Artificial Neural Networks: On Order and Time

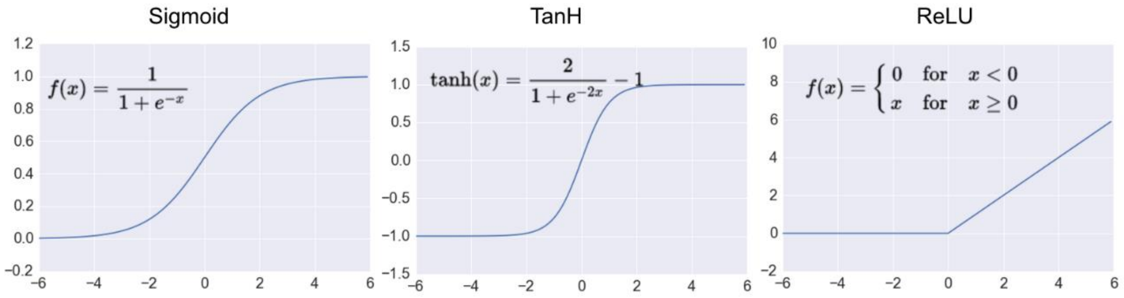
Alexander G. Ororbia II

Introduction to Machine Learning

CSCI-635

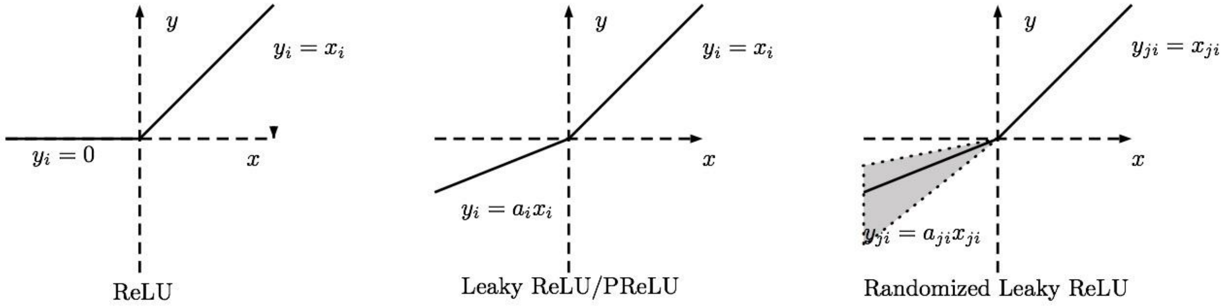
12/1/2023

Non-Standard Activations



Linear Rectified Unit (Relu)

Not smooth / not differentiable everywhere, ***Benefit***: Hard sparsity
Issues: Dead units, explosive weight updates
Parametric Relu (PRelu), leaky Relu: “Learn” slope of activation function



The magic behind deep learning

THE HUMAN-IN-THE-LOOP

Manual, Exhaustive Search

Manual Search

- Fast if you know what you are doing!
- Explore a few configurations, based on literature/heuristics
- Select lowest validation loss configuration

Deep tuning!

