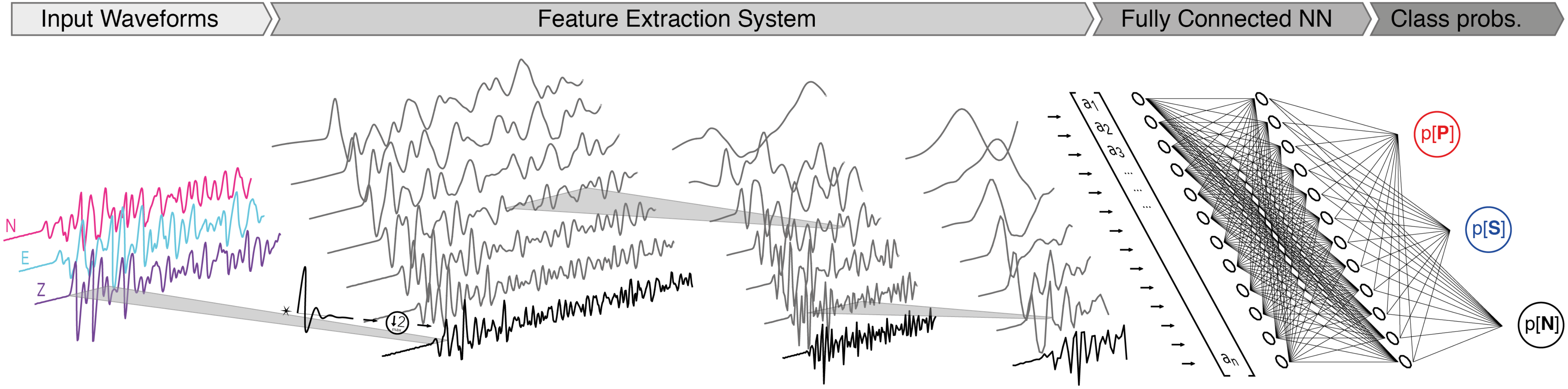
- Input images directly (or seismograms), without need for feature extraction
- Learnable feature extraction systems
- Hierarchical network of learned filters to be convolved with input (convolution layer)
- Periodic decimation of information between layers to span all length scales (pooling layer)



