Intro. to (Adversarial) Machine Learning

Ben Zhao, Blase Ur, David Cash

November 26, 2018

CS 232/332

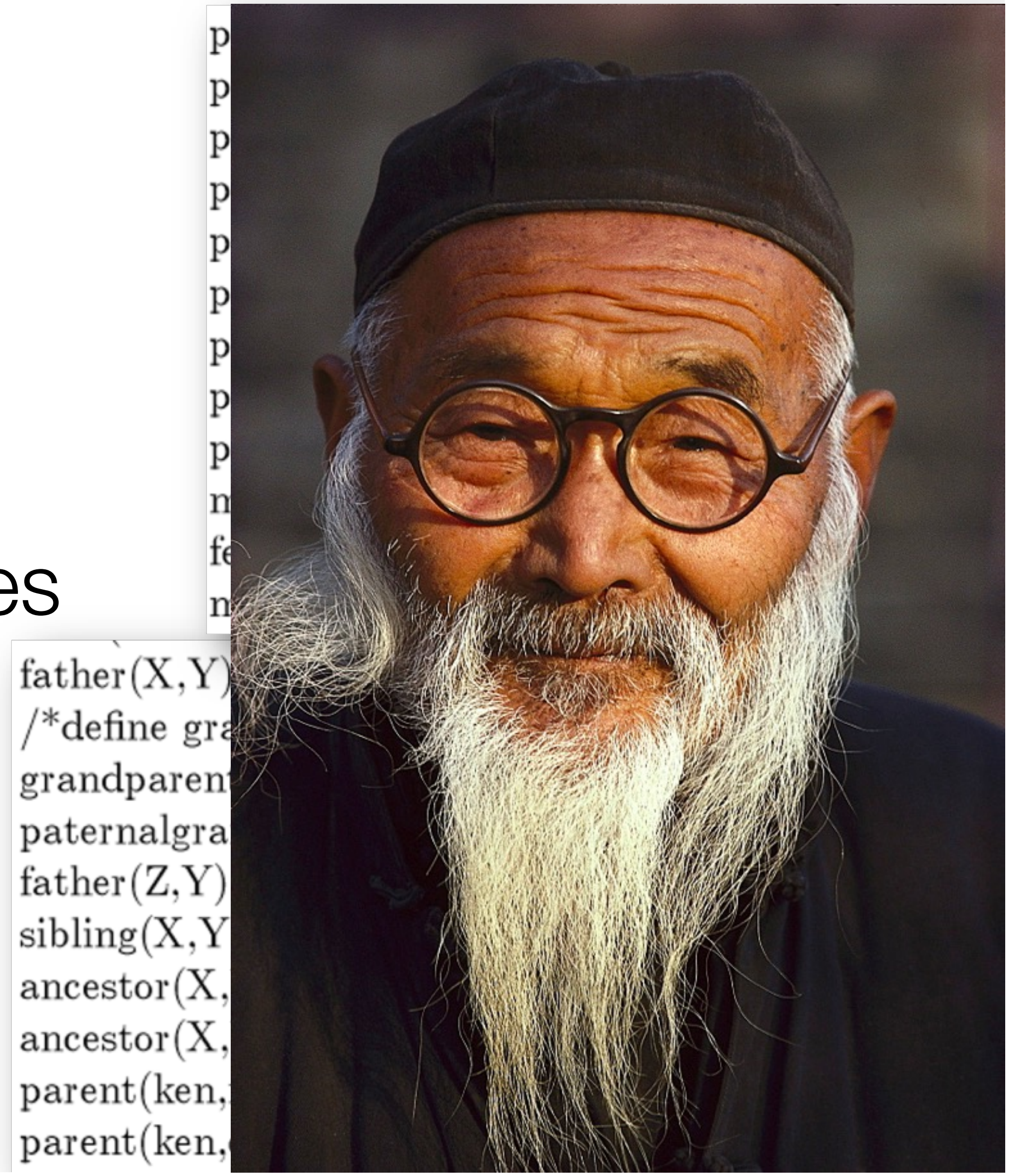
THE UNIVERSITY OF CHICAGO

This Week: Whirlwind Flyover of ML

- Today
  - What is machine learning?
  - Learning system models
  - Linear classifiers
  - Deep neural nets
  - Basic Attacks: poisoning and evasion
- Wednesday
  - Advanced adversarial attacks

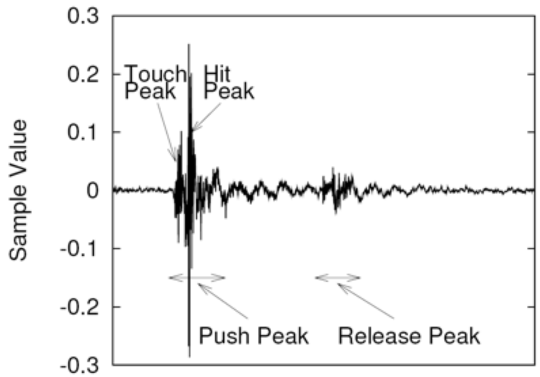
ML Has Come a Long Way…

- Artificial Intelligence in1996
- Making choices by searching through all possible outcomes
- Prolog!



ML Has Come a Long Way…

- Early 2000's
  - Attention shifted to classification problems
- Statistical ML takes off with surprising results
  - Decision trees, SVMs, Bayes, HMMs
  - *e.g.* recovering random passwords from keyboard acoustics (unsupervised learning on HMMs)
  - *e.g.* reconstructing images from monitor reflections

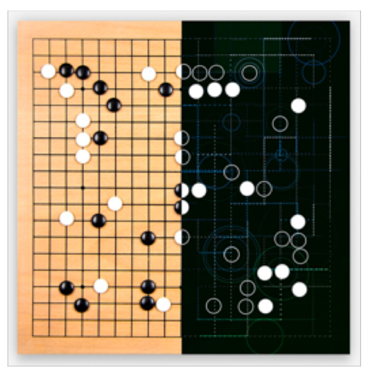
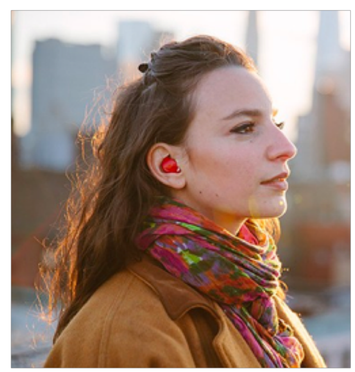