Grid Search

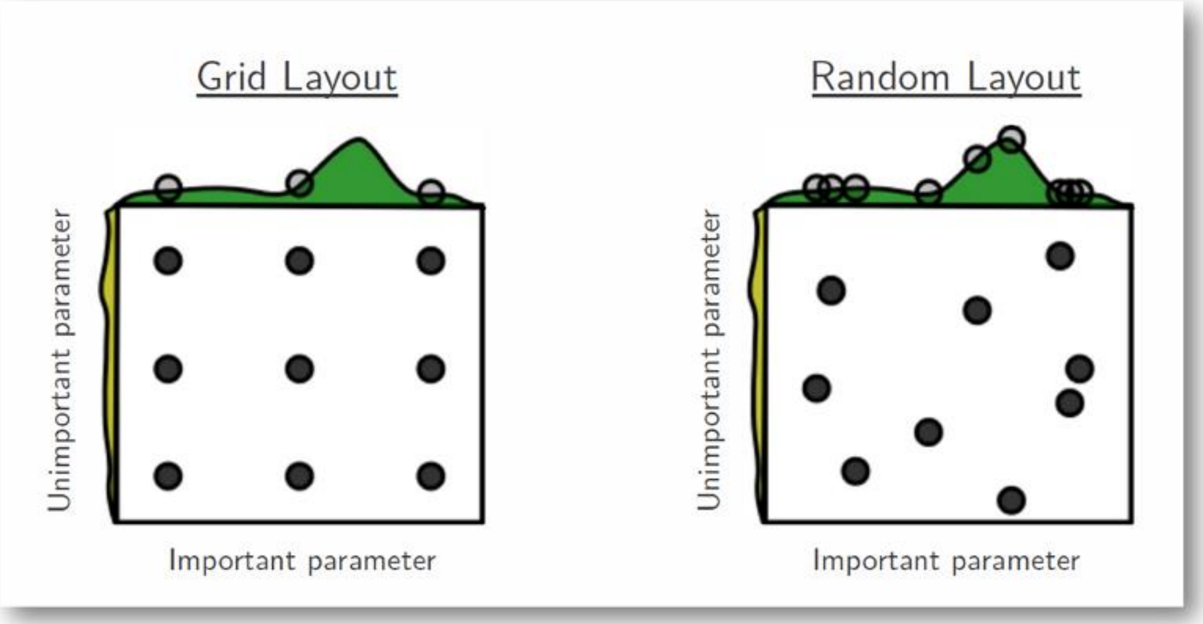
- Compose an n-dimensional hypercube, where along each axis is a hyper-parameter
  (length determined by max & min values to explore)
- Exhaustively calculate loss/error for each configuration (or combination of meta-parameter values) in hypercube
  - Choose lowest error/minimal loss configuration as optimal model
  - Loss/error is calculated on a held-out validation/development set (or in held-out set in cross-fold validation schemes)
- Will ultimately find optimal model (given coarseness of grid-search)
  - Takes long time!

Random Search

Draw *k* sample configurations from hypercube & calculate validation loss for each (w/o replacement)

- Repeat *T* trials, can use optimal of each trial to inform subsequent trials
- Could "guide" or "target" next set of random samples based on best last found point (a guided stochastic search)

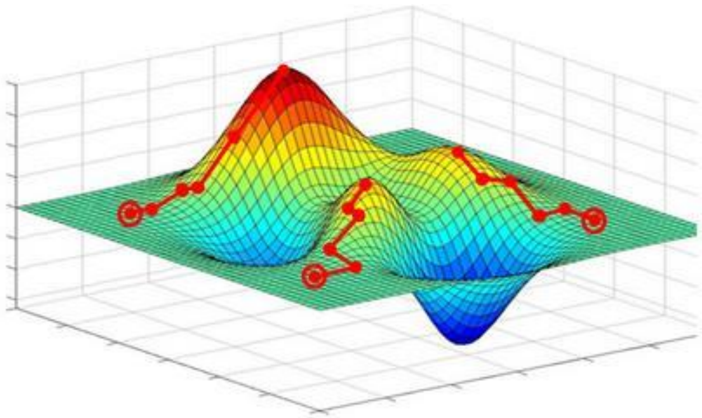
Surprisingly effective (over manual search) & faster than grid search



Bayesian Optimization: Meta Machine Learning

Use machine learning to do your research for you…

- Sequential Model Optimization (*SMO*)
- Gaussian Processes for surface-response modeling
- Gradient-based: Use another ANN
  - How do we tune this higher-level parametric model?
  - ***Meta-meta-meta-….-machine learning??***



High-level idea:

- Build a meta-model (w/ some prior that encodes intuition about hyper-parameter space)
- Draw samples from space (i.e., run few model configurations)
- Update meta-model using these samples
- Meta-model selects next best point to evaluate
  - Balancing criterion, i.e., minimal error & minimal compute time

Bayesian Optimization: Meta Ma… e Learning

Use machine learning to d… …

Sequential Model O

Gaussian P

Gradient-b

How do we

parametric

Meta-meta-

High-level ide

Build a meta-m

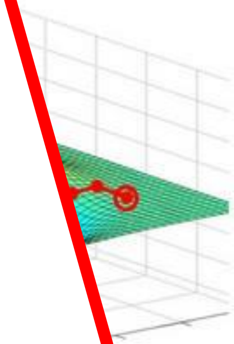
encodes intuiti

Draw samples fr

Update meta-mo

M… selec

erion, … compute time



Now all we need is for our models to write our papers for us, submit them to the conferences on our behalf, and…

…then go to class for us too! Neat!

Deep Thinking!