Ross et al. (2018), BSSA

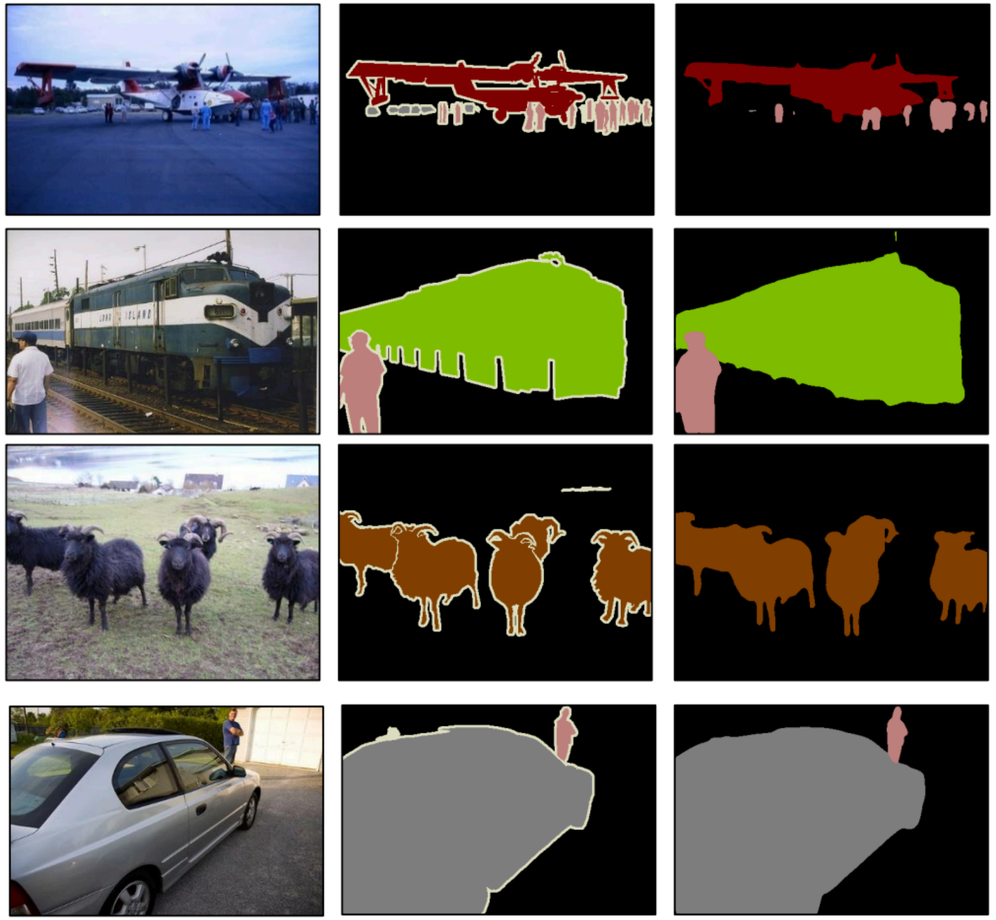
Convolutional neural networks

- Excellent at invariant pattern recognition (e.g. translation, reflection, distortion)
- Capable of generalizing the knowledge contained in extremely large datasets
  - Not necessary for input to exactly match something previously observed
- Major limitation is large amounts of labeled data samples



