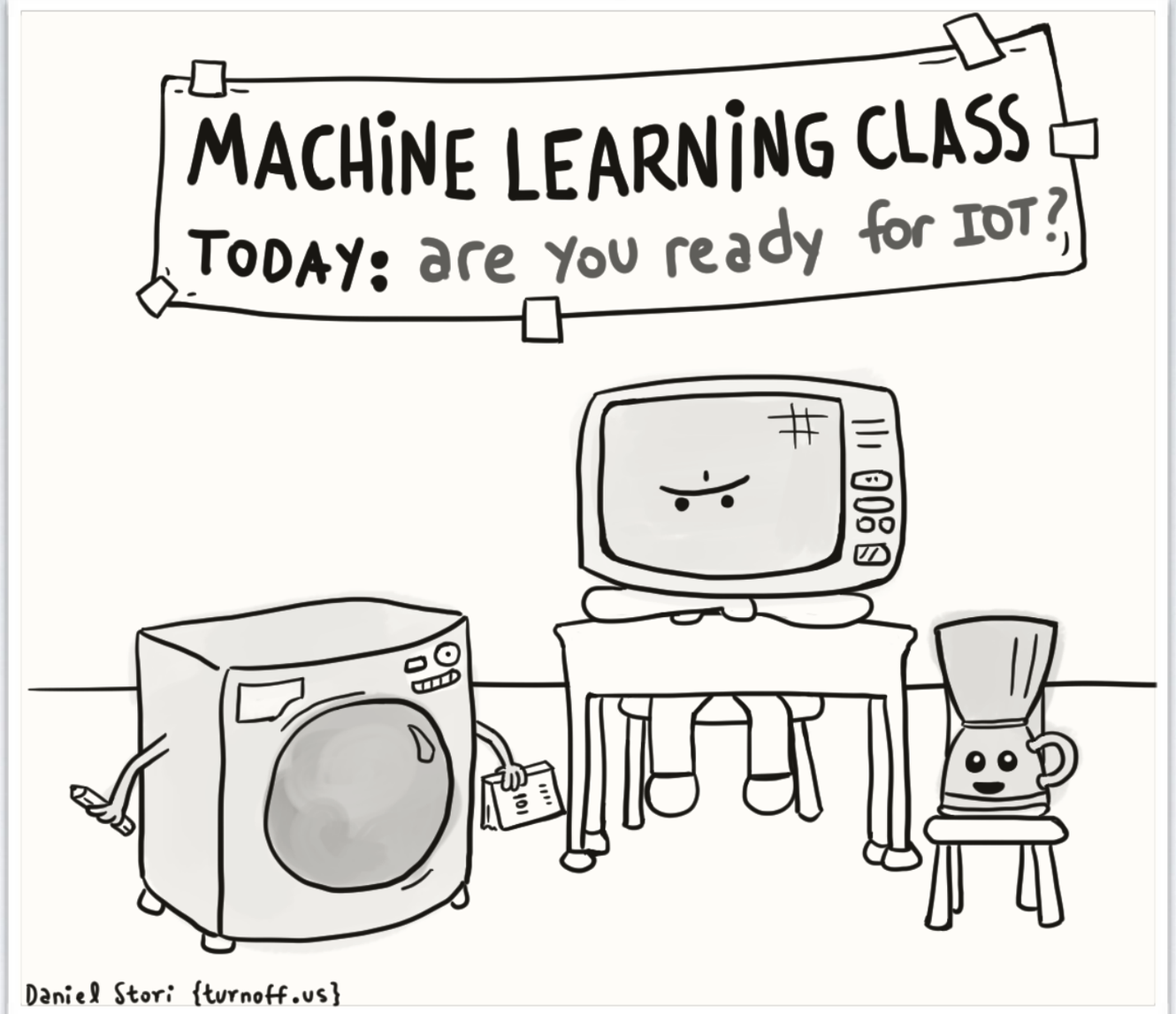


COMPUTATIONAL SELF-HELP (MACHINE LEARNING)

Artificial Intelligence and Global Risks

Leilani H Gilpin (MIT) and Matias Aranguiz (SJTU)

MOTIVATION

- Some computational learning questions
  - What can be learned efficiently?
  - What is inherently hard to learn?
  - A general model of learning?
- Complexity
  - Computational complexity : time and space
  - Sample complexity : amount of training data needed to learn successfully
  - Mistake bounds: number of mistake before learning successfully

from Mehryar Mohri

GOALS

- Understand how machine learning works (and what it cannot do)
    - Problems of Learning (little statistics, only when necessary)
        - Neural Networks
        - Reinforcement learning
        - Anomaly Detection
- Understand real problems
    - Adversarial examples

BEWARE : WIZARD OF OZ

BLOOM TAXONOMY

1. Remembering/Memorizing defined as the knowing of previously learned material or retrieving, recognizing, and recalling relevant knowledge.

2. Understanding defined as being able to comprehend facts by comparing and interpreting main ideas within the learned material.

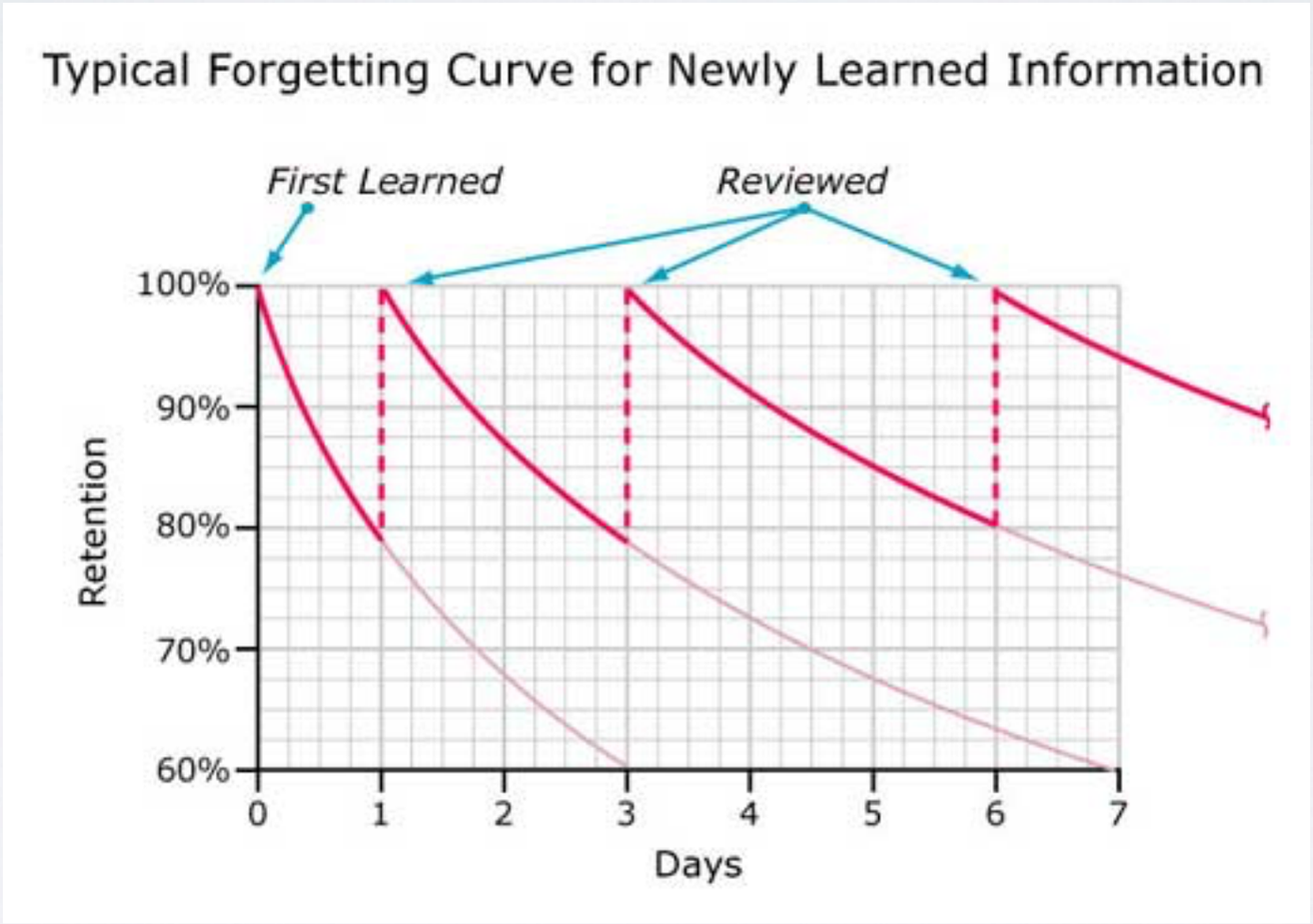
3. Applying defined as the ability to use learned material in a new or unprompted way of abstraction and to solve a newly defined problem.