It is a matter of posing the problem

What is the low-level representation of your sample?
(i.e., low-level features, inputs, or sensors)
Is there an output we are interested in?
- Regression: a real-valued target
- Categorization: a discrete target

Deep digit recognition!

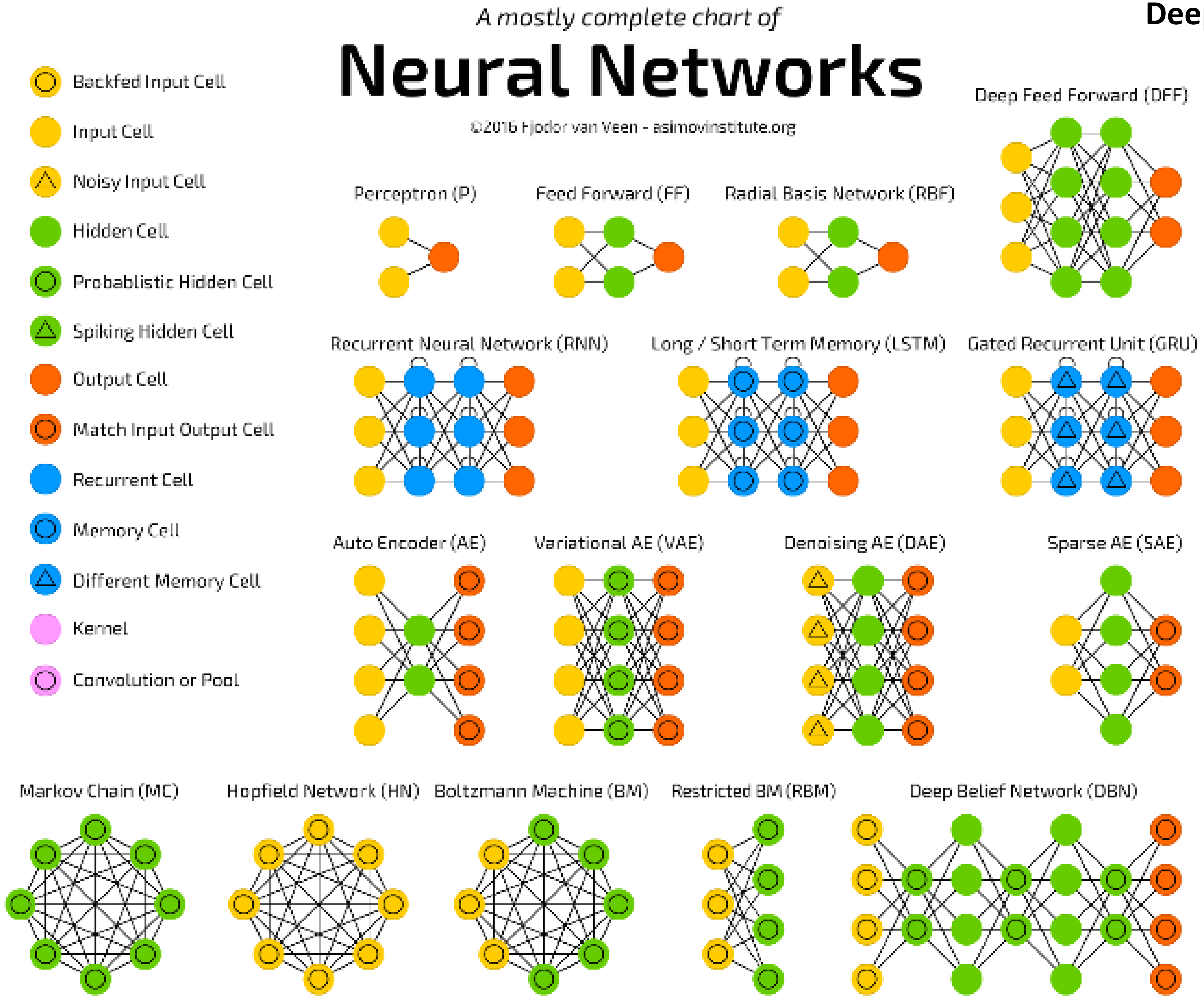
How much data do you have?

More data is better! (MNIST is 60K)
Only a small sample?
Go Bayesian Neural Networks!