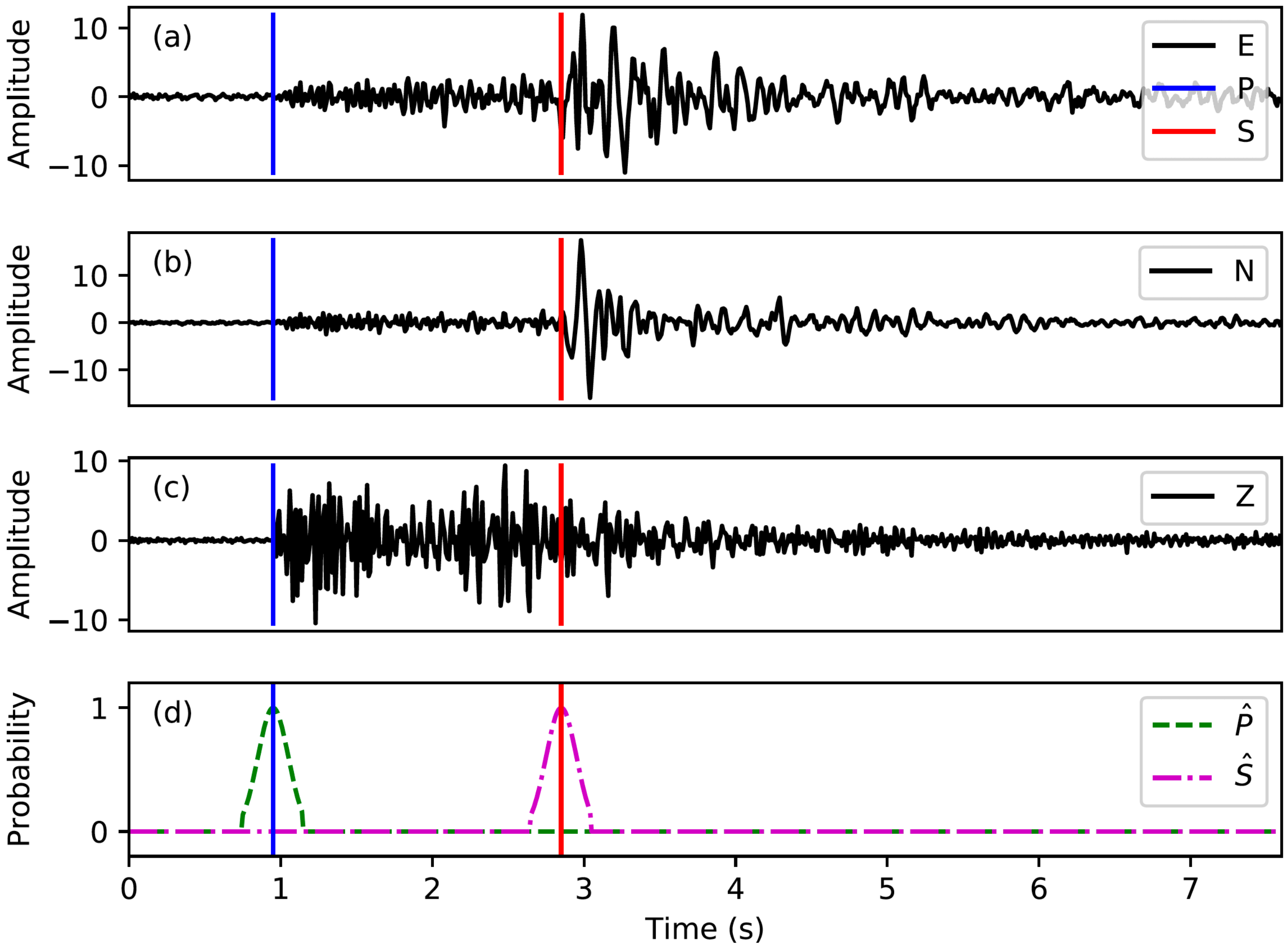
Ross et al. (2018), BSSA

One step further: semantic segmentation



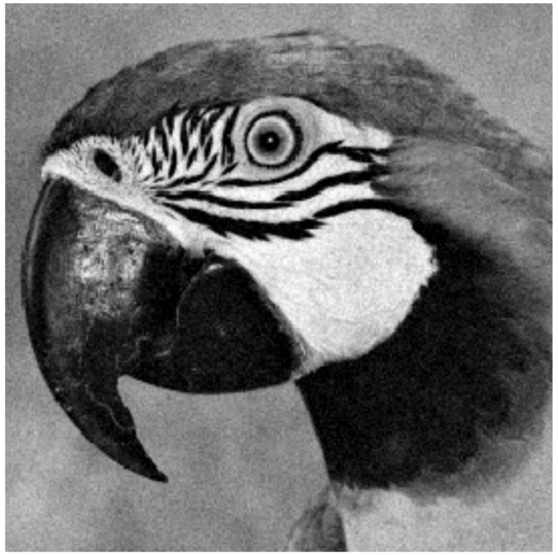
But: need tons of images with pixel-wise labels

Fu et al. 2019



Zhu and Beroza (2018)

Image denoising w/ deep convolutional networks



Noisy training image,
$\sigma = 10$ (max level)