Machine Learning (& AI) Today …

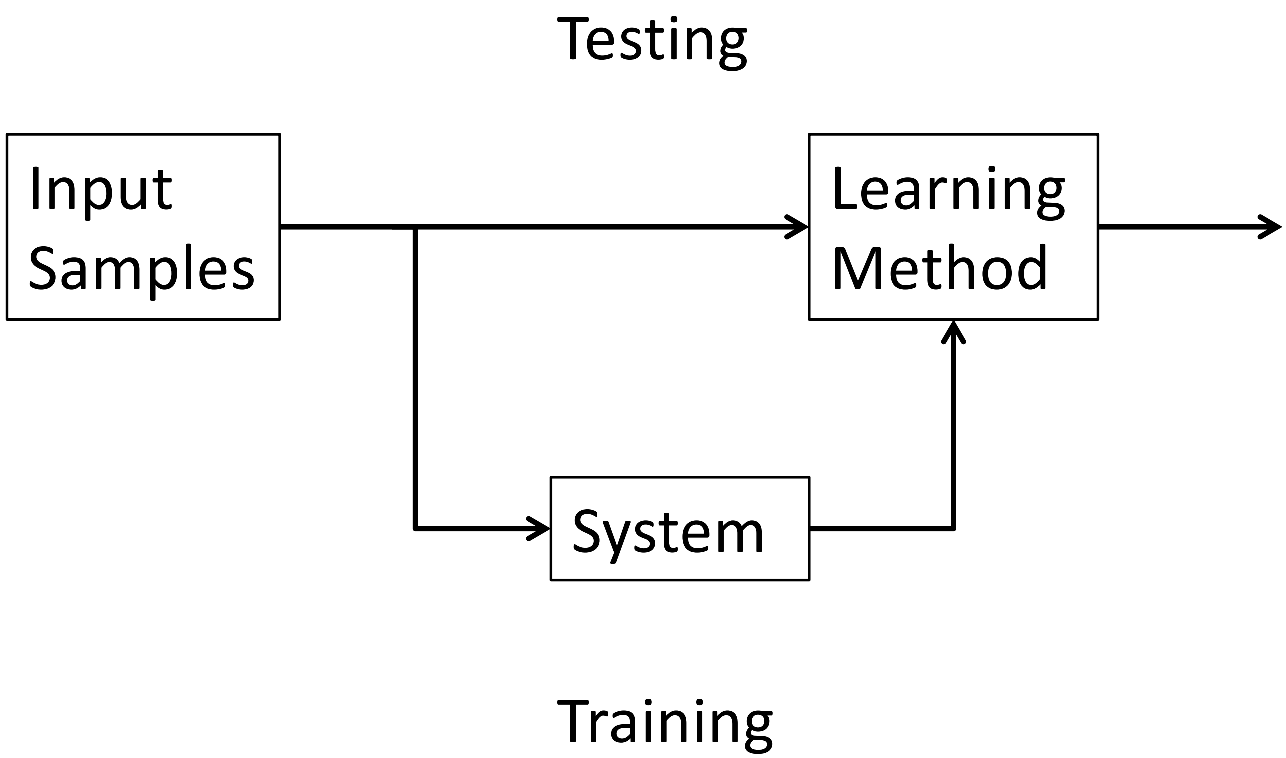
- 2018: Everything is better with deep learning!
  - Real time voice translation
  - Recommendation systems
  - Fraud monitoring
  - Multimedia synthesis and manipulation