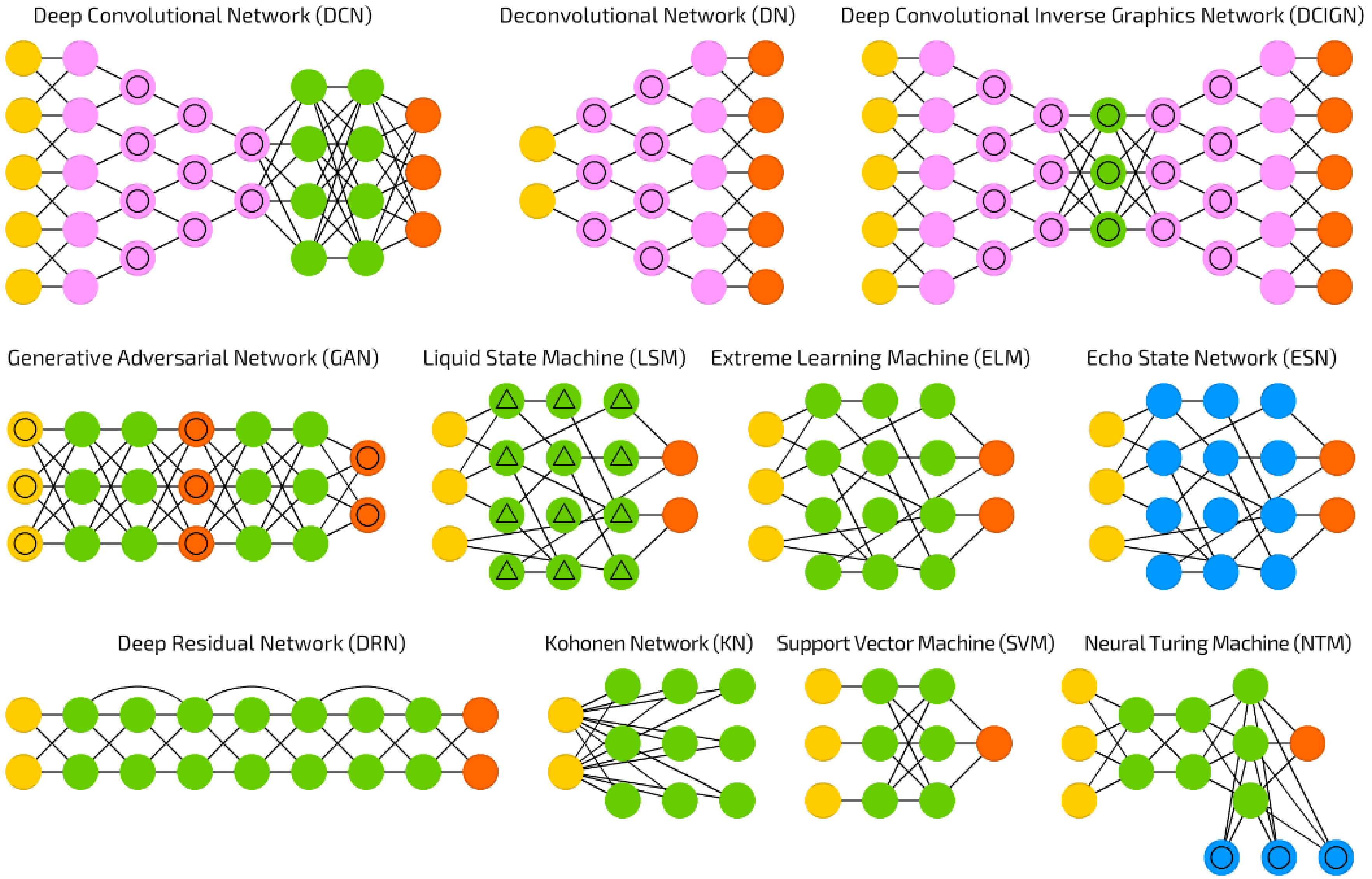
What kind of hardware do you have?

Multi-CPU settings
GPUs
Specialized hardware?
FPGAs, TPUs?

Deep zoo!