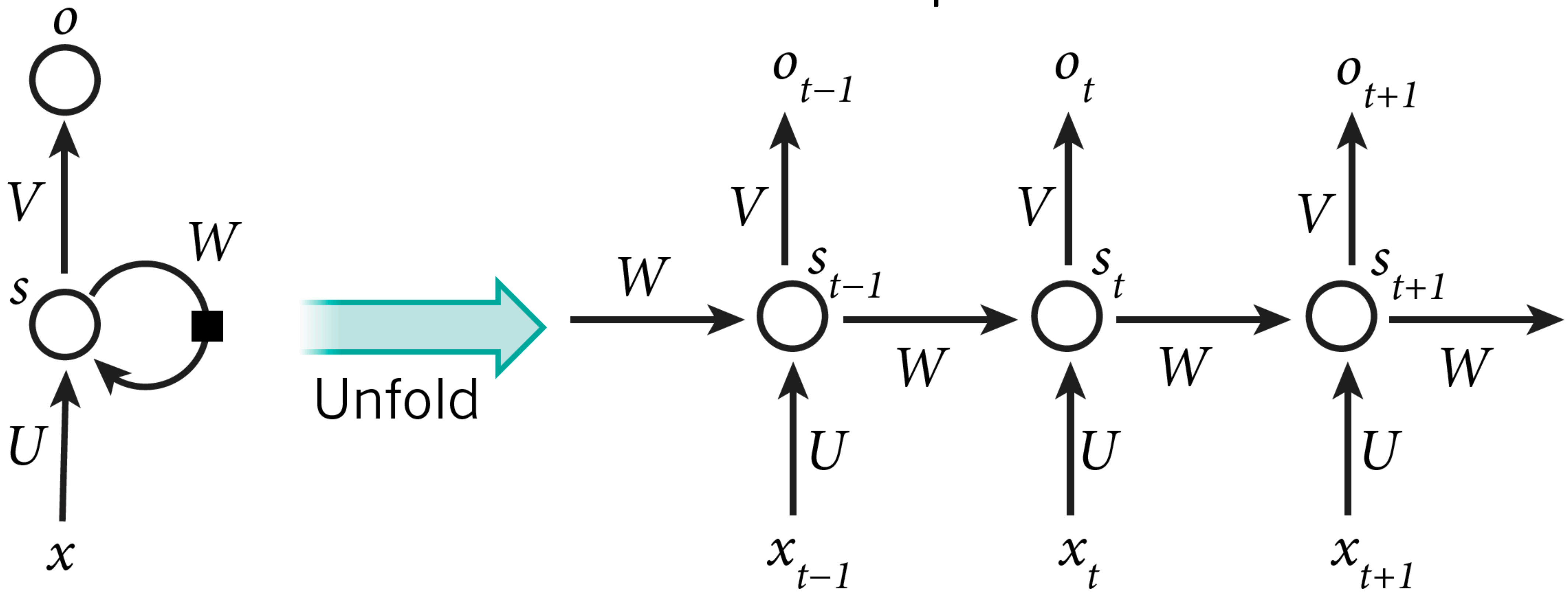
Noisy test image,
$\sigma = 90$

Test image, denoised
by CNN

Test image, denoised
by BF-CNN

Mohan et al. 2020, ICLR

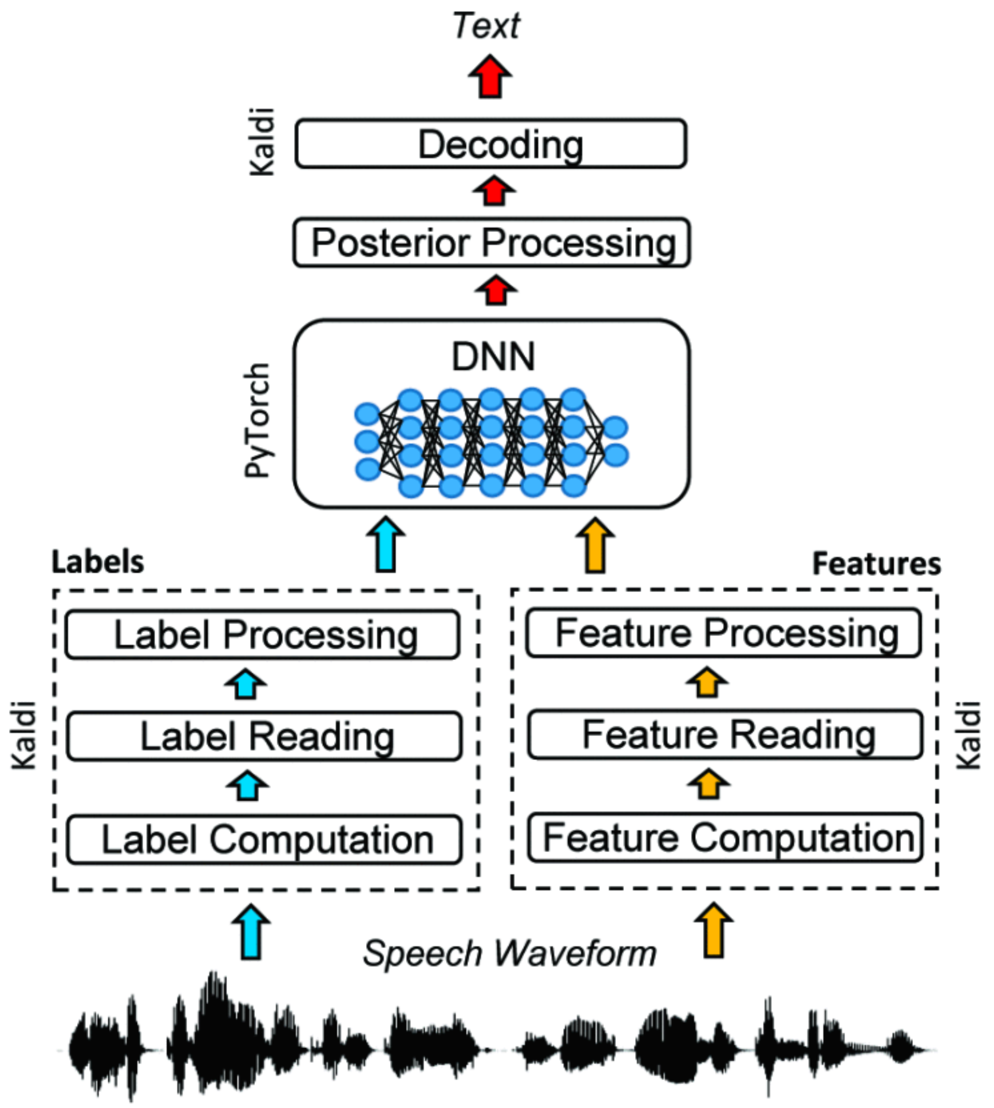
Recurrent networks act on sequential datasets



