Machine Learning

Amir Globerson - Tel Aviv University

Learning in neural computation & cognition

- One of the hallmarks of cognition
- Still not sufficiently understood
- Machine learning is a field that:
  - Designs algorithms for learning
  - Builds theory for understanding learning
  - Not necessarily biologically or cognitively inspired

State of the Art

- Face recognition - Human accuracy
- Text based image search - Works very well in some cases.
- Speech recognition - Functional for search.

Costumes?

- Game playing.

- Self driving cars

Challenges

- Translation:

On November 9, Israeli Prime Minister Benjamin Netanyahu congratulated President-elect Donald Trump through a video message, in which the Israeli leader could barely contain his giddiness at the prospect of a friendlier White House.

ב 9- בנובמבר, ראש ממשלת ישראל, בנימין נתניהו, בירך הנשיא הנבחר דונלד טראמפ באמצעות הודעת וידאו, שבו המנהיג הישראלי הצליח אך בקושי להכיל הסחרחורת בפרוספקט של הבית הלבן ידידותית יותר.

On 9 November, Israeli Prime Minister Benjamin Netanyahu congratulated President-elect Donald Trump through a video message, in which the Israeli leader was able to barely contain dizziness prospect of the White House more friendly.

- Conversational agents (e.g., Turing test)
- <u>Robotics</u>: Assistance, chores, risk

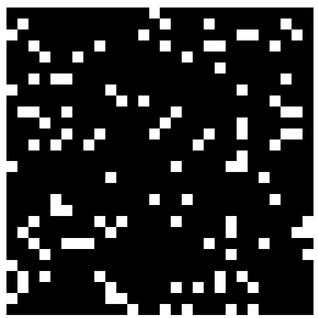
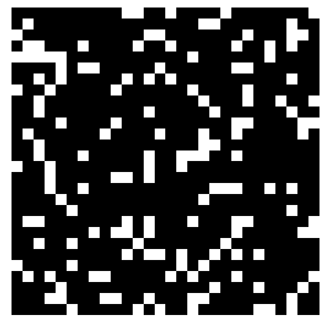
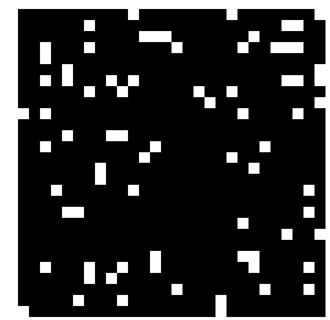
ML Research

- Always ask: how can I use data to learn rules
- Considerations:
  - What type of data do I have?
  - What rules do I want to learn?
- Abstract away…
  - Define a clean mathematical formulation
  - Design algorithms that make sense
  - Analyze their generalization properties

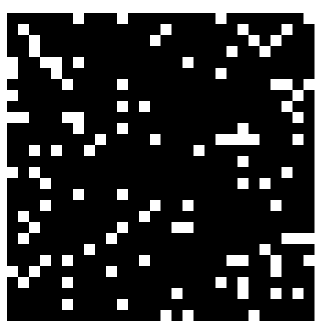
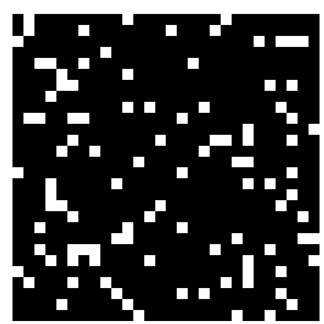
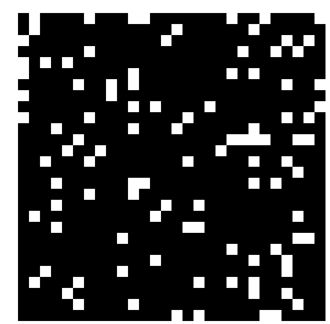
A learning problem

- Here are two classes.

- Class 1

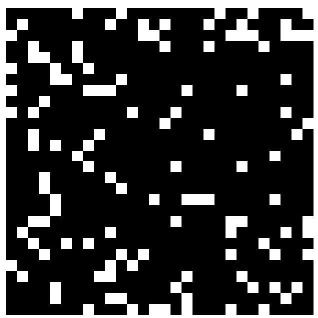




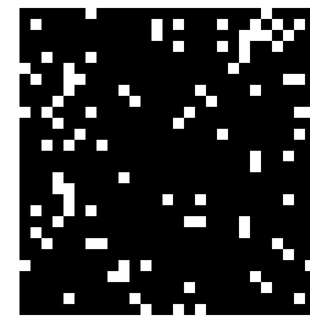
- Class 2

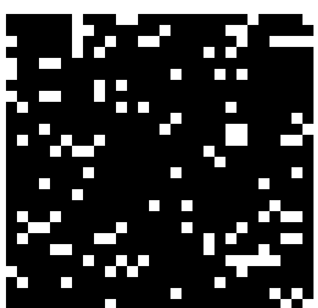
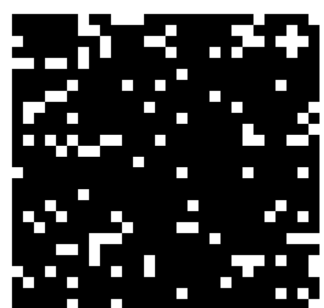
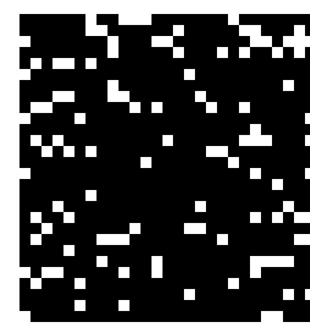
Test