4. Analyzing defined as the ability to examine a problem area and identify the various components (breaking the problem down).

5. Evaluating defined as the ability to make judgments based on criteria or standards or to combine parts to form a new concept or idea.

6. Creating defined as the ability to integrate learning from different areas into a plan solving a problem and to propose alternative solutions.

HUMANS FORGET, MACHINES DON'T



HOW WE LEARN

Task	Human	Computer
Riding a bike	Just do it	?
Speak a language	Practice, talk with native speakers	word embeddings
Memorize	Repetition	Save and lookup
Recognize objects in Images	sometimes black box sometimes trial and error	neural network

MACHINE LEARNING (I)

WHAT IS MACHINE LEARNING

- Broadly defined as, "computational methods using experience to improve performance or make predictions." - Mohri, Rostamizadeh and Talwalkar

- Experience is past information available to the learner (data)

- Quality and size of data is crucial

- Machine learning consists of designing efficient and accurate prediction algorithms

MACHINE LEARNING REVOLUTION

- Multiple features at a time with different weights
  - Don't have trial and error
- Possible because of data, computational power and fast querying
- "Automatically" process complex datasets - get "structure"
- Lack of labels

APPLICATIONS AND PROBLEMS

- Text or document classification, e.g., spam detection;
- Natural language processing, e.g., morphological analysis, part-of-speech tagging, statistical parsing, named-entity recognition;
- Speech recognition, speech synthesis, speaker verification;
- Optical character recognition (OCR);
- Medical diagnosis;
- **Headline text selection**
- Computational biology applications, e.g., protein function or structured prediction;
- Computer vision tasks, e.g., image recognition, face detection;
- Fraud detection (credit card, telephone) and network intrusion;
- Games, e.g., chess, backgammon;
- Unassisted vehicle control (robots, navigation);
- Recommendation systems, search engines, information extraction systems.

CLASSES OF PROBLEMS

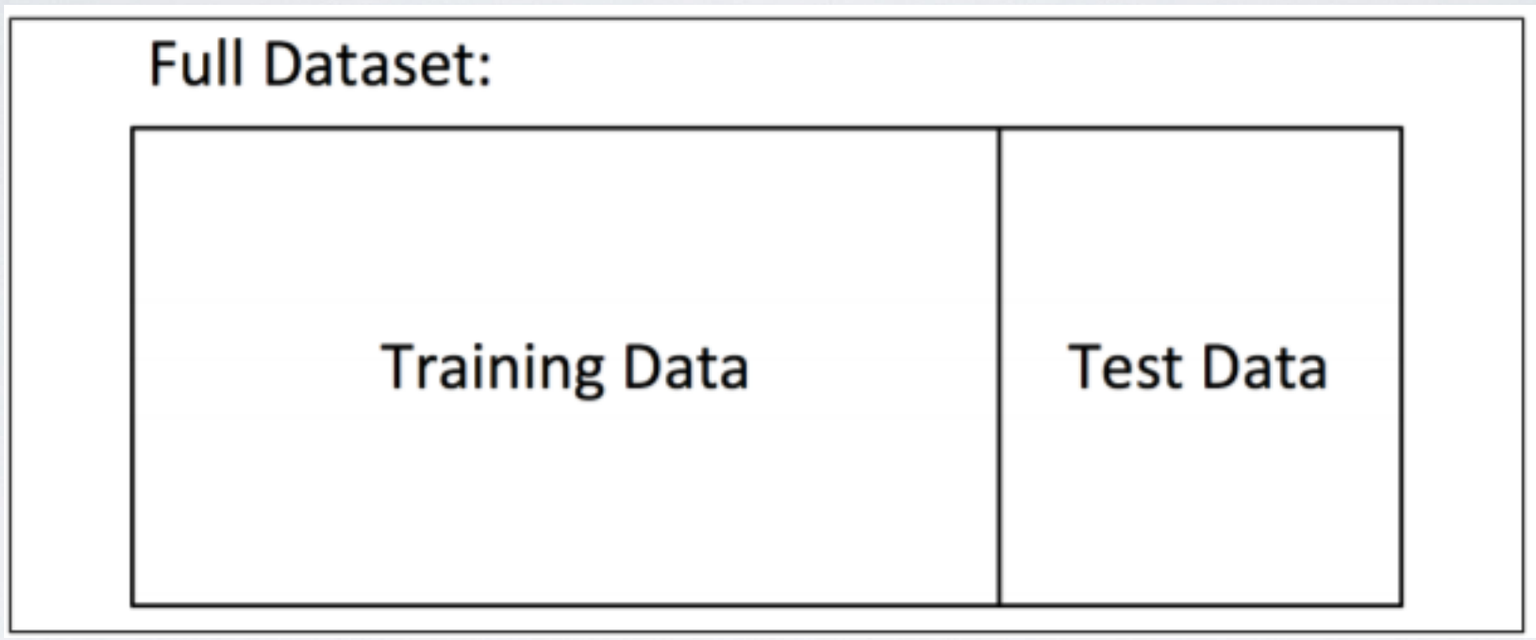
- Classification
- Regression
- Ranking
- Clustering
- Dimensionality
- Object - Generating accurate predictions for unseen items and of designing efficient and robust algorithms to produce these predictions, even for large-scale problems.

KEY TERMS

- Examples- Items or instances of data used for learning or evaluation.
- Features -The set of attributes, often represented as a vector, associated to an example.
- Labels- Values or categories assigned to examples.
- Training sample- Examples used to train a learning algorithm.
- Validation sample- Examples used to tune the parameters of a learning algorithm when working with labeled data.
- Test sample- Examples used to evaluate the performance of a learning algorithm.
- Loss function - A function that measures the difference (loss) between a predicted label and a true label.

$$L(y, y') = 1_{y' \neq y} \qquad L(y, y') = (y' - y)^2$$

SPLITTING



"fair evaluation"

SO WHAT DO WE NEED

- Data size - small, medium, large
- Hypothesis about the data
- FEATURES
- Test set “similar to” training set

WHAT HAS GONE WRONG

- Learn the wrong thing
- Garbage in, garbage out
- More data is not necessarily right
    - Bag of words is king

NEURAL NETS (IN THEORY)

WHAT'S THE BIG DEAL WITH DEEP LEARNING?

- Performing arithmetic really fast

- Explicitly following a set of instructions

- Similarities with the way the brain works

WHAT'S THE BIG DEAL WITH DEEP LEARNING?