LeCun et al. 2015, Nature

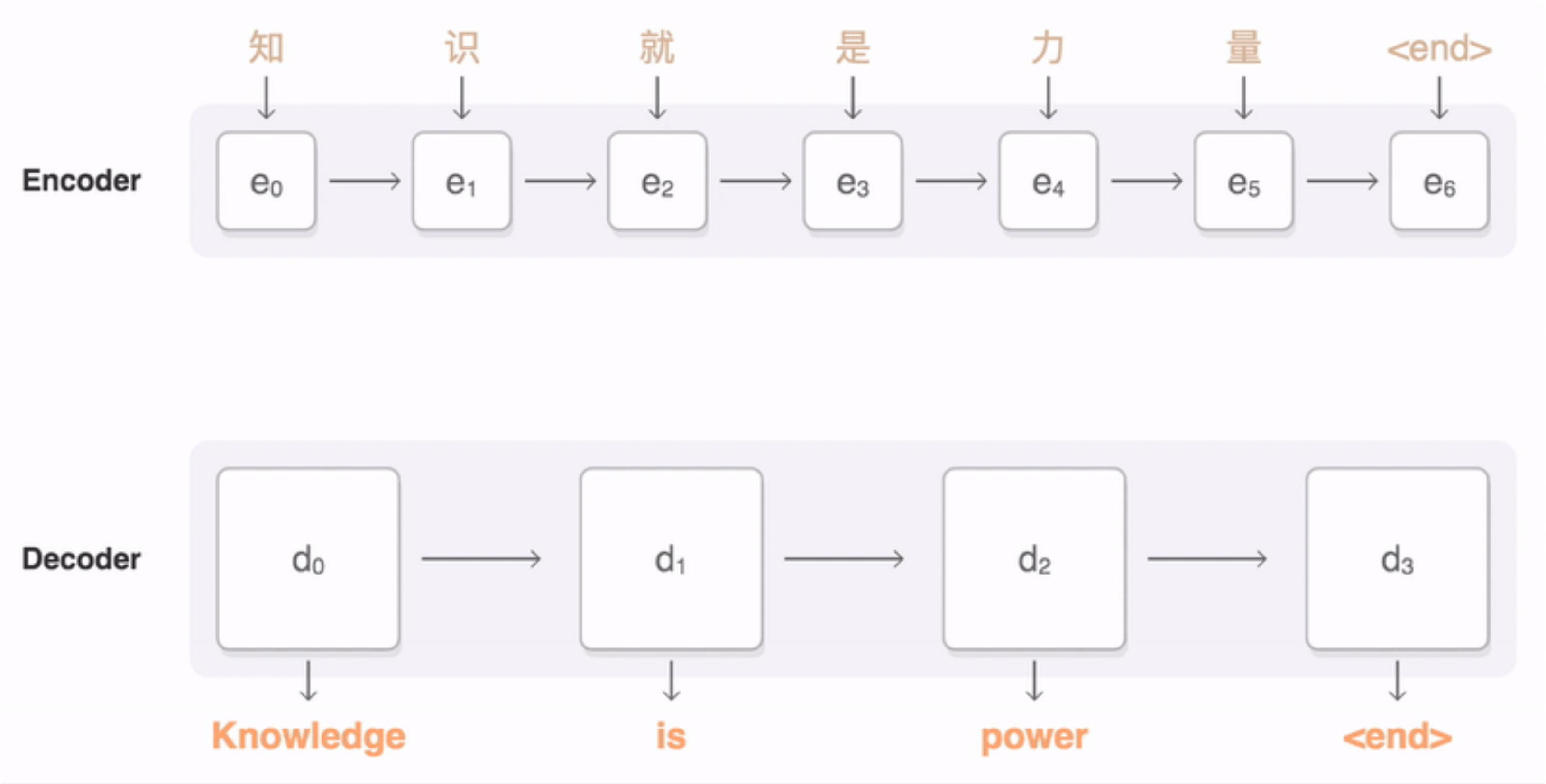
- Many types of physical datasets are multivariate time series
- Doesn't need to have time; could just be positioning/context (e.g. words in a sentence)
- RNN variants (Long Short-term memory, LSTM; Gated recurrent unit, GRU) have shared parameters across all sequence elements