- R&D focused on accelerating DL

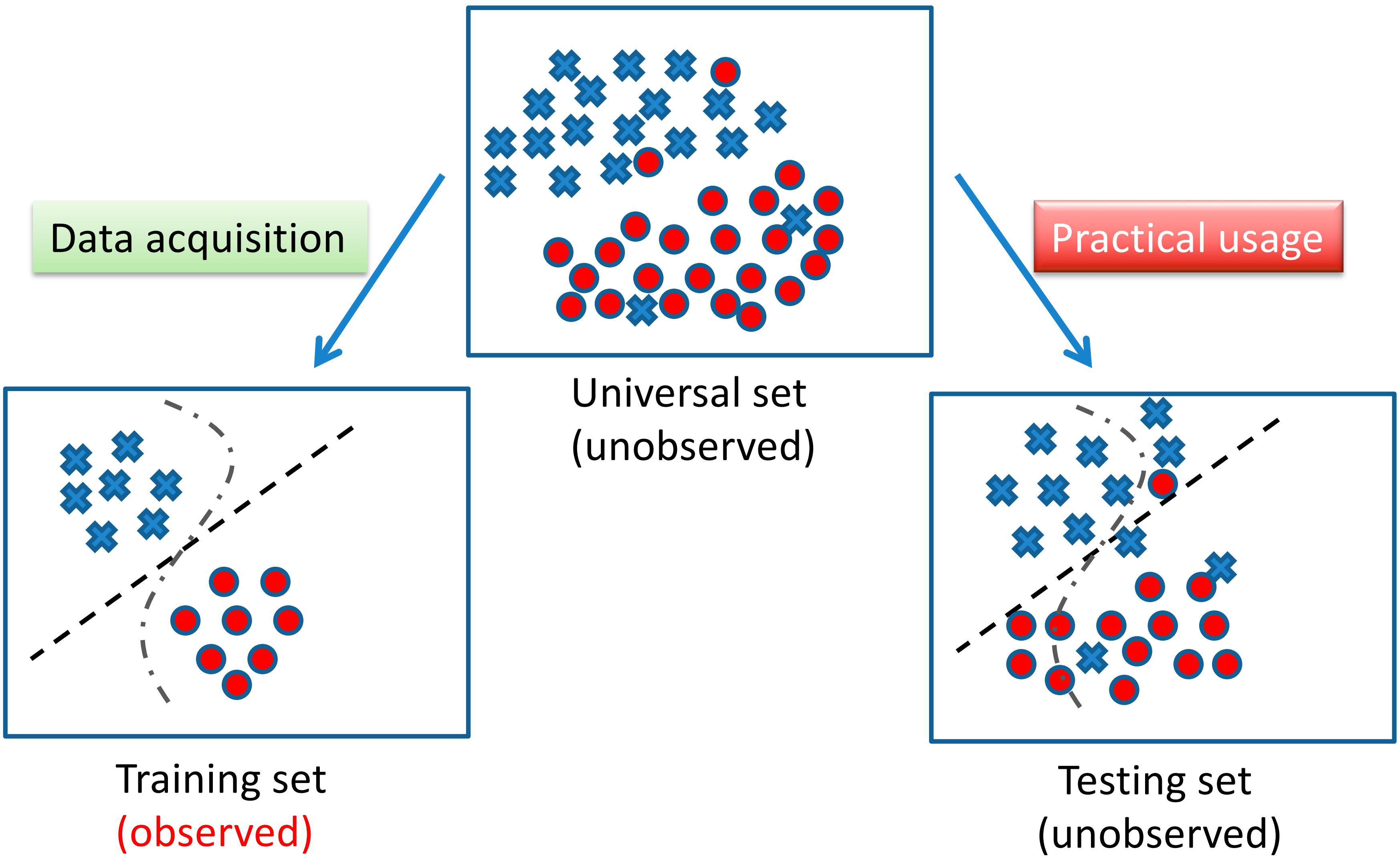




Learning System Model



(Supervised Learning)

Training Models from Data



Training and Testing

- Training: process of making system able to learn

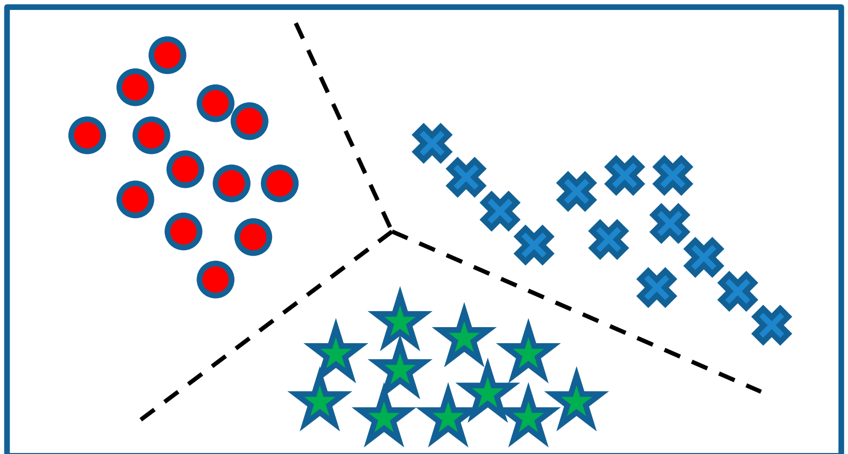
- No free lunch rule
  - No one model works best for all problems

  - Model: simplified representation of reality that discards unnecessary details (based on assumptions we make)

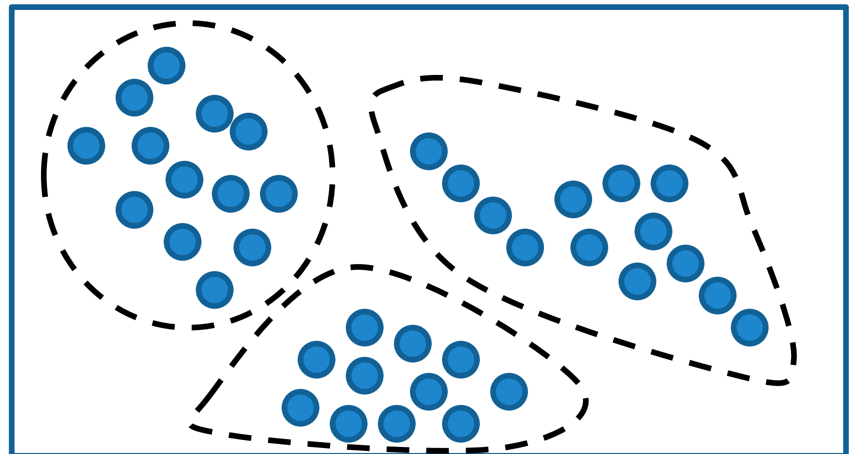
Algorithms

- Supervised learning
  - Prediction
  - Classification (discrete labels), Regression (real values)
- Unsupervised learning
  - Clustering
  - Probability distribution estimation
  - Finding association (in features)
  - Dimension reduction
- Semi-supervised learning
- Reinforcement learning
  - Decision making (robot, chess machine)