- Class 1

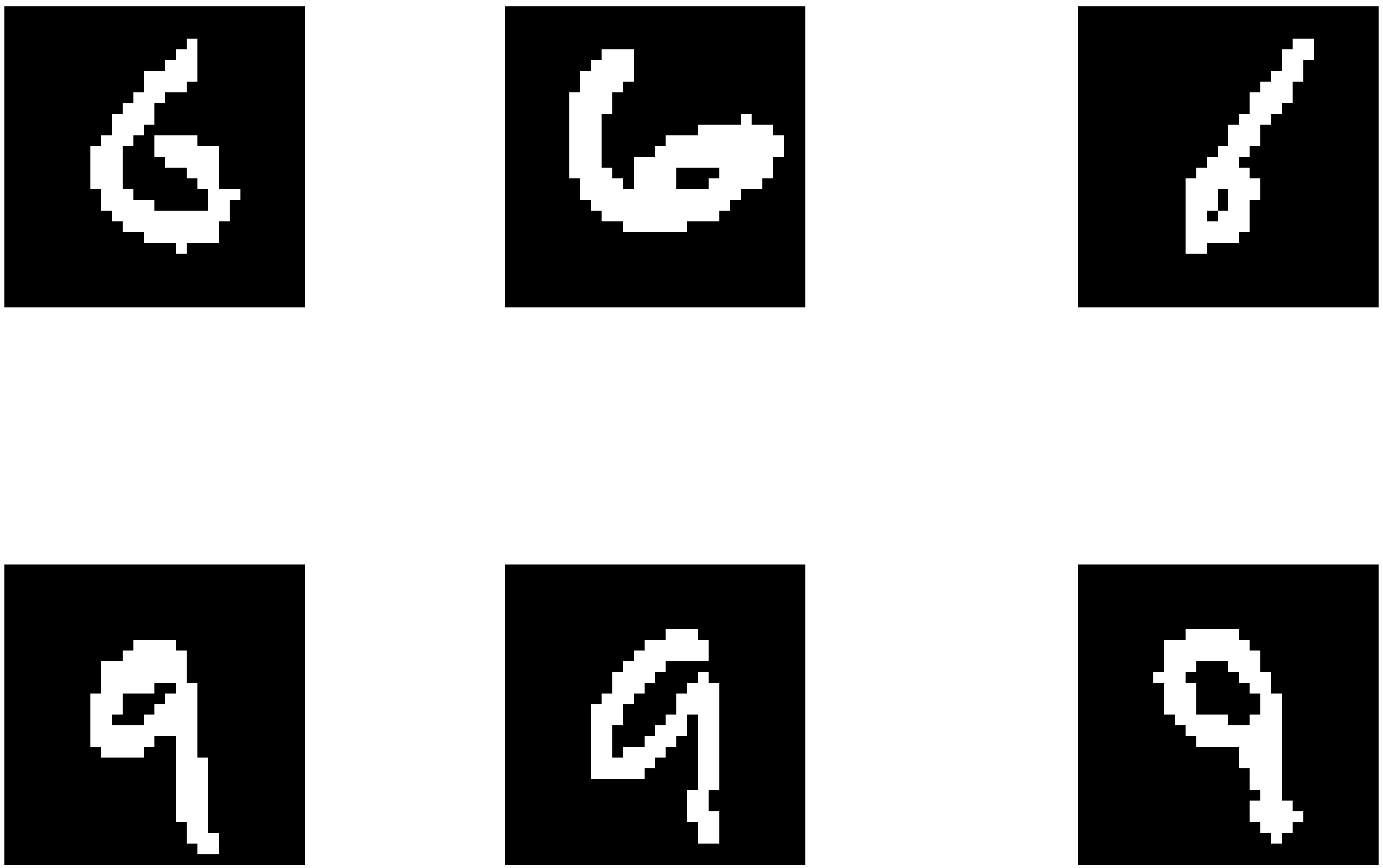




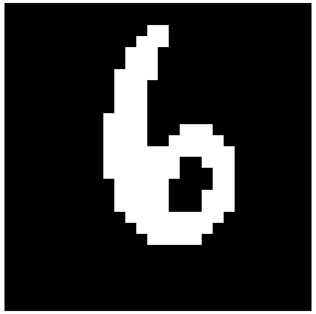
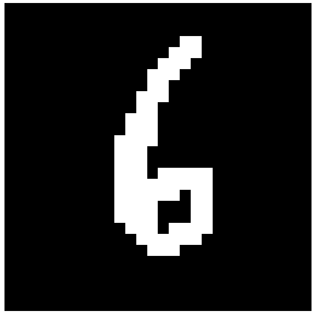
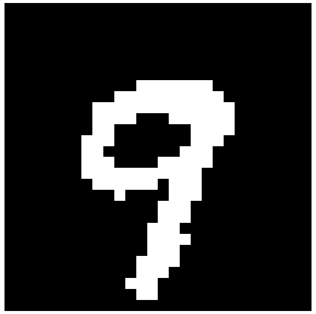
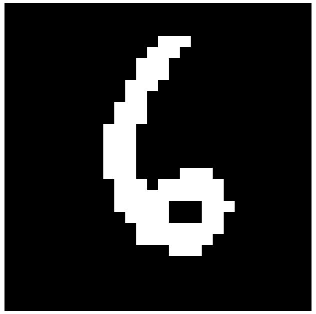
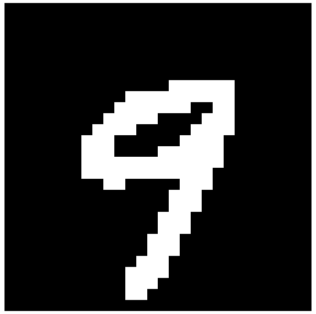
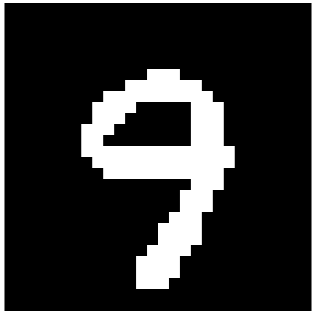
- Class 2

Train After reshuffling

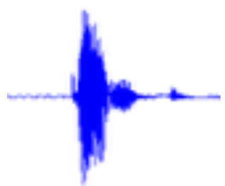


Test

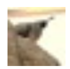
Learning mappings

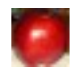
- Learn mappings between inputs and interpretations

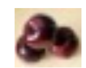
	Wash the dishes
Wash the dishes	שטוף את הכלים
	Plates
	Malignant Tissue
Water the plants	

Supervised Learning

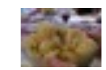
Labeled Data

 quail

 apple

 apple

 corn

 corn

Features

1.1 -0.5 0 0 0.3 …	-1 0 1.2 -0.4 0.1 …	1.1 -0.5 0 0 0.3 …	-1 0 1.2 -0.4 0.1 …	x
quail	apple	apple	corn	y

Model Class

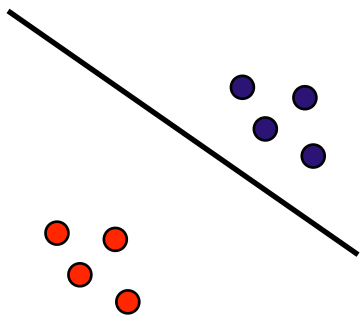
Consider classifiers of the form $y = f(x; w)$