RNNs achieve outstanding performance on speech recognition



Ravanelli et al. 2019

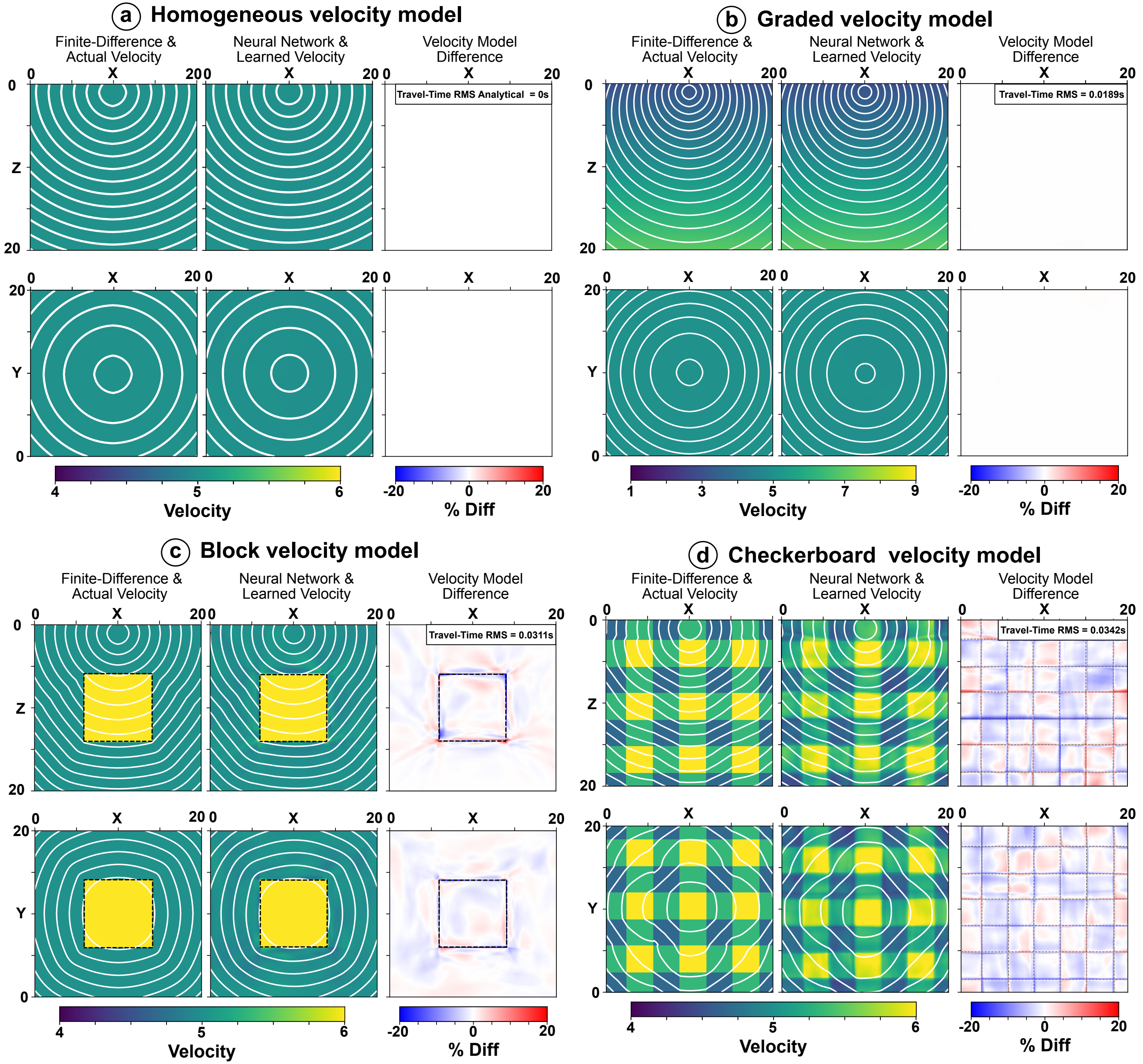
Encoder/decoder models are excellent at language translation



Source: Google

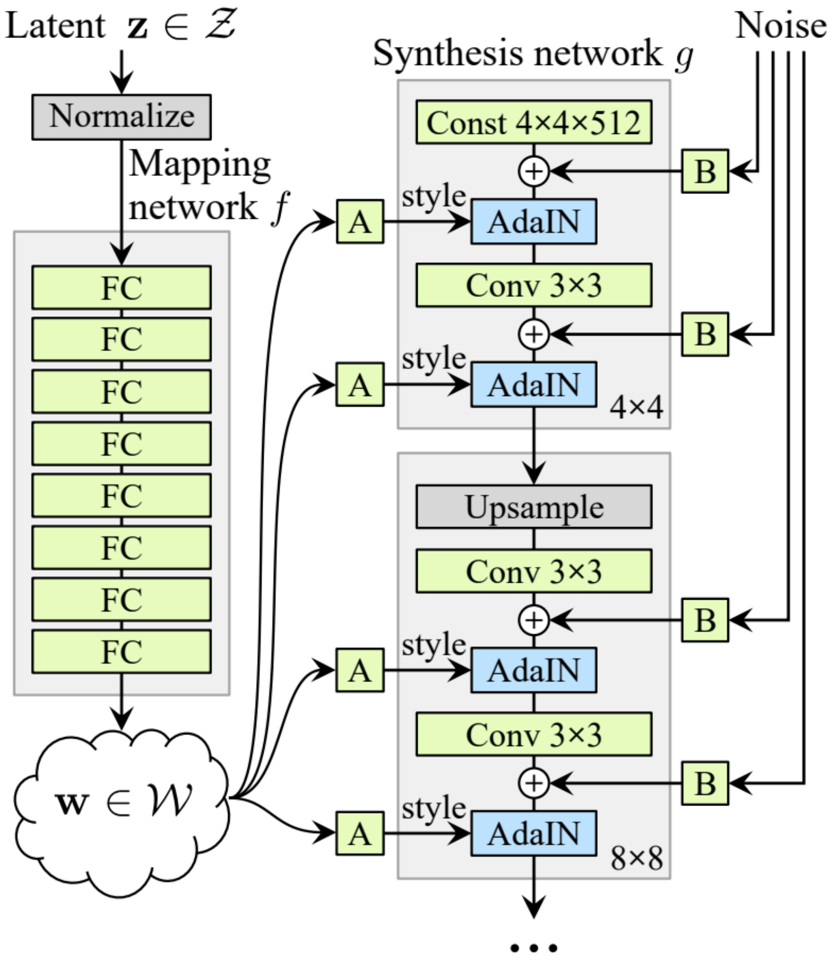
One RNN encodes a sentence into a latent representation.
2nd RNN decodes it into target language.