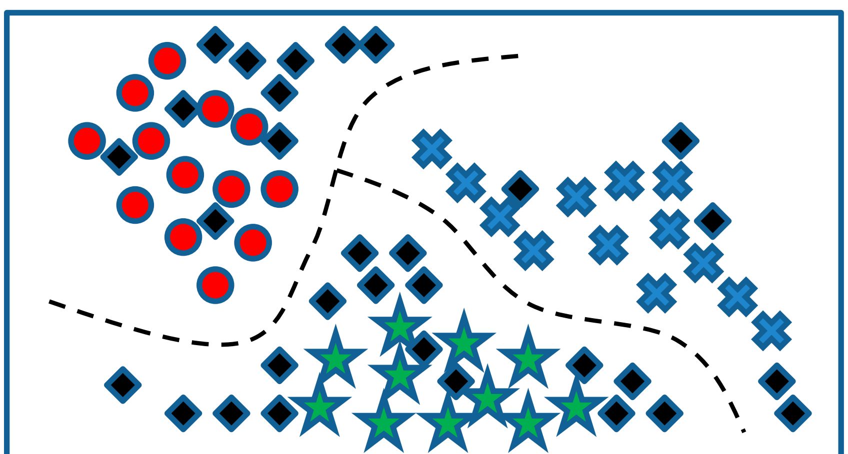
Algorithms



Supervised learning



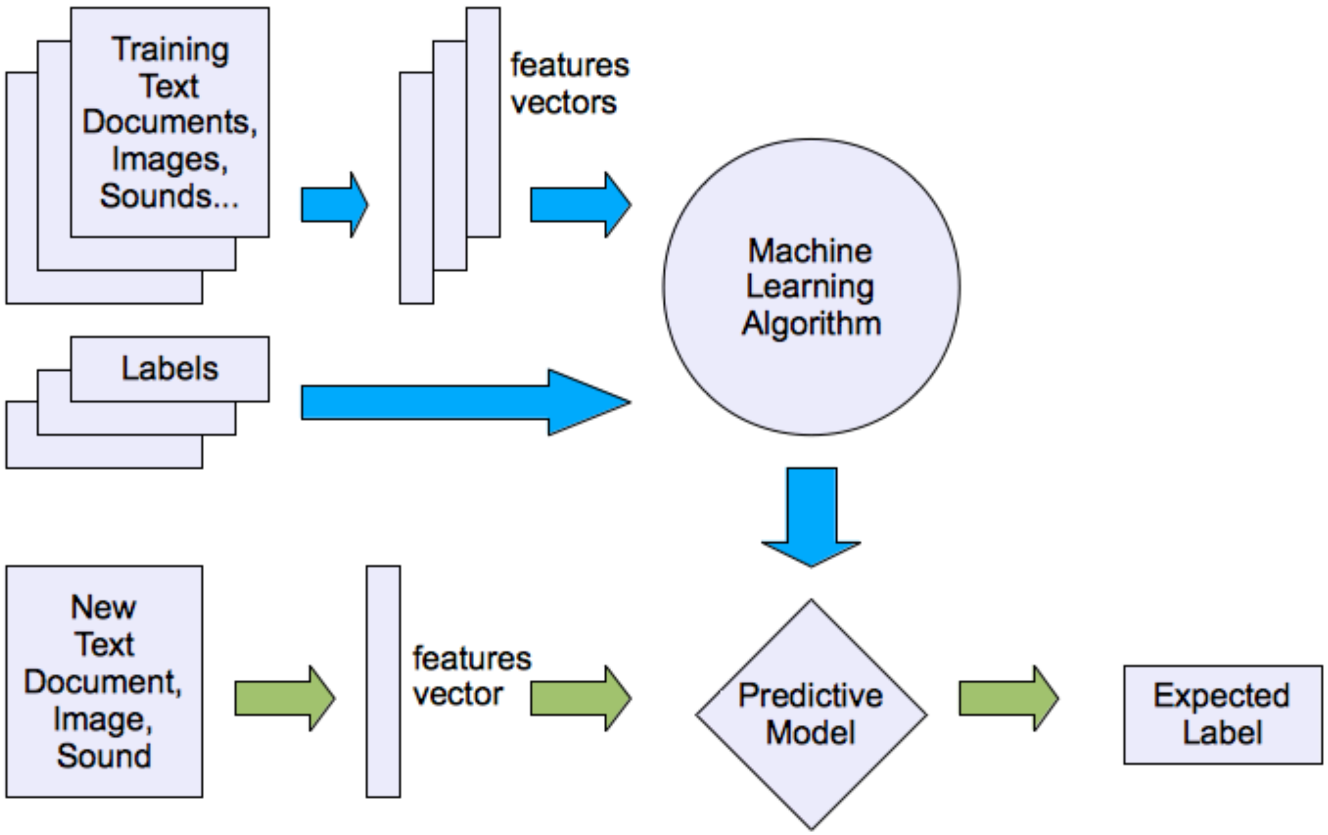
Unsupervised learning



Semi-supervised learning

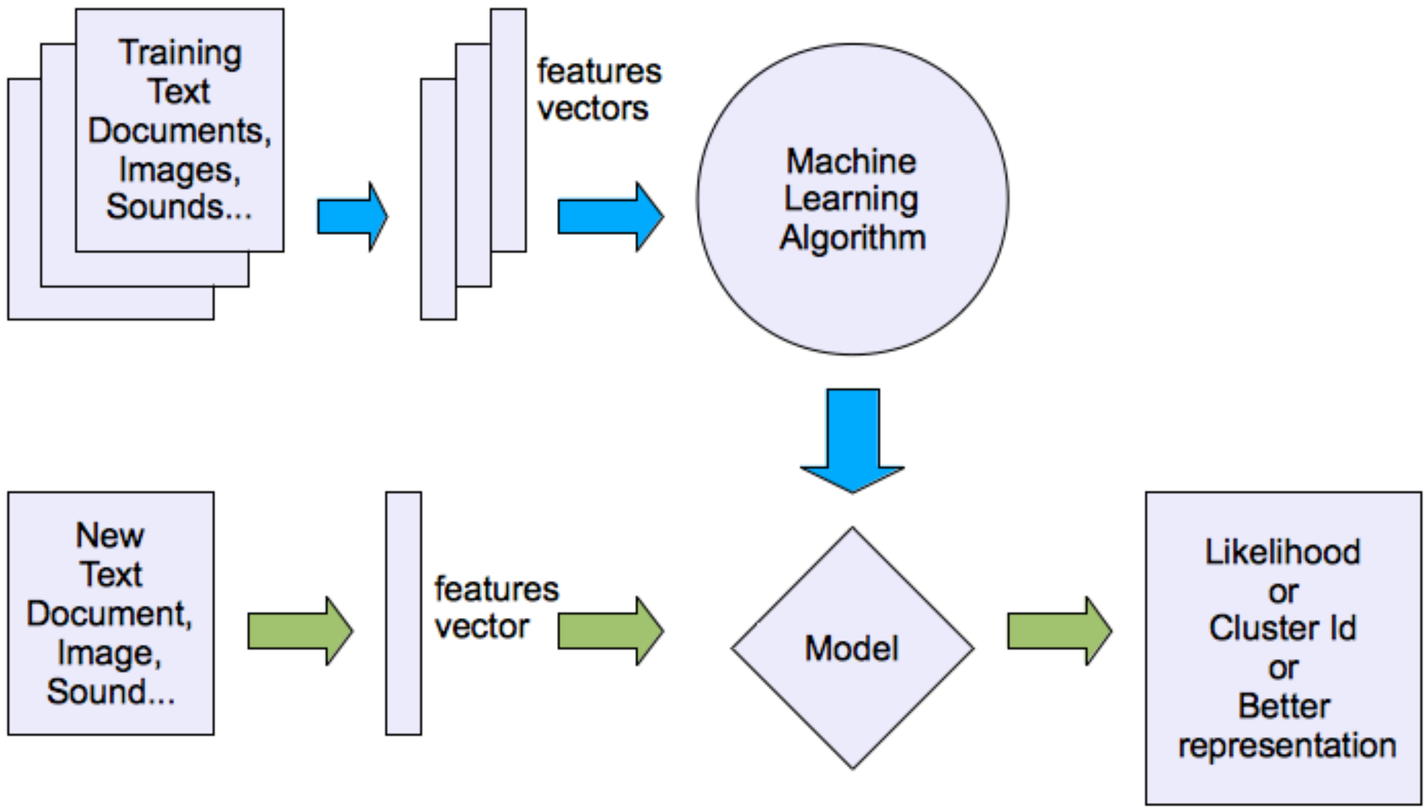
Machine learning structure

- Supervised learning



Machine learning structure

- Unsupervised learning



What are we seeking?