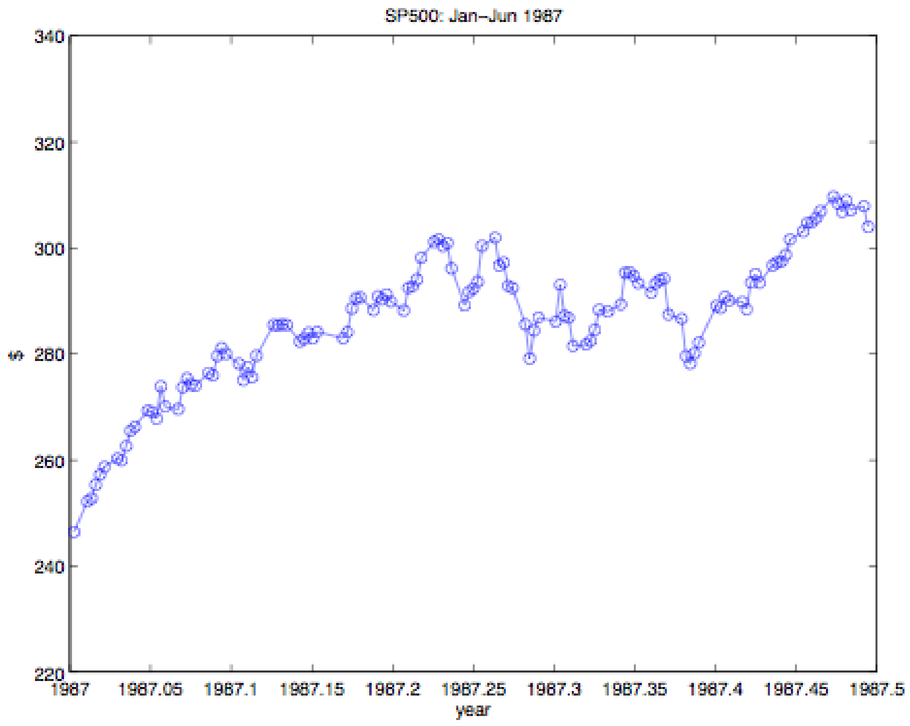
http://www.asimovinstitute.org/neural-network-zoo/



http://www.asimovinstitute.org/neural-network-zoo/

Violating the first ‘i” in i.i.d….

RECURRENT NEURAL NETWORKS


SP500: Jan–Jun 1987
340
320
300
280
260
240
220
$
1987
1987.05
1987.1
1987.15
1987.2
1987.25
1987.3
1987.35
1987.4
1987.45
1987.5
year

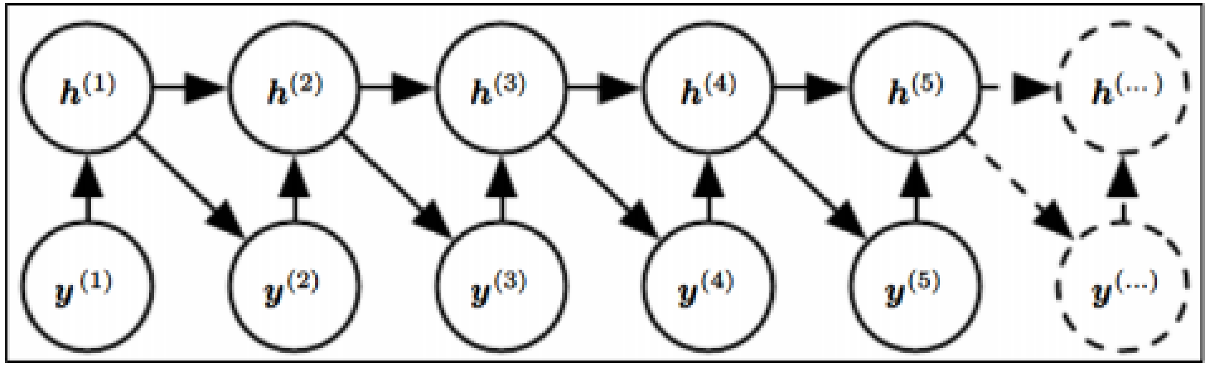
RNNs process sequential data

- Recurrent Neural Networks are a family of neural networks for processing sequential data
- RNN and CNN are both specialized architectures
  - RNN is specialized for processing a sequence of values $\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(\tau)}$
    - Just as CNN is specialized for processing a grid of values such as an image
  - RNNs can scale to much longer sequences than would be impractical for networks without sequence-based specialization
  - RNNs can also process variable-length sequences
    - Just as a CNN can scale to images with large width/height and process variable size images

White Board Time!

(Turning MLPs into RNNs)

Efficient parameterization based on $h^{(t)}=f(h^{(t-1)},x^{(t)};\theta)$



QUESTIONS?