EikoNet: a deep learning approach to ray tracing



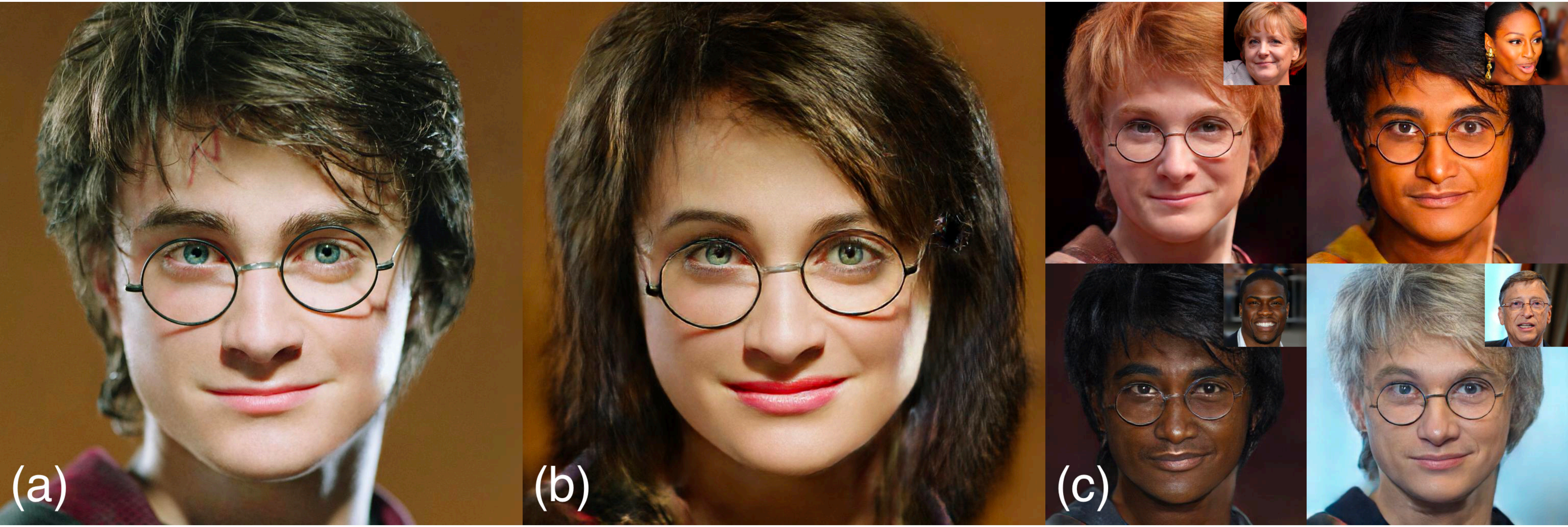
Smith et al. (2020), arXiv: 2004.00361

Generative adversarial networks are a major advance in AI



(b) Style-based generator

Karras et al. (2016)



Viazovetskyi et al. (2020)

Many types of supervised models:

- How to choose what algorithm to use?
  - Mathematical structure of the data
  - Amount of data available
  - How good is your ground truth?
  - Computational demands
  - Interpretability of learned features

- State of the art models are often not the most practical

- Requires knowledge of different algorithms (often experience-based)

Training a supervised learning model

- ML distinguishes between parameters and hyperparameters
  - Parameters characterize the model itself. Learned from the data
  - Hyperparameters affect model performance; not necessarily learned.
    - e.g. optimization, model architecture, regularization weights,

- Training a ML model is the process by which the parameters are learned

- Deep neural networks are generally trained with stochastic gradient descent

- Involves iteratively updating parameters to minimize prediction error on ground truth. Always measure performance on out-of-sample test data!

Training a supervised learning model

- How much data do I need? Very problem/model/algorithm specific.
- Often want 10x as much data as number of model parameters to fight overfitting
- Modern deep neural networks have 10^6-10^7 parameters
- Lots of regularization strategies for dealing with

What if I don’t have (much) ground truth?

- Sim2real approaches often are effective
  - Synthesize training data from a model that you trust
  - Train the model with synthetic data
  - Apply trained model to real data

- Transfer learning:
  - Train a model on a surrogate (plentiful) dataset to learn coarse structure
  - Fine tune model on limited actual dataset

Why did ML research explode over the last decade?