Learning

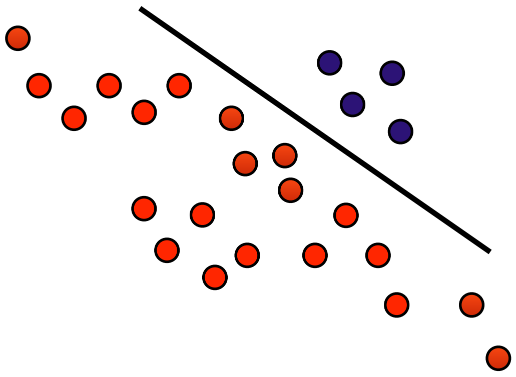
Find w that works well on the training data

Learning Theory

Training. Find a classifier

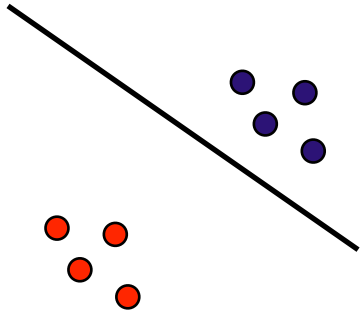


Testing. How well did I do?

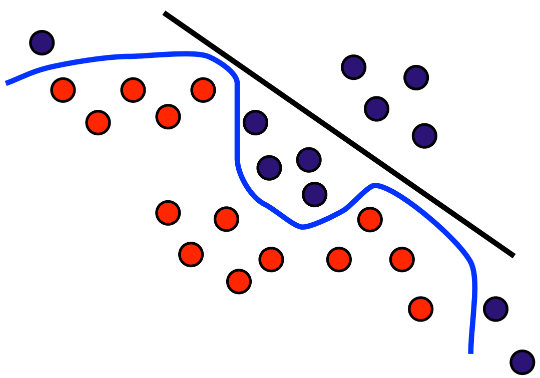


Learning Theory

Training. Find a classifier

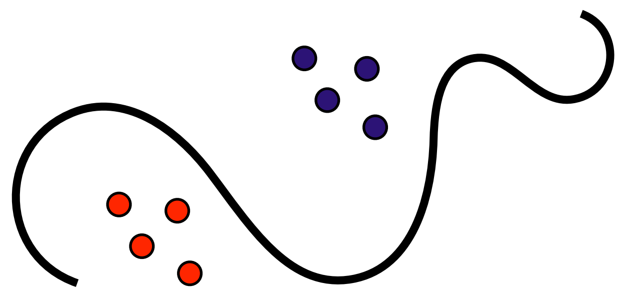


Testing. How well did I do?

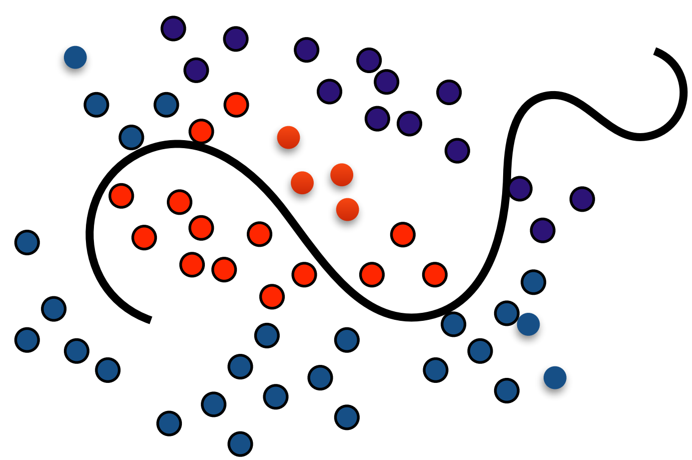


Learning Theory

Training. Find a classifier

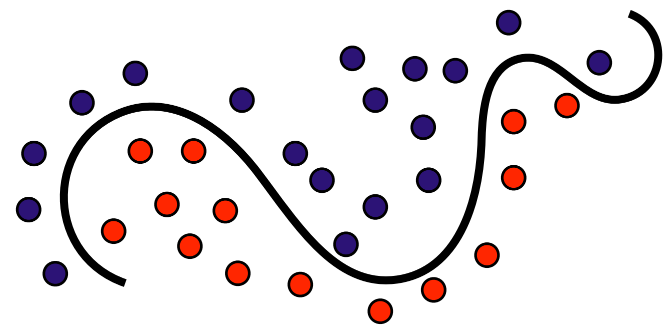


Testing. How well did I do?

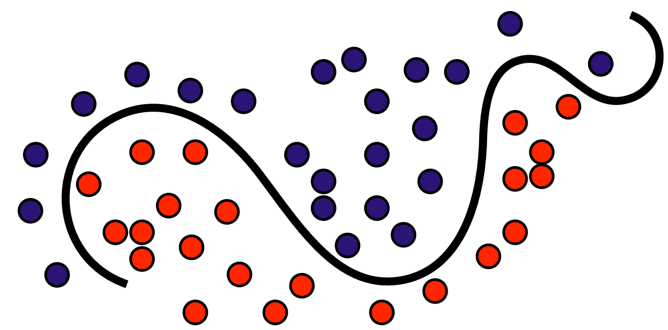


Learning Theory

Training. Find a classifier



Testing. How well did I do?



Learning Theory

- What does training error tell us about generalization error?

- Intuitively:
  - As you get more data, training and generalization error become more _______?
  - As you learn with a richer set of models training and generalization error become more _______?

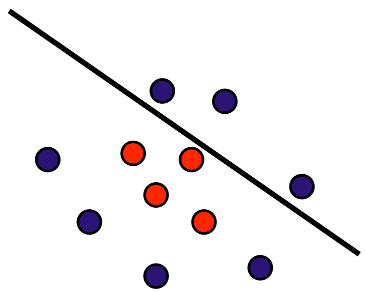
A Generalization Bound

- Suppose you have a set of H possible classifiers
- Your training error is e_{trn}
- Generalization error will be (with probability greater than $1 - \delta$

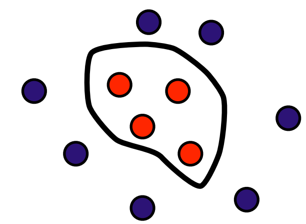
$$e_{tst} \leq e_{trn} + \sqrt{\frac{1}{2n} \log \frac{2H}{\delta}}$$

- What happens with infinite classes?