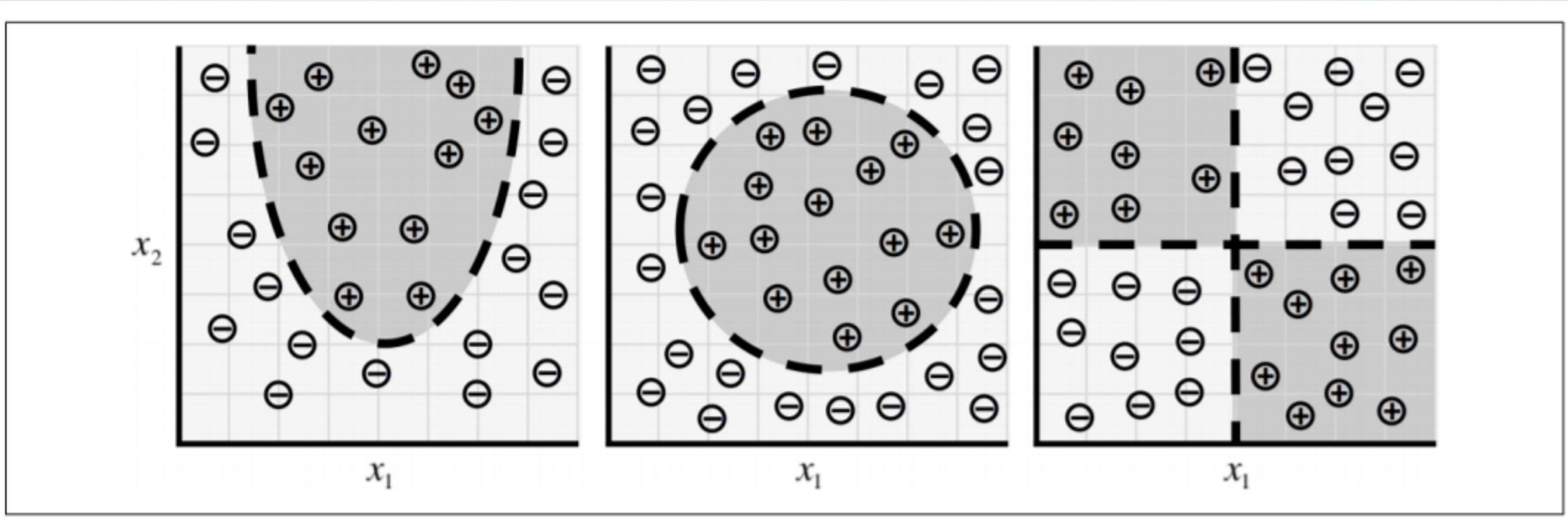
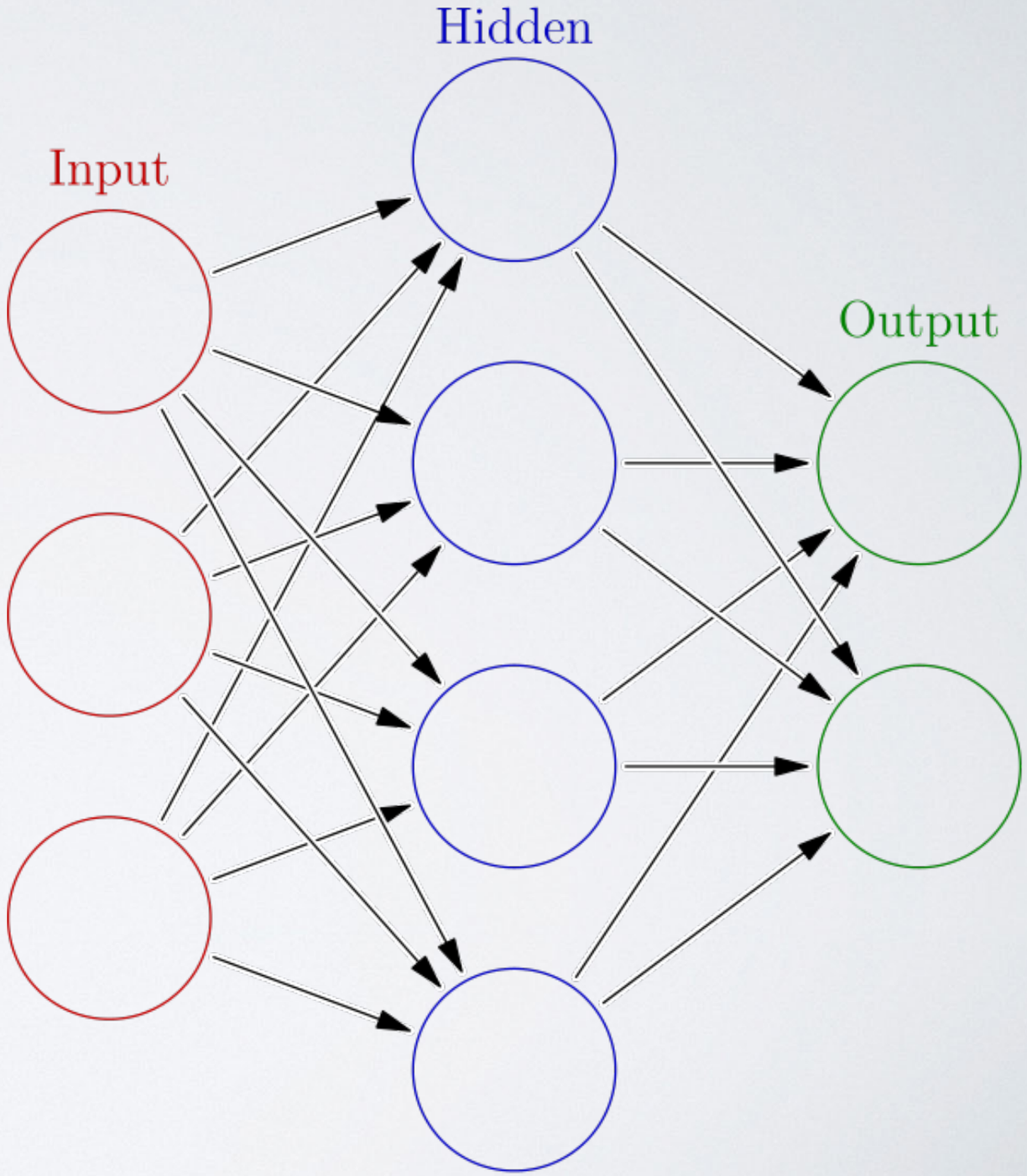


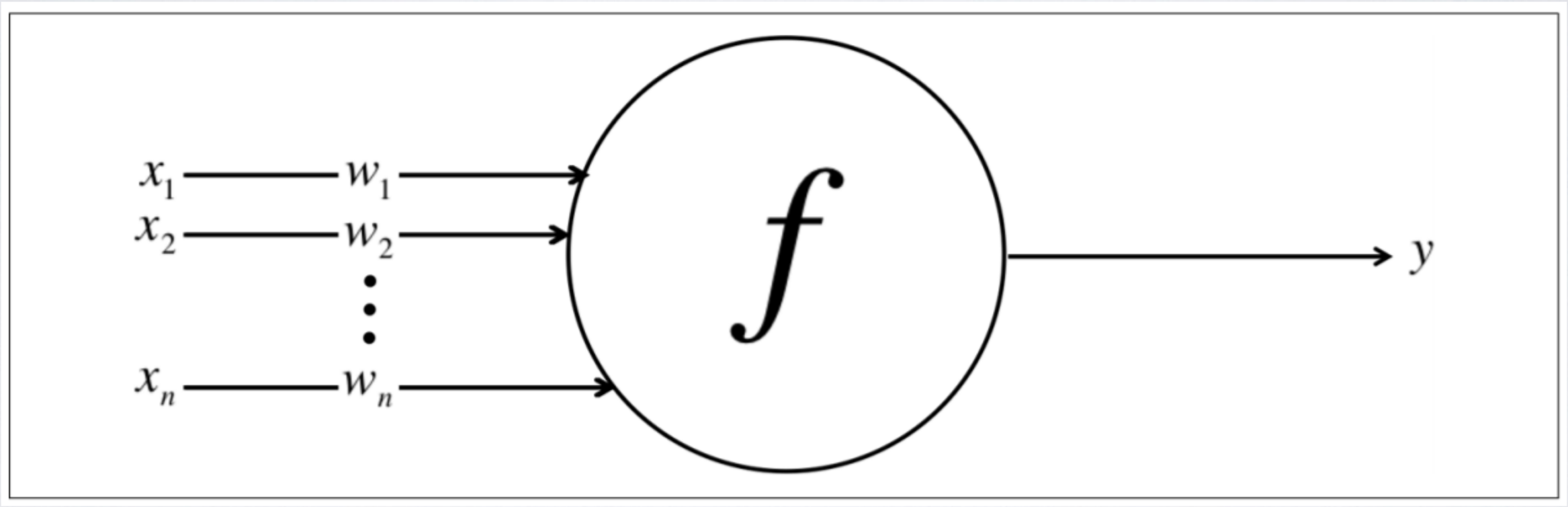
Figure 1-5. As our data takes on more complex forms, we need more complex models to describe them

KEY TERMS

- Neuron - has an input, weight, function, produces an output
- Connections - between neurons



WHAT'S THE BIG DEAL WITH DEEP LEARNING? AND HOW?



WHAT'S THE BIG DEAL WITH DEEP LEARNING? AND HOW?

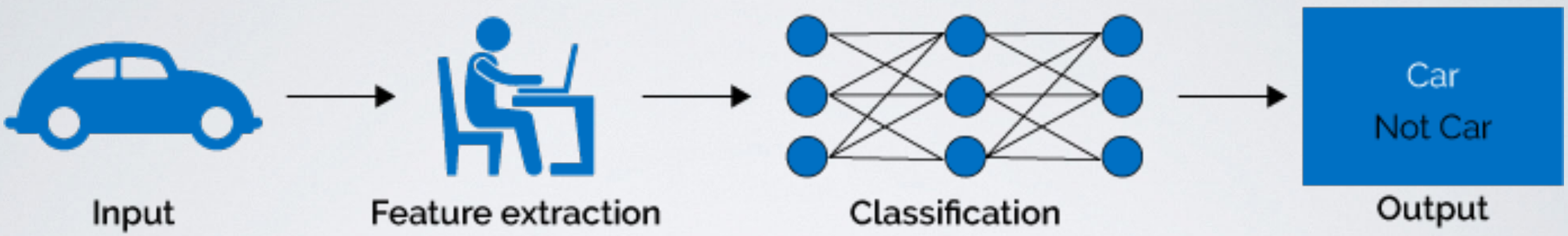
1. The layers of neurons that lie sandwiched between the first layer of neurons (input layer) and the last layer of neurons (output layer) are called the hidden layers. This is where most of the magic is happening

2. More often than not, hidden layers have fewer neurons than the input layer to force the network to learn compressed representations of the original input.

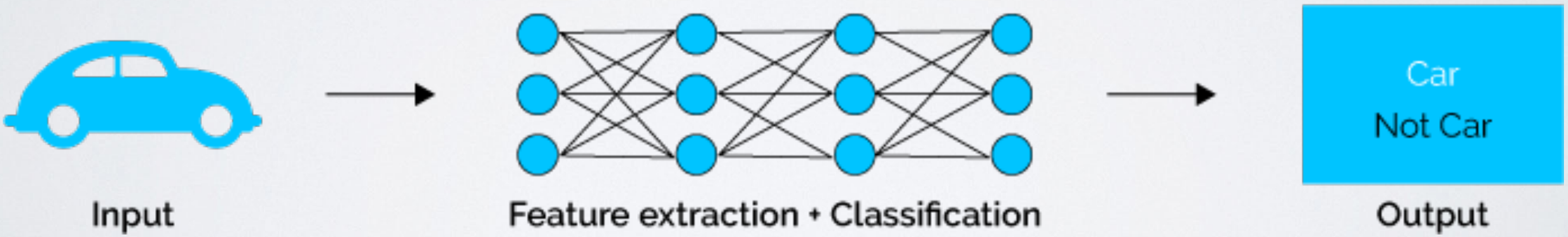
3. It is not required that every neuron has its output connected to the inputs of all neurons in the next layer.

4. The inputs and outputs are vectorized representations. For example, you might imagine a neural network where the inputs are the individual pixel RGB values in an image represented as a vector

Machine Learning



Deep Learning



THIS ISN'T NEW

- *Perceptrons* - by Minsky and Paper in 1968

- A perceptron is a kind of artificial neural network

THIS ALSO ISN'T NEW

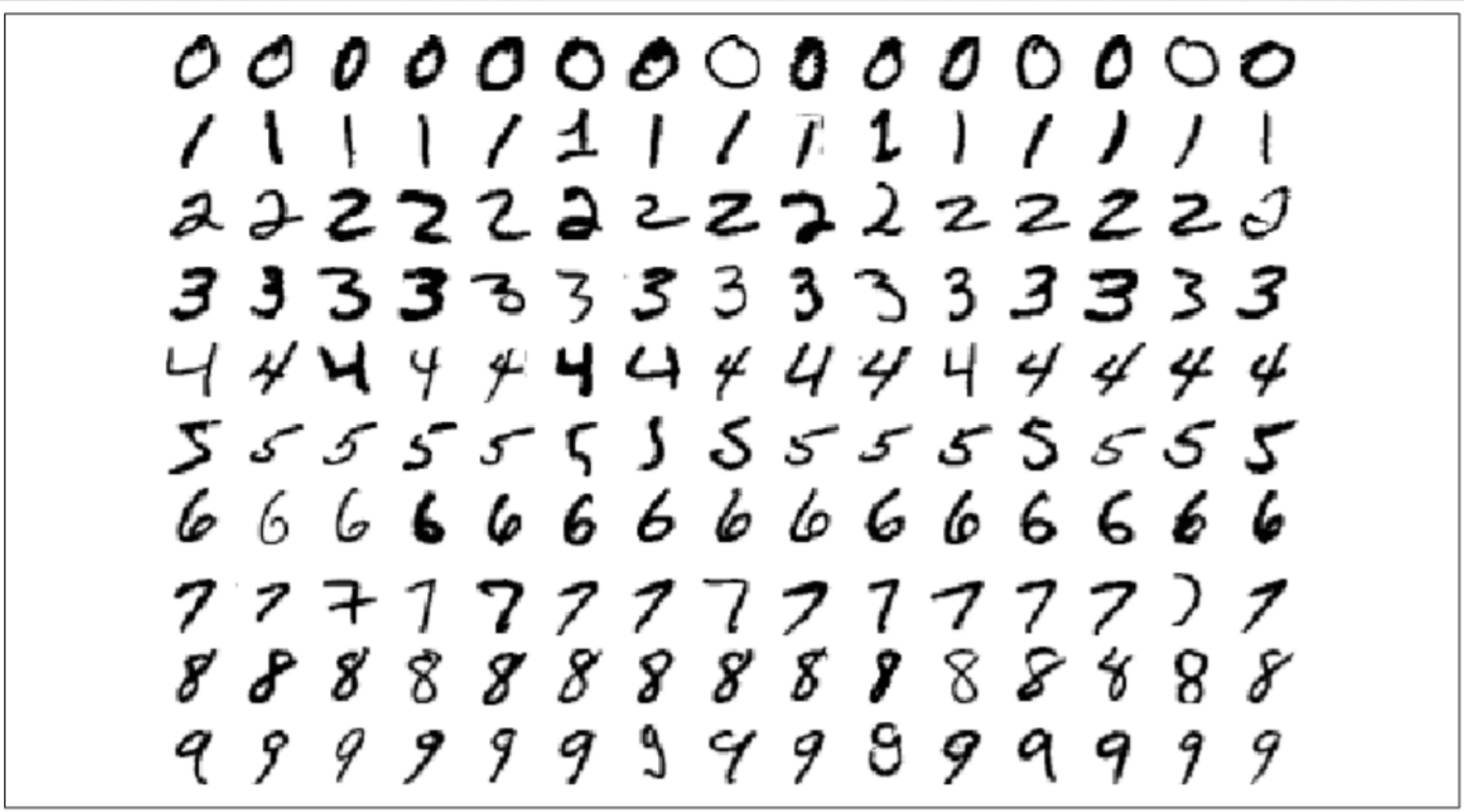


Figure 1-1. Image from MNIST handwritten digit dataset[2]

WHAT SHOULD WE CARE ABOUT

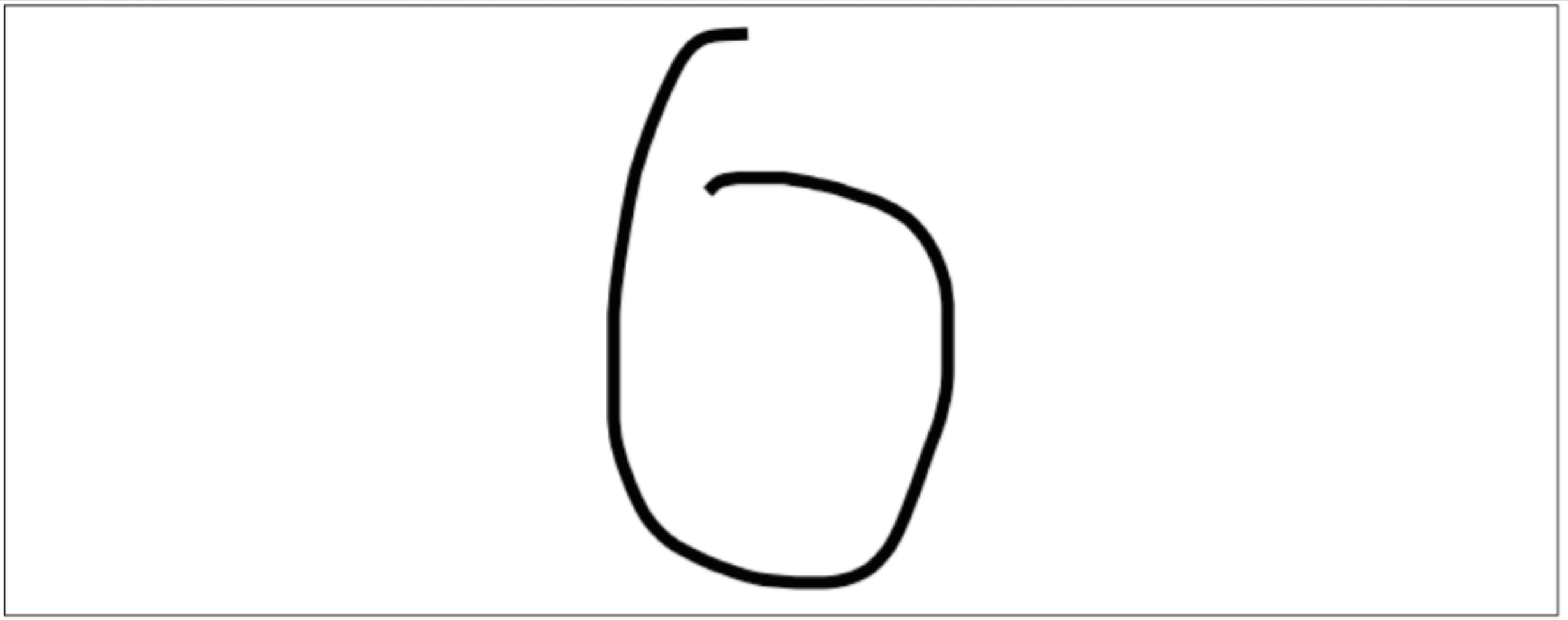


Figure 1-2. A zero that's algorithmically difficult to distinguish from a six

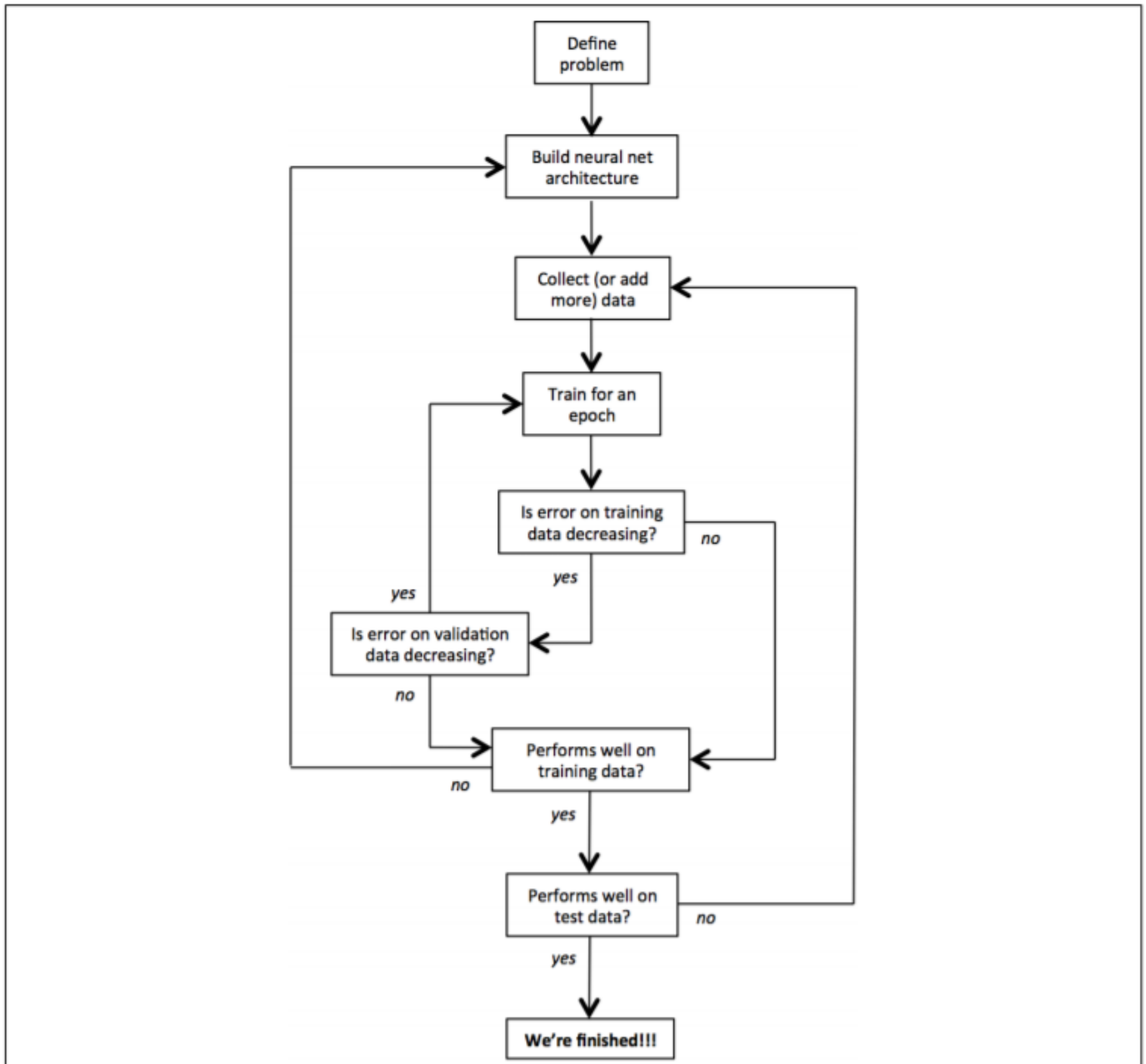
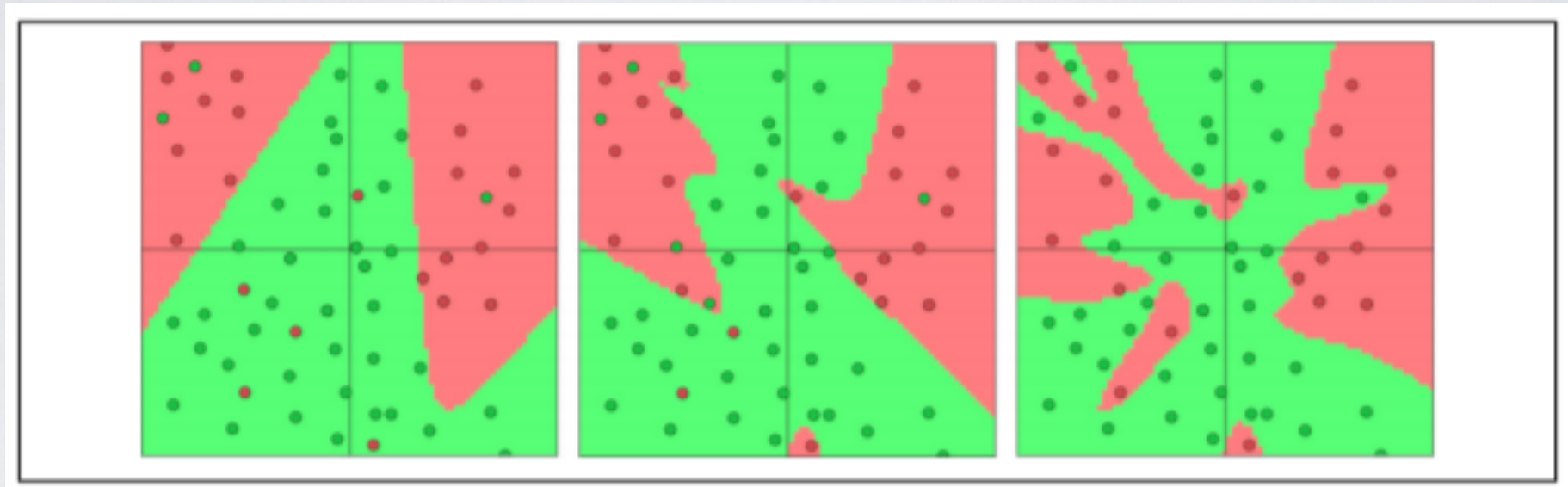
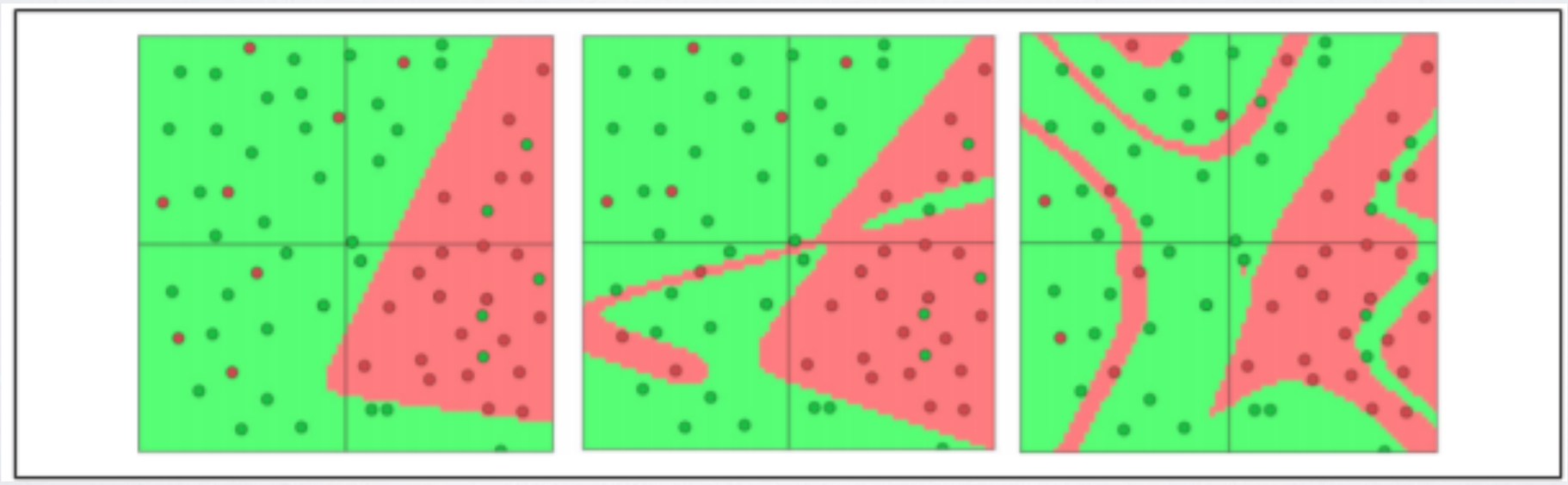


Figure 2-14. Detailed workflow for training and evaluating a deep learning model

INTUITION?



1 hidden layer with 3,6, 20 neurons



network with 1,2, or 4 hidden layers

Table 1: **Glossary**

Neural network: A network of simple neuron-like processing units that collectively perform complex computations. Neural networks are often organized into layers, including an input layer that presents the data (e.g, an image), hidden layers that transform the data into intermediate representations, and an output layer that produces a response (e.g., a label or an action). Recurrent connections are also popular when processing sequential data.

Deep learning: A neural network with at least one hidden layer (some networks have dozens). Most state-of-the-art deep networks are trained using the backpropagation algorithm to gradually adjust their connection strengths.

Backpropagation: Gradient descent applied to training a deep neural network. The gradient of the objective function (e.g., classification error or log-likelihood) with respect to the model parameters (e.g., connection weights) is used to make a series of small adjustments to the parameters in a direction that improves the objective function.

Convolutional network (convnet): A neural network that uses trainable filters instead of (or in addition to) fully-connected layers with independent weights. The same filter is applied at many locations across an image (or across a time series), leading to neural networks that are effectively larger but with local connectivity and fewer free parameters.

Model-free and model-based reinforcement learning: Model-free algorithms directly learn a control policy without explicitly building a model of the environment (reward and state transition distributions). Model-based algorithms learn a model of the environment and use it to select actions by planning.

Deep Q-learning: A model-free reinforcement learning algorithm used to train deep neural networks on control tasks such as playing Atari games. A network is trained to approximate the optimal action-value function $Q(s, a)$, which is the expected long-term cumulative reward of taking action a in state s and then optimally selecting future actions.

Generative model: A model that specifies a probability distribution over the data. For instance, in a classification task with examples X and class labels y, a generative model specifies the distribution of data given labels $P(X|y)$, as well as a prior on labels $P(y)$, which can be used for sampling new examples or for classification by using Bayes' rule to compute $P(y|X)$. A discriminative model specifies $P(y|X)$ directly, possibly by using a neural network to predict the label for a given data point, and cannot directly be used to sample new examples or to compute other queries regarding the data. We will generally be concerned with directed generative models (such as Bayesian networks or probabilistic programs) which can be given a causal interpretation, although undirected (non-causal) generative models (such as Boltzmann machines) are also possible.

Program induction: Constructing a program that computes some desired function, where that function is typically specified by training data consisting of example input-output pairs. In the case of probabilistic programs, which specify candidate generative models for data, an abstract description language is used to define a set of allowable programs and learning is a search for the programs likely to have generated the data.

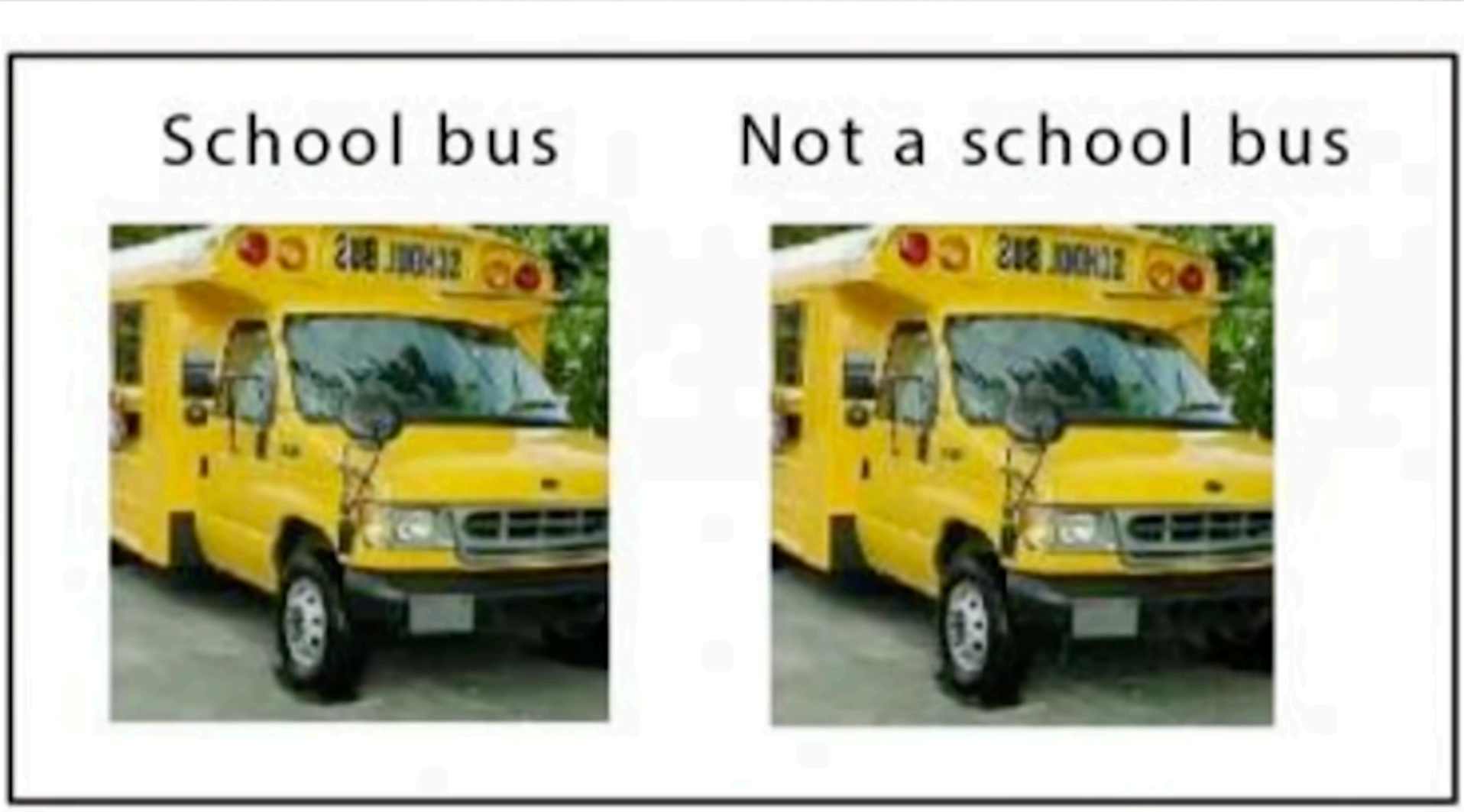
SO WHAT DO WE NEED

- Data size - large
  - ImageNet - 10 million images (1 million with bounding boxes)
- ~~Hypothesis about the data~~
- ~~FEATURES~~
- Test set "similar to" training set

WHAT HAS GONE WRONG

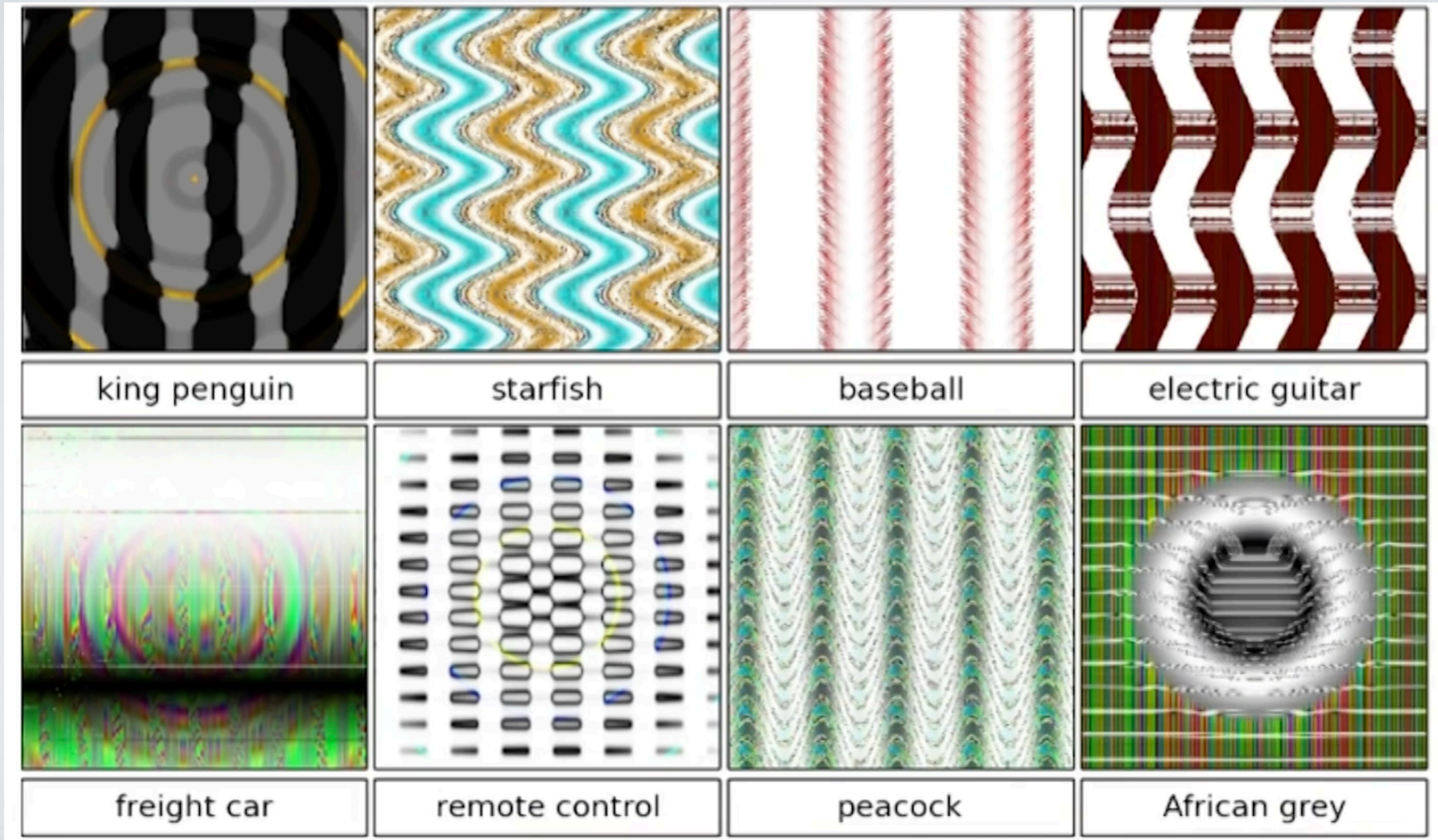
- Learn the wrong thing
- Garbage in, garbage out
- More data is generally right
- Trial and error instead of understanding

ADVERSARIAL EXAMPLES



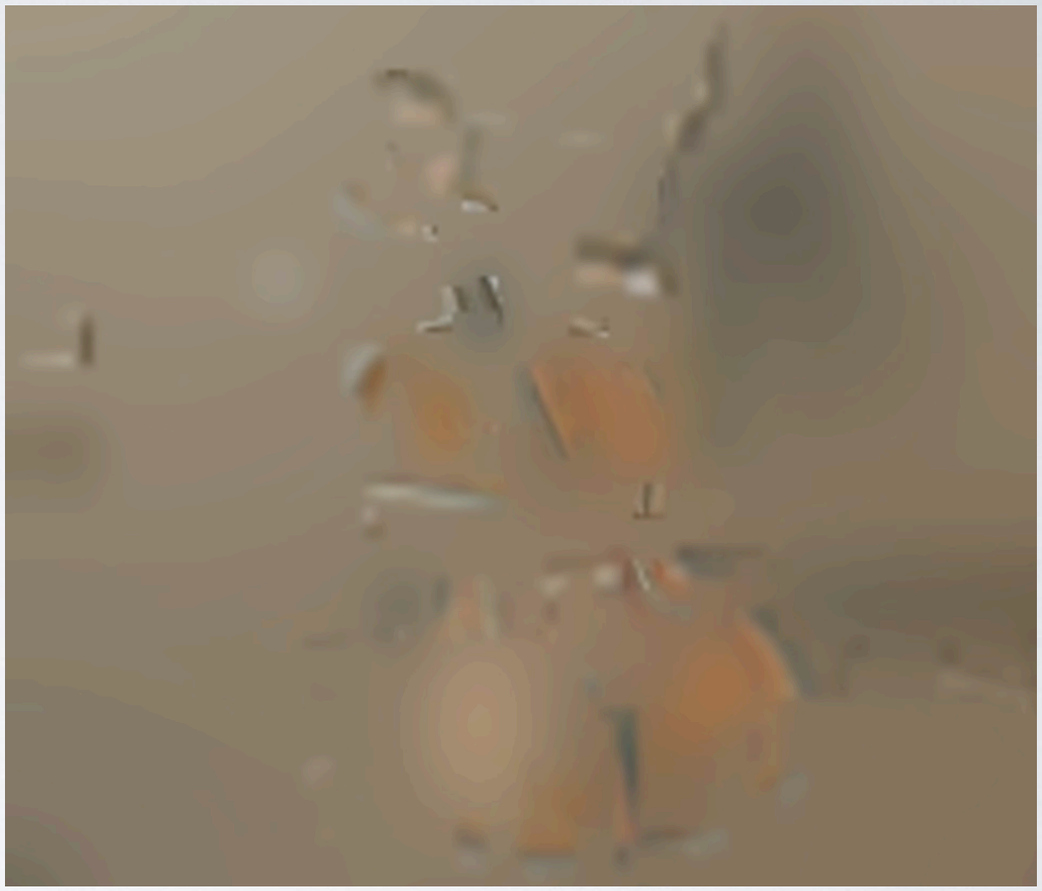
Szegedy et al. 2014

ADVERSARIAL EXAMPLES



ADVERSARIAL EXAMPLES

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean



Theoretical Impediments to Machine Learning
A position paper

Judea Pearl, University of California, Los Angeles
November 2016

ADVERSARIAL EXAMPLES

```
input : “A penguin eats food”

This perception is UNREASONABLE

REASONING:
A penguin is an animal that lives in
Antartica and eats enough to eat.
Food is an animal that lives in the
refrigerator and eats food.

So a penguin cannot reasonably be
located at the same location as food.
```

EXPLAINING REASONABLENESS

input : “The table has a few books on it and the table itself is white”

This perception is REASONABLE

REASONING:
Books are typically found at the same location as a table. So it is reasonable for books to be located on a table.

Furthermore, tables are typically made of materials that can be colored white. So it is reasonable for a table to be white.

work with Cagri Zaman

EXPLAINING REASONABLENESS


```
input : “Mailbox crossing the street”


This perception is unreasonable.

==================================

A mailbox is an object or thing that
cannot move on its own.  So it is
unreasonable for a mailbox to cross
the street.
```

REINFORCEMENT LEARNING

MACHINE LEARNING (II)

BIG IDEA

What if you don't have a supervisor?

BIG IDEA

- There is no supervisor, only a reward signal
- Feedback is delayed
- Time MATTERS
- Agent's actions affect the data it receives

CLASSES OF PROBLEMS

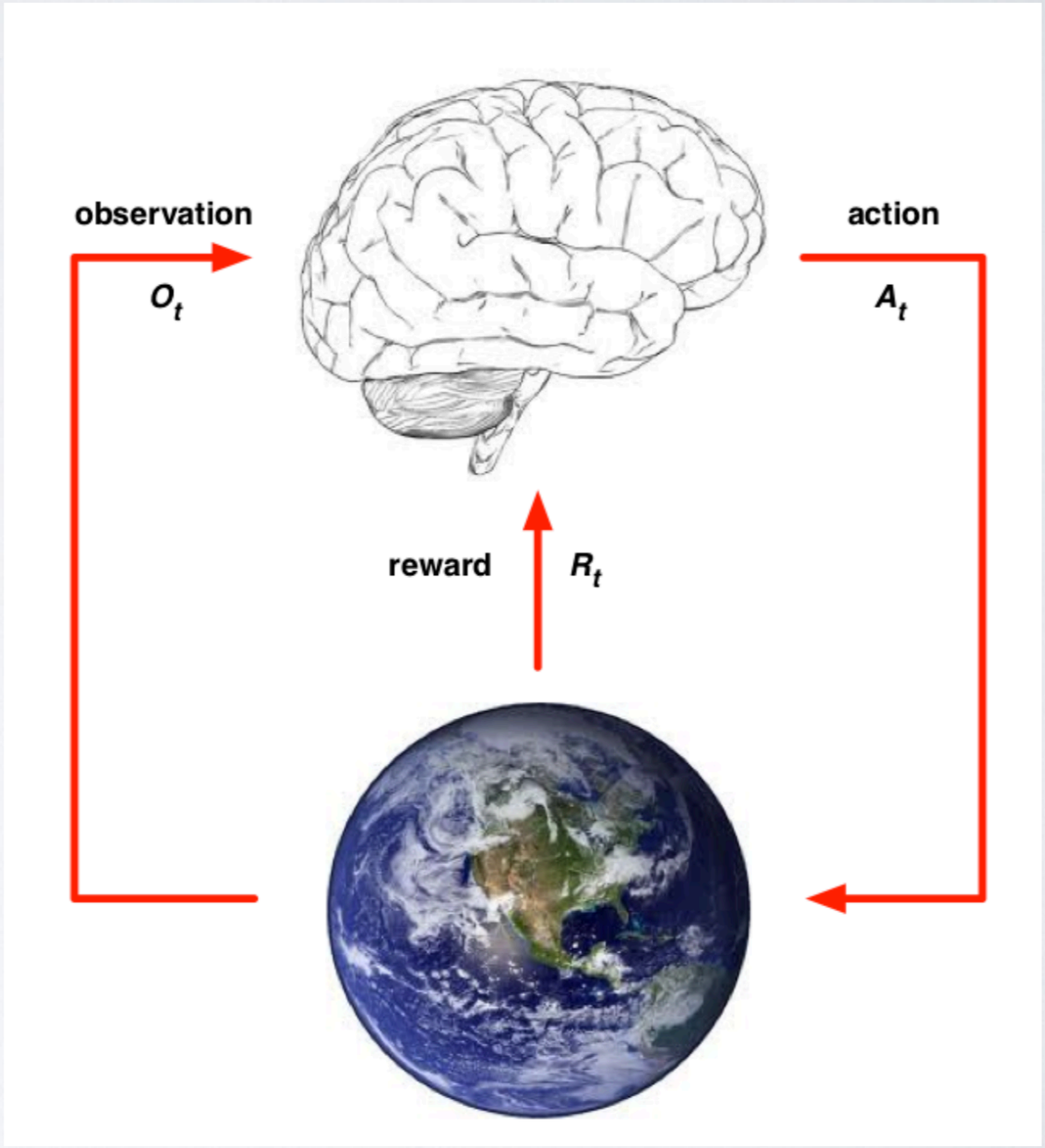
- Playing Backgammon
- Managing investments
- Controlling a power station
- Making a robot walk
- Playing Atari games

KEY TERMS

- Agent (state) - Agent's internal representation
- Policy - The agent's behavior (A map from state to action).
- Actions - What the agent can do (for example, take one step left or right).
- Environment - Environments private representation. (E.g. Whatever data the environment uses to pick the next observation/reward).
- Reward - A scalar feedback signal to indicate how well an agent is doing at step t
- Markov Decision Process - Extension of Markov chains

PROBLEM FORMULATION

- At each time, t, the agent
  - Executes action At
  - Receives observation Ot
  - Receives scalar reward Rt
- The environment
  - Receives action At
  - Emits observation Ot+1
  - Emits scalar reward Rt+1
- Increment t