IMAGENET

14,197,122 images, 21841 synsets indexed

Explore Download Challenges Publications Updates About

Not logged in. Login I Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.
Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

Deng et al. (2009)

Source: www.image-net.org

Why did ML research explode over the last decade?

IMAGENET

14,197,122 images, 21841 synsets indexed

Explore Download Challenges Publications Updates About

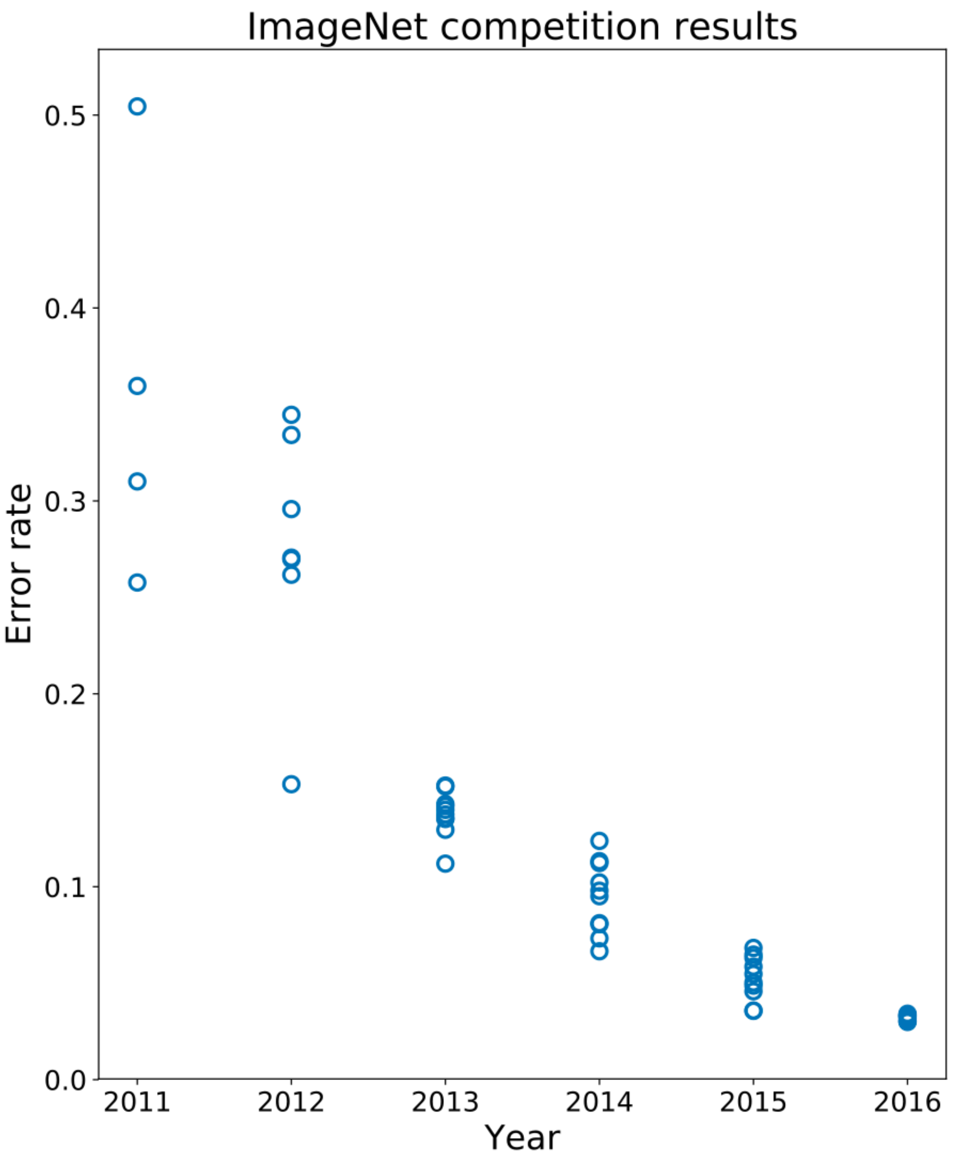
Not logged in. Login I Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.
Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

Deng et al. (2009)

Source: www.image-net.org

Why did ML research explode over the last decade?



Deng et al. (2009)

Source: www.image-net.org

Summary

- ML is a rapidly growing field. Can’t possibly cover w/ 1 lecture
  - Strongly recommend taking a formal ML course
- Has led to major advances in all areas of science and technology
- Provides tools for (e.g. Bergen et al. 2019):
  - Automation
  - Accelerating compute
  - Data discovery
- Will be a key component of geophysics research going forward