- Supervised: Low E-out or maximize probabilistic terms

$$error = \frac{1}{N}\sum_{n=1}^{N}[y_n \neq g(x_n)]$$

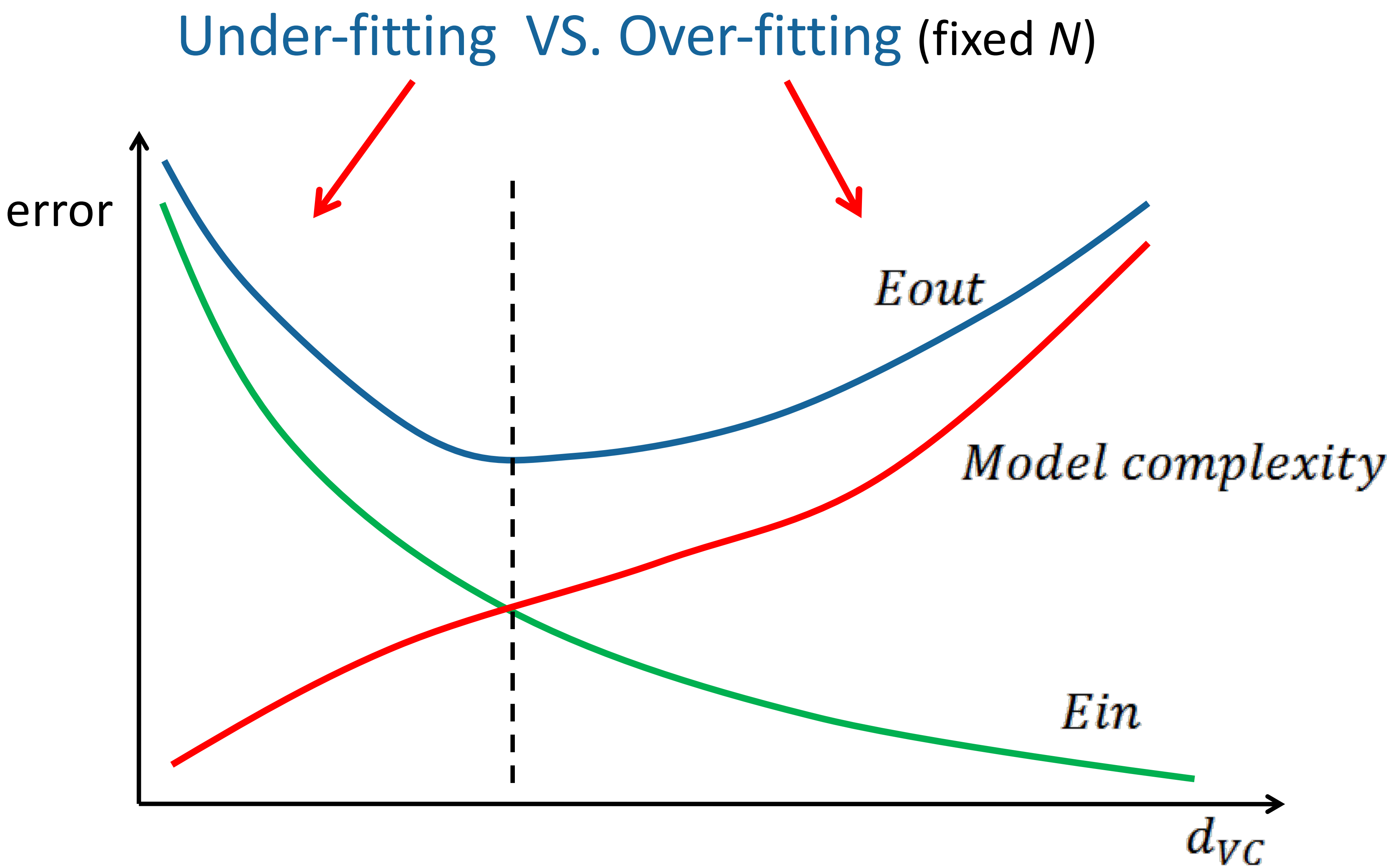
E-in: for training set

E-out: for testing set

$$Eout(g) \leq Ein(g) \pm O\left(\sqrt{\frac{d_{VC}}{N} \ln N}\right)$$

- Unsupervised: Minimum quantization error, Minimum distance, MAP, MLE(maximum likelihood estimation)

What are we seeking?

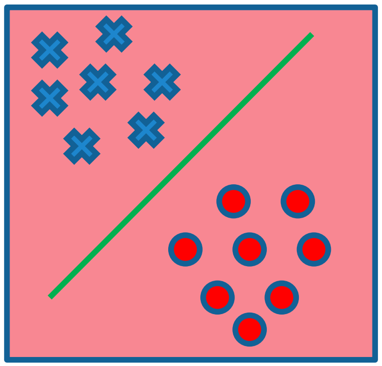


Learning Techniques

- Supervised learning categories and techniques
  - **Linear classifier** (numerical functions)
  - **Parametric** (Probabilistic functions)
    - Naïve Bayes, Gaussian discriminant analysis (GDA), Hidden Markov models (HMM), Probabilistic graphical models
  - **Non-parametric** (Instance-based functions)
    - K-nearest neighbors, Kernel regression, Kernel density estimation, Local regression
  - **Non-metric** (Symbolic functions)
    - Classification and regression tree (CART), decision tree
  - **Aggregation / ensemble methods**
    - Bagging (bootstrap + aggregation), Adaboost, Random forest