Learning Theory

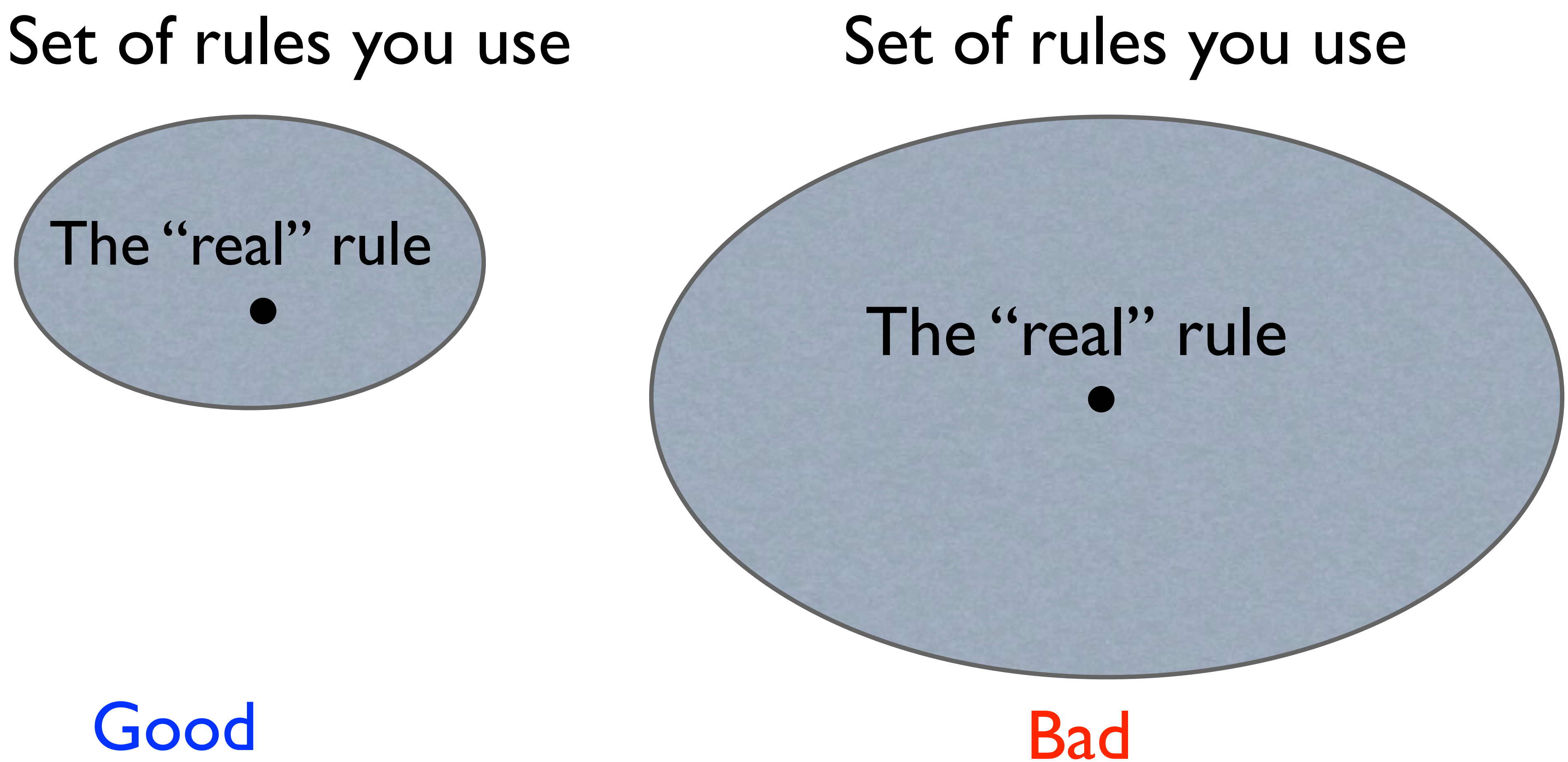


- You will not learn well if:
  - Your classifiers are not a good description of the data
  - Your classifiers are too complex



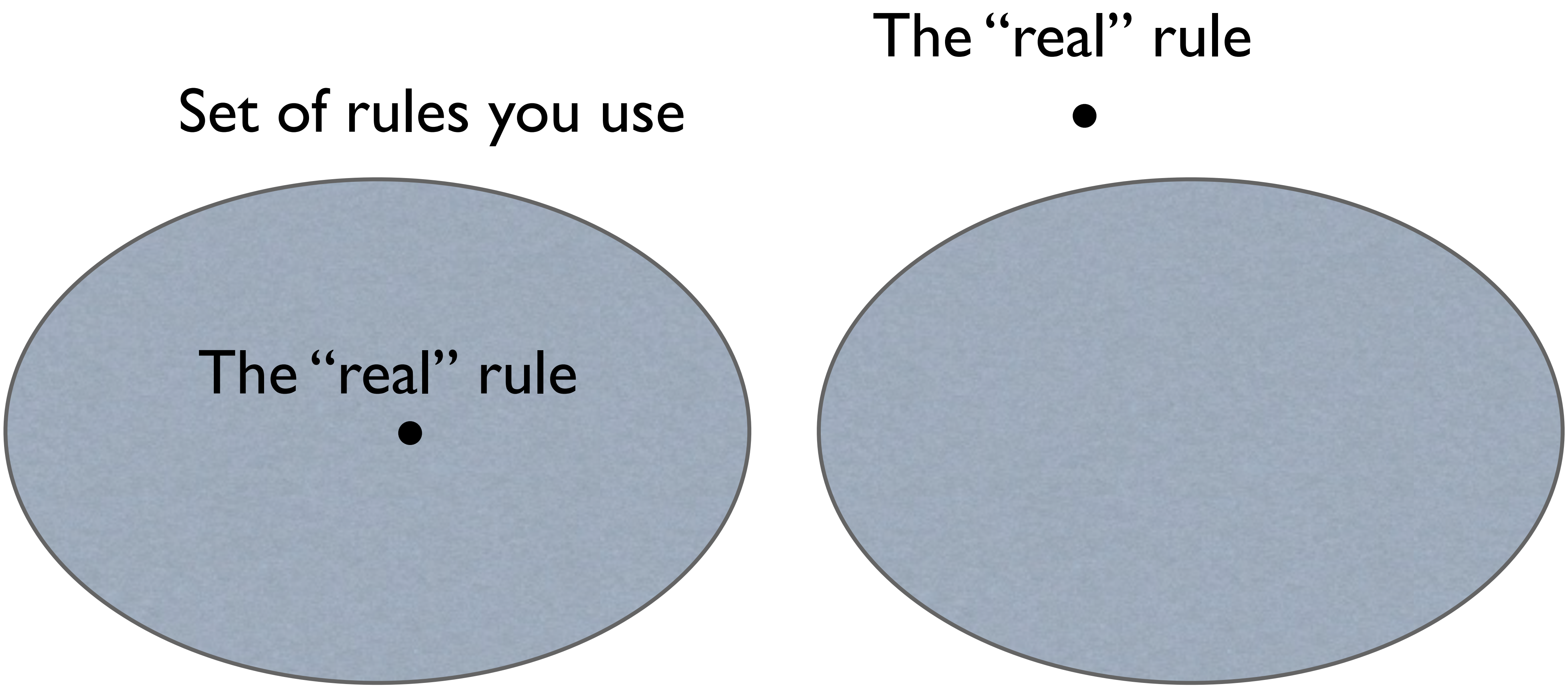
- Good approach: choose the simplest class that contains the rules you will need to learn
- Impossible without prior knowledge