Learning Techniques

- Linear classifier



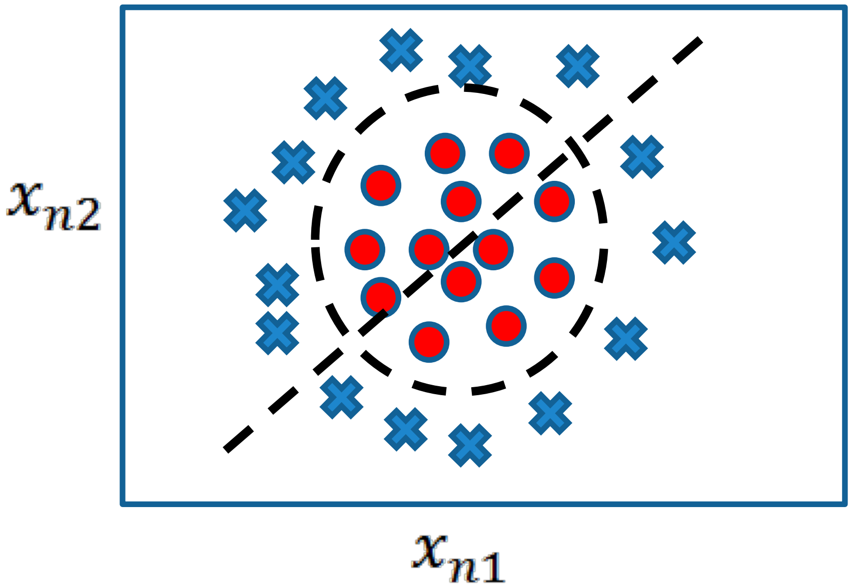
$$g(x_n) = sign(w^T x_n)$$

, where *w* is an *d*-dim vector (learned)

- Techniques:
  - Perceptron
  - Logistic regression
  - Support vector machine (SVM)
  - Ada-line
  - Multi-layer perceptron (MLP)

Learning Techniques

- Non-linear case



$$x_n = [x_{n1}, x_{n2}]$$

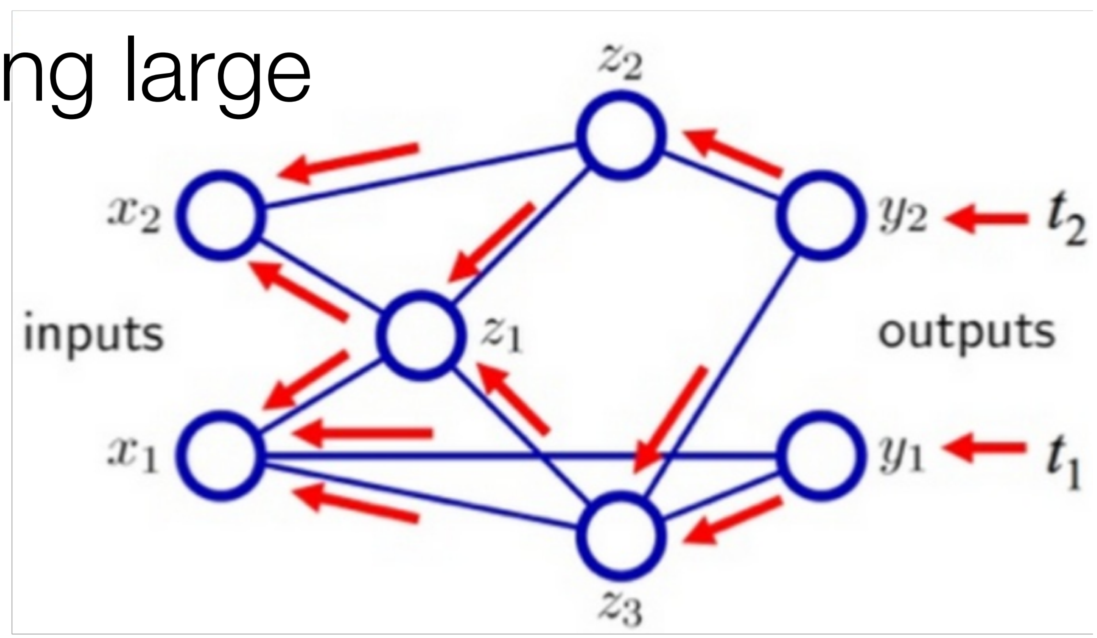
$$x_n = [x_{n1}, x_{n2}, x_{n1} * x_{n2}, {x_{n1}}^2, {x_{n2}}^2]$$

$$g(x_n) = sign(w^T x_n)$$

- Support vector machine (SVM):
  - Linear to nonlinear: **Feature transform** and **kernel function**

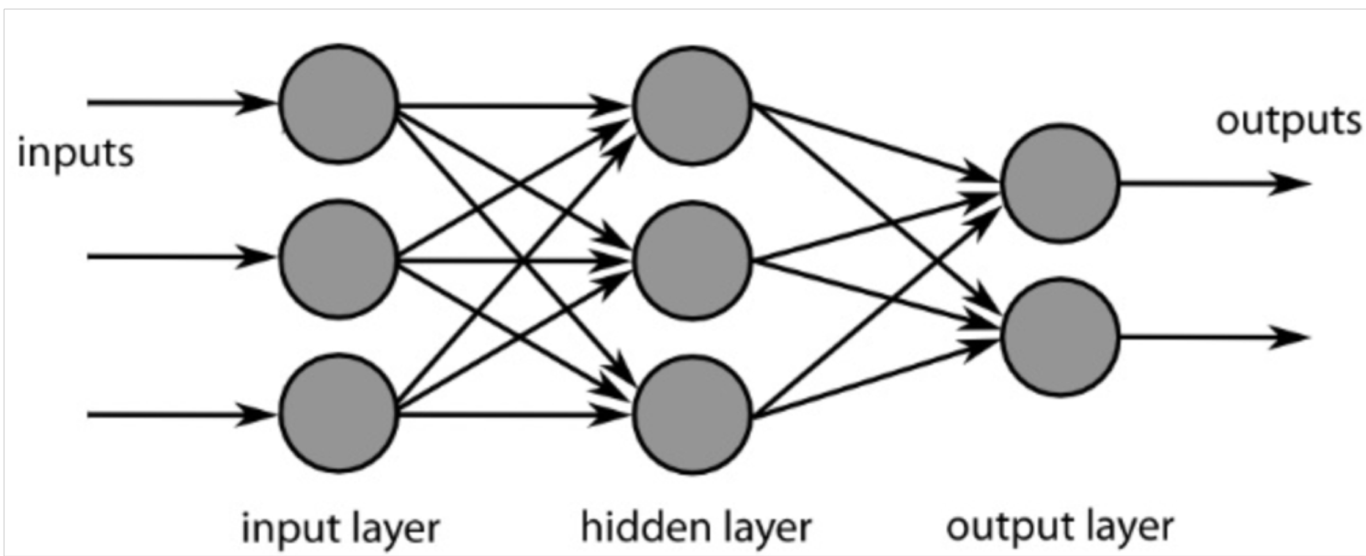
Deep Neural Networks

- Powerful models that try to emulate human neurons
- Multi-layers of neuron/units
  - (Mostly) linear combinations
- Iterative training using large labeled datasets
  - Backpropagation



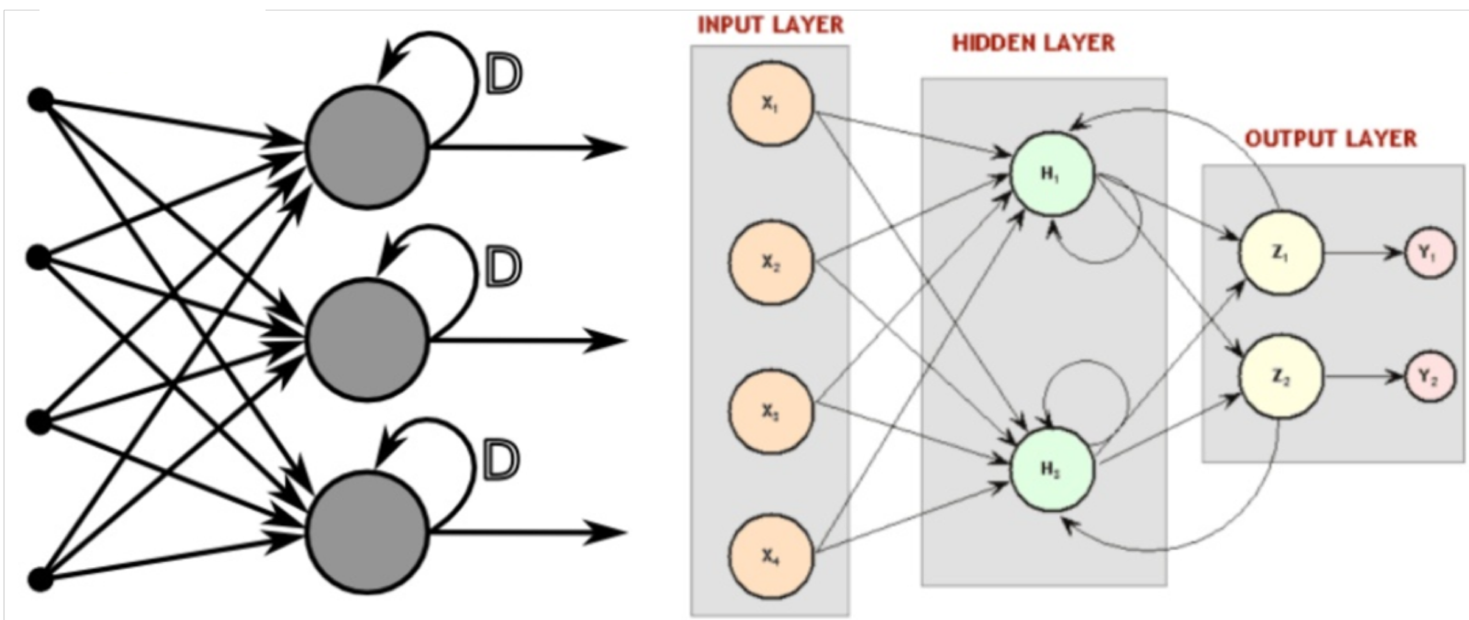
DNN Architectures: CNNs

- "Convolutional," feed-forward neural networks
  - Connections between units do not form directed cycle
  - "traditional" DNNs focused on image recognition