Scenarios



Set of rules you use
The "real" rule
Good
Set of rules you use
The "real" rule
Bad

Scenarios

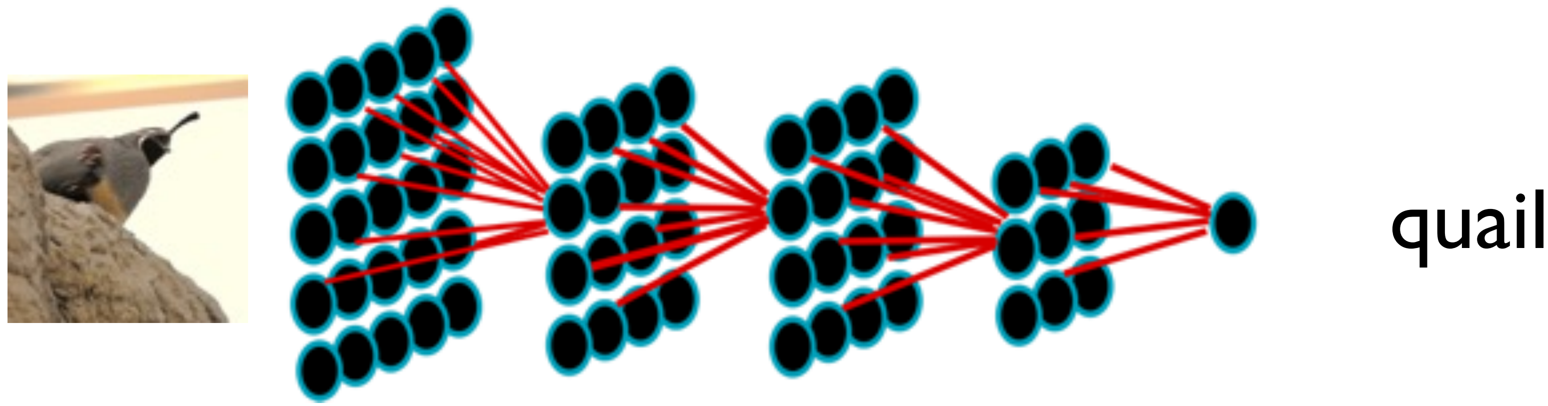


Bad

<u>Terrible</u>

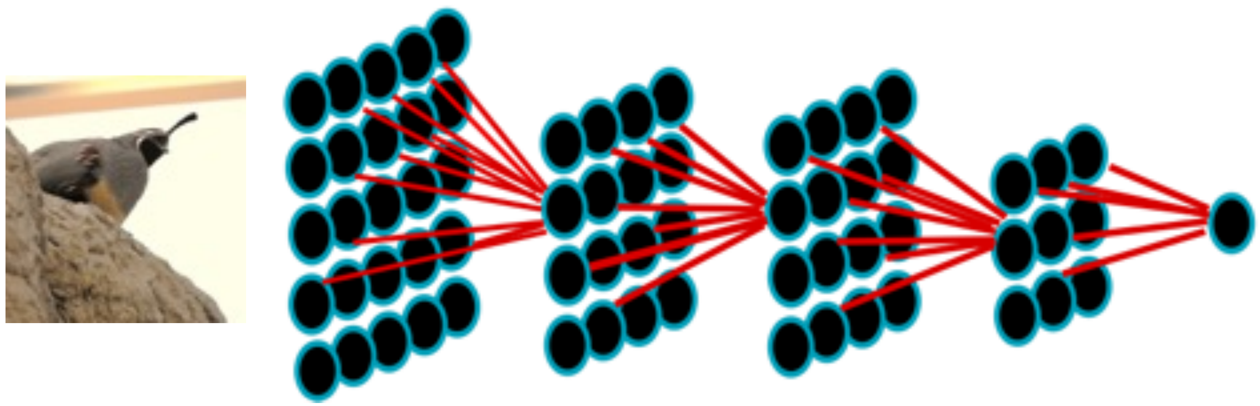
A model for AI

Deep Learning



- Not really neural networks. Don't use spikes. Each layer is a linear function of previous, plus some non-linearity.
- But, can basically represent any function (with enough units). Is that a good thing?

Some History

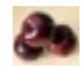
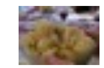


quail

- Had some limitations, and “lost” to support vector machines in the 90s.
- Re emerged in full force since 2000 as “deep learning”. Main reason: amazing performance in image recognition.

The Recipe

Labeled Data

 quail
 apple
 apple
 corn
 corn

Features

1.1 -0.5 0 0 0.3 ...	-1 0 1.2 -0.4 0.1 ...	1.1 -0.5 0 0 0.3 ...	-1 0 1.2 -0.4 0.1 ...	x
quail	apple	apple	corn	y

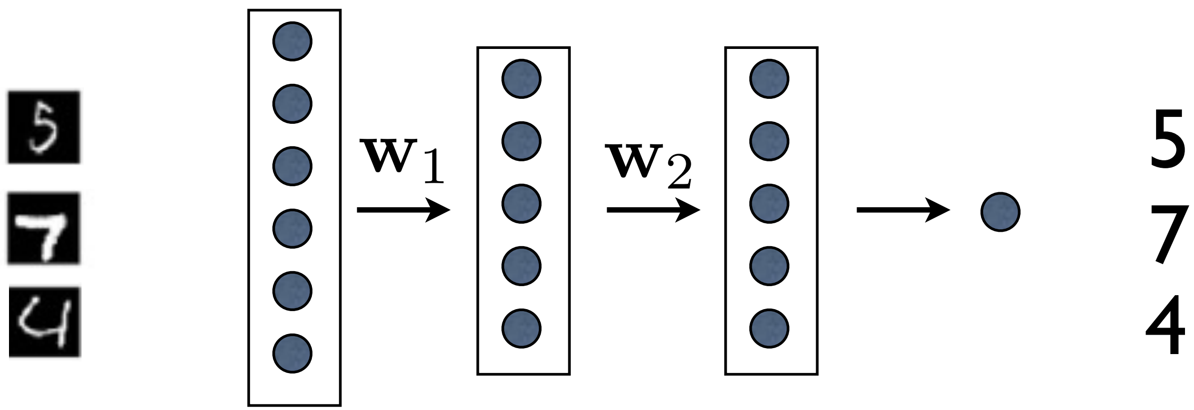
Model Class

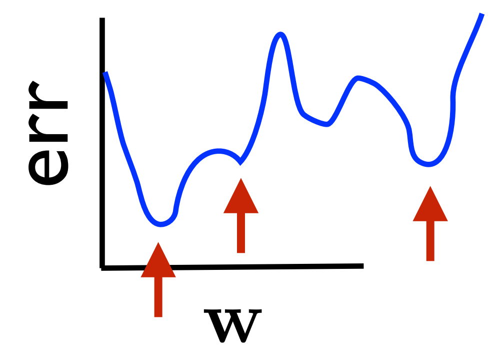
Consider classifiers of the form $y = f(x; w)$

Learning

Find w that works well on the training data

Learning and optimization

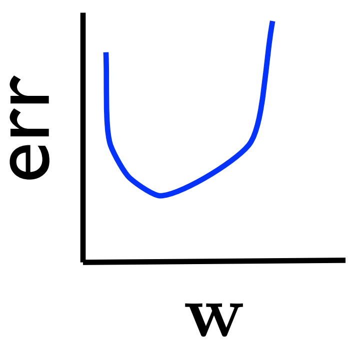


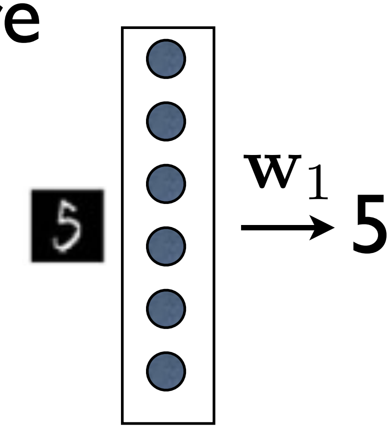


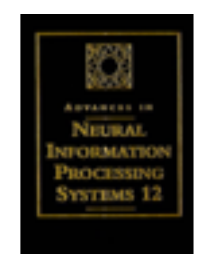
- Find weights that minimize error
- Objective has multiple local minima
- Computationally impossible to find the global optimum (NP hard)
- Serious problem in practice

The (temporary) decline of Neural Networks

- Optimization was a key difficulty with multi-layer architecture
- Support vector machines (early 90s) are one layer architecture that:
  - Can be globally optimized efficiently
  - Works well in practice
- Led to decrease in interest in NN







Why did deep learning win?

- Huge labeled datasets became available
- Focus on certain simple training algorithms
- The some magic happens, which we do not yet understand!
- For images, the deep learning architecture is closely inspired by visual cortex, and this is where the main win is.

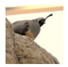
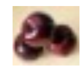
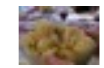
Unsupervised Learning

- Supervised data needs labels.
- But life is more like this:

- Many images. Very few labels.

Unsupervised Learning

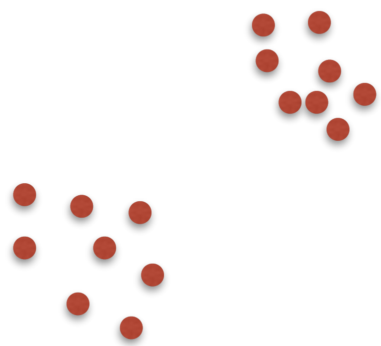
Unlabeled Data

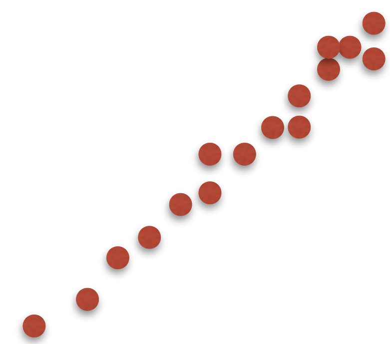
- What's it good for?
  - Learning to generate images/text/music etc
  - Learning useful features
  - If we have some labeled data, we can use them jointly (semi supervised learning)

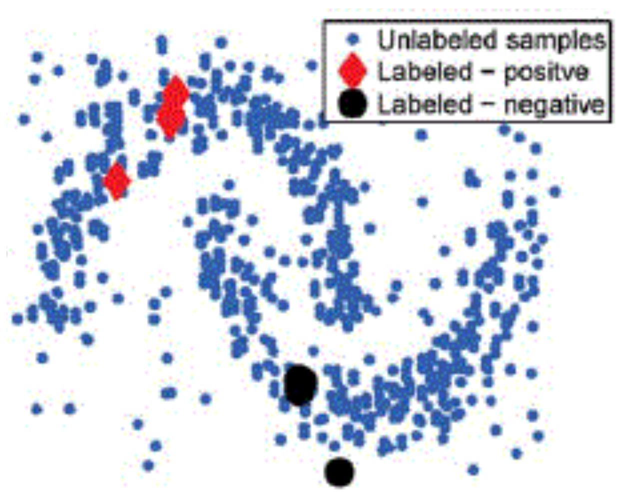
Unsupervised Learning

Clustering



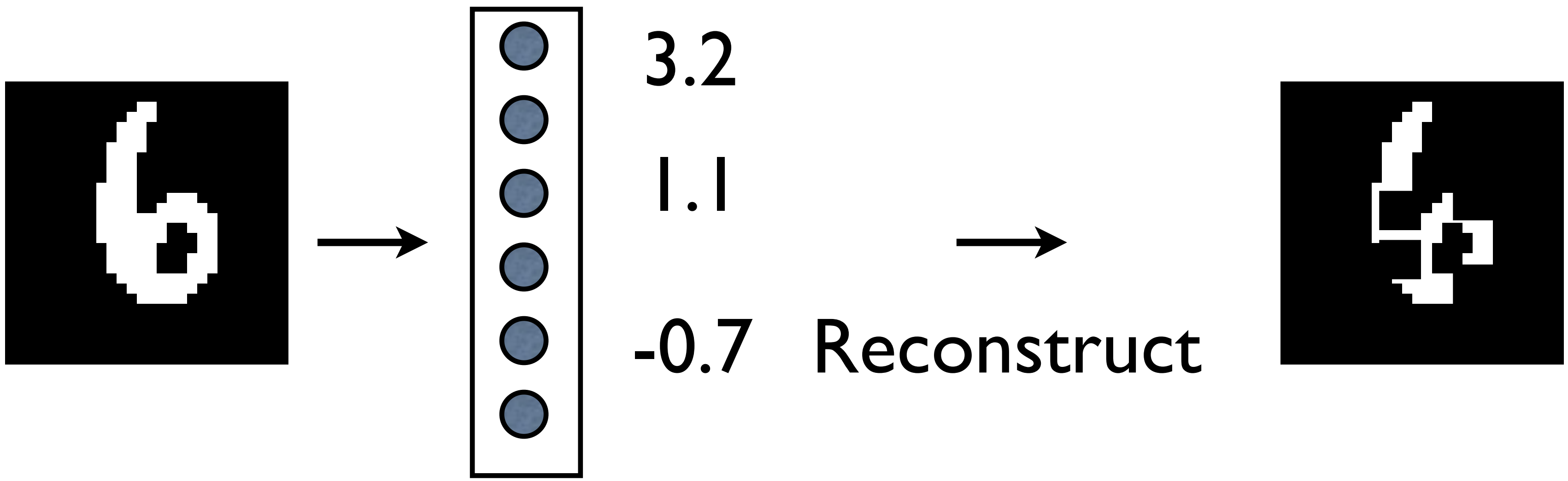
PCA





Graph based SSL

Autoencoders



Goal: Learn features that yield good reconstruction

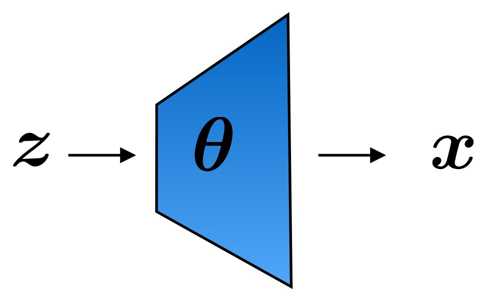
Needs to be a “small” set of features. Why?

If function is linear, you get PCA!

Generative Adversarial Networks

- An approach proposed by Goodfelllow and colleagues.

- Consider a generative model:



$$G(\boldsymbol{z};\boldsymbol{\theta})$$

- Here mapping from z to x is deterministic. Only source of stochasticity is z.

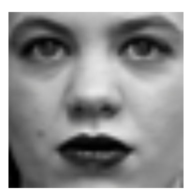
GANs Approach

- The model generates x distributed as $p(\boldsymbol{x};\boldsymbol{\theta})$
- We observe data distributed as $p_D(\boldsymbol{x})$
- In principle, we want to tune $\boldsymbol{\theta}$ such that data and model distribution are close.
- ML is one way of doing this, but is hard to estimate.
- GAN is another.

The GAN Game

- Need some function to measure similarity between data and model distribution.

- Key idea: need to identify cases where the model generates “unreal” points.



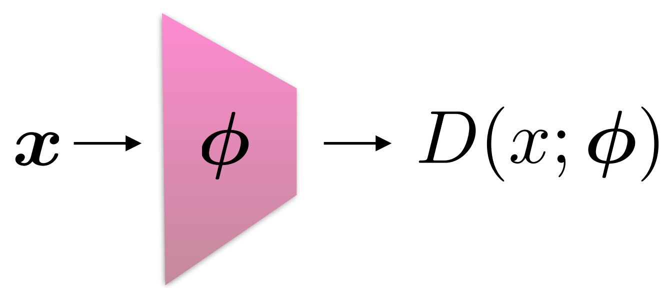
- If you can tell which is real and which is “fake” then model is not perfect.

The GAN Game

- Need some function to measure similarity between data and model distribution.
- The GAN idea: let a “discriminator” function try to identify inputs **x** as **real** or **model**.
- If discriminator fails, we have a good model!
- Formally, discriminator is a function D(x) from x to [0,1].
- D(x) models the probability that x is **real**

The Discriminator

- In practice, the discriminator will be some function from X to [0,1], parameterized by ϕ

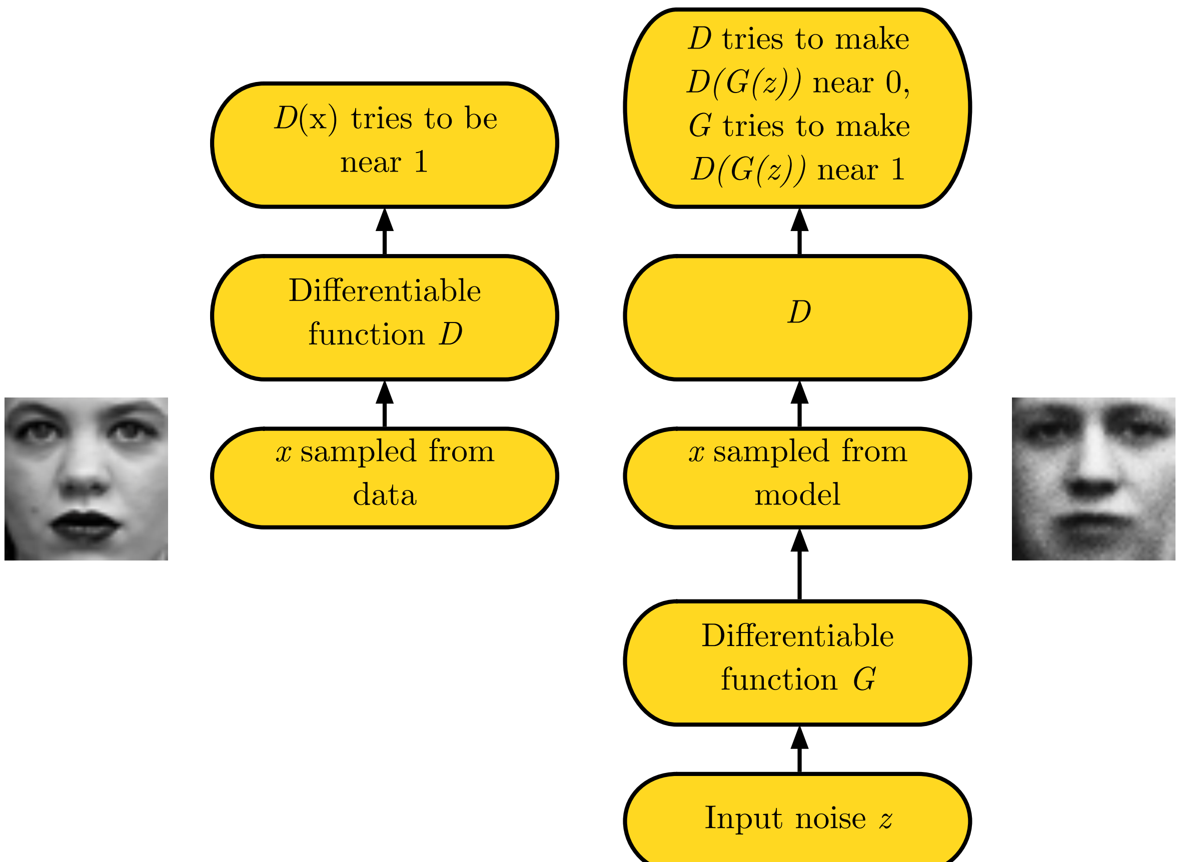


- Usually some multilayer network.

- Interestingly, this is a limited classifier. Cannot discriminate with arbitrary power.

The GAN Game

- From Goodfellow tutorial.



The GAN Optimization Problem

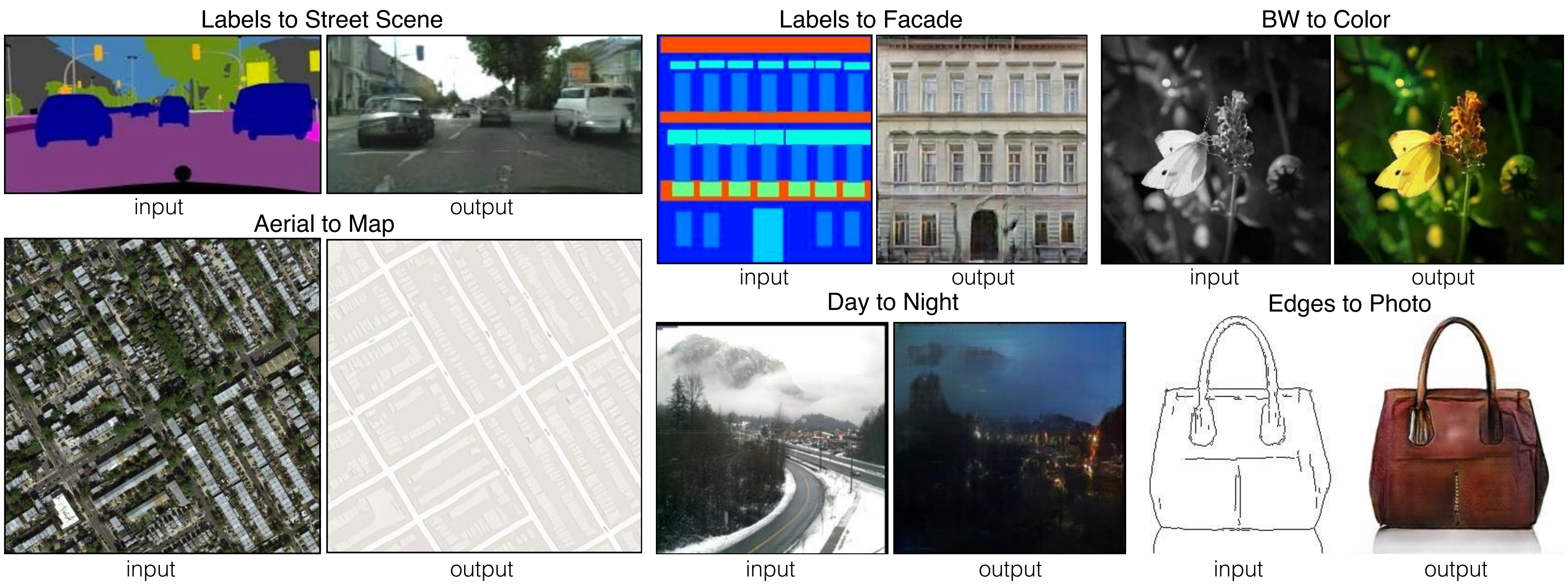
- The generator wants to find parameters that confuse the discriminator as much as possible.
- Denote the "accuracy" of the discriminator by V(D,G)
- Namely, if data is generated by G and we use discriminator D, what is the error.
- Then we want to solve: $\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} V(D(\boldsymbol{\phi}), G(\boldsymbol{\theta}))$
- Just need to specify V.

Generating Faces

- These people do not exist!

Some Examples

- From Isola et al., conditional GANs



Text to Image Generation

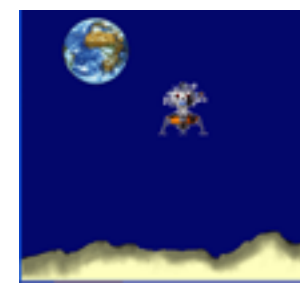
- From Zhang et al., StackGAN

Reinforcement Learning

- We often want to learn how to act in an environment:
  - Self driving cars
  - Playing games
  - Dialogue
- Our actions will affect the world
- What is the best policy?
- How do we learn?

Reinforcement Learning

- One of the hottest topic of research and applications currently.
- Led to:
  - Winning Go.
  - Playing video games as well as humans.
  - Self driving cars.
- Dialogue is far behind…



Teaching ML

- Many toolboxes for learning:
  - TensorFlow (Google)
  - MXNet (Amazon)
  - CNTK (Microsoft)
- Let you train and use models.
- Many nice demos