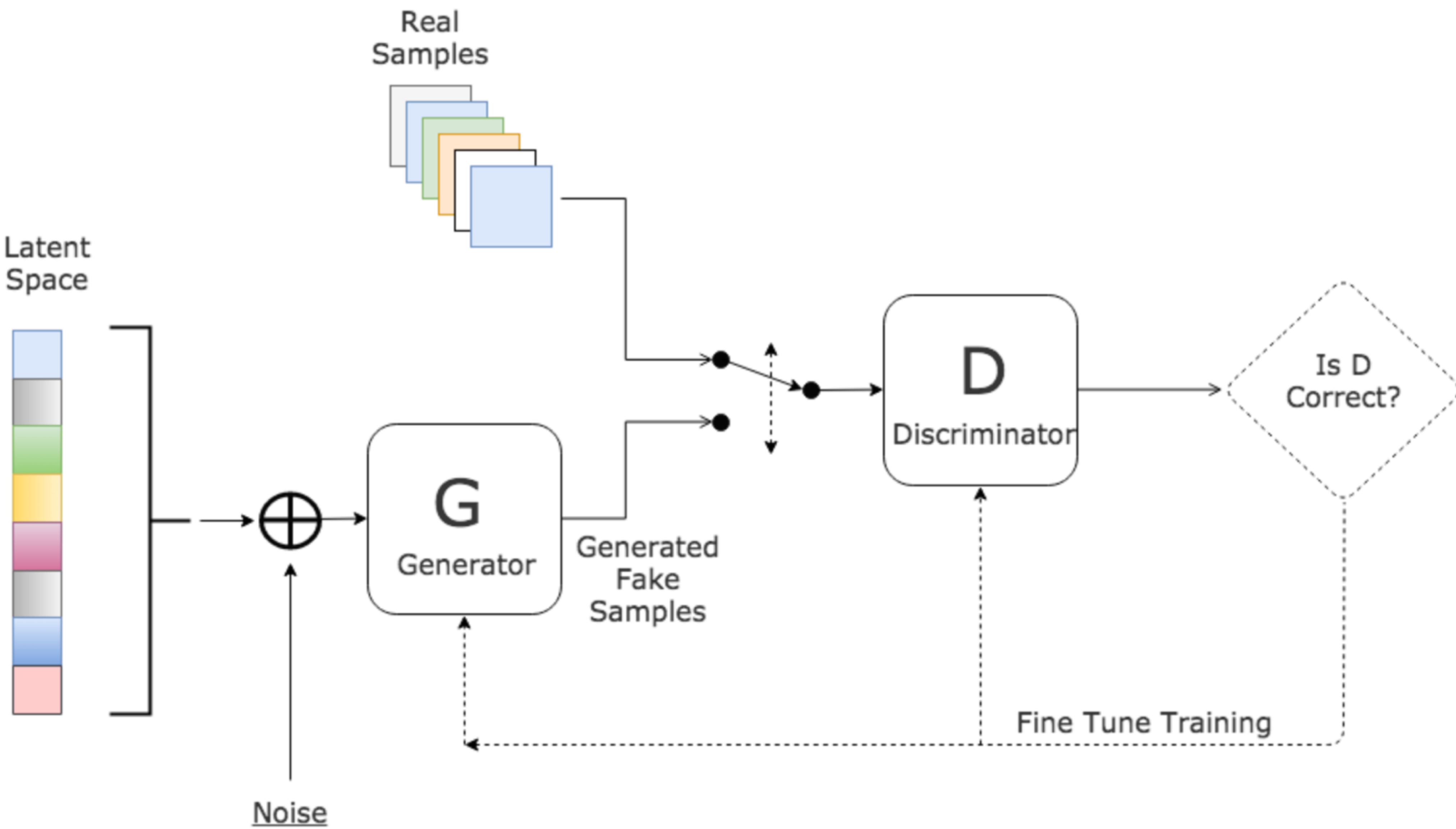


DNN Architectures: RNNs

- Recurrent neural nets (RNNs)
  - Most popular: Long/short-term Memory (LSTMs)
  - Designed for capturing sequences, e.g. language, handwriting, temporal data

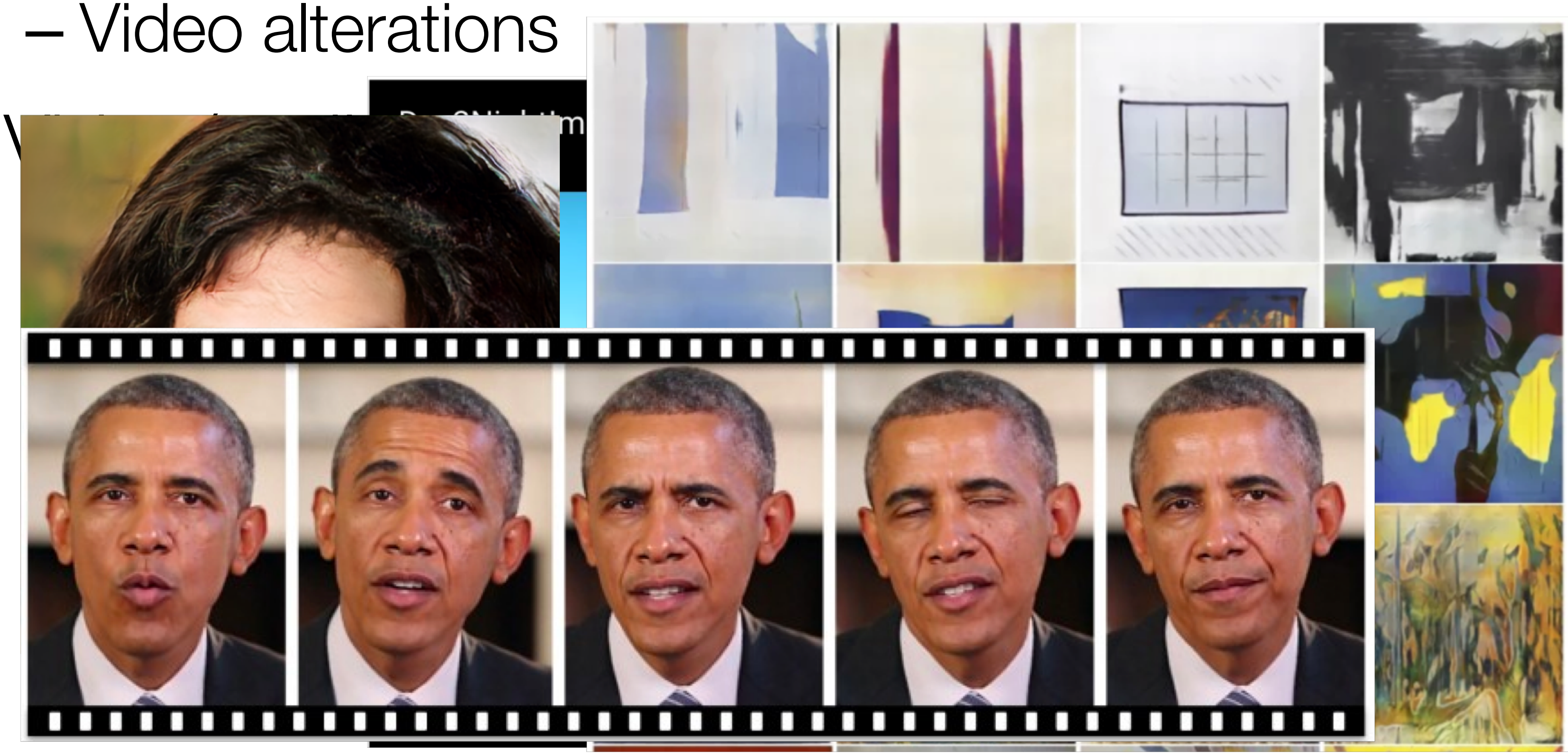


DNN Architectures: GANs



Media Manipulation using DNNs

- Much of this in last 16 months using GANs
  - Content generation
  - Video alterations
- V



Applications? What Can't ML Do?

- Medicine: cancer / disease diagnosis
- Robots, mobile devices, self-driving cars
  - Object detection and recognition
- Computer vision
  - Weapons systems, surveillance

(Medicine through surveillance, bracketed:) introducing errors

- Network & systems management / network attacks
  - Network IDS, anomaly detection, malware detection
  - Resource allocation: e.g. TCP congestion control
- Human behavior modeling / human mimicry
- Financial fraud detection / fraud generation
- Automating the law / manipulating the law

General Attacks on ML Systems

- Attack model
  - White box vs. black box
  - Access to training?
- Basic attacks
  - Data poisoning
  - Evasion
- Case study: ML detection of malicious crowdsourcing workers/campaigns
  - *Man vs. Machine: Adversarial Detection of Malicious Crowdsourcing Workers*, Wang et al, USENIX Security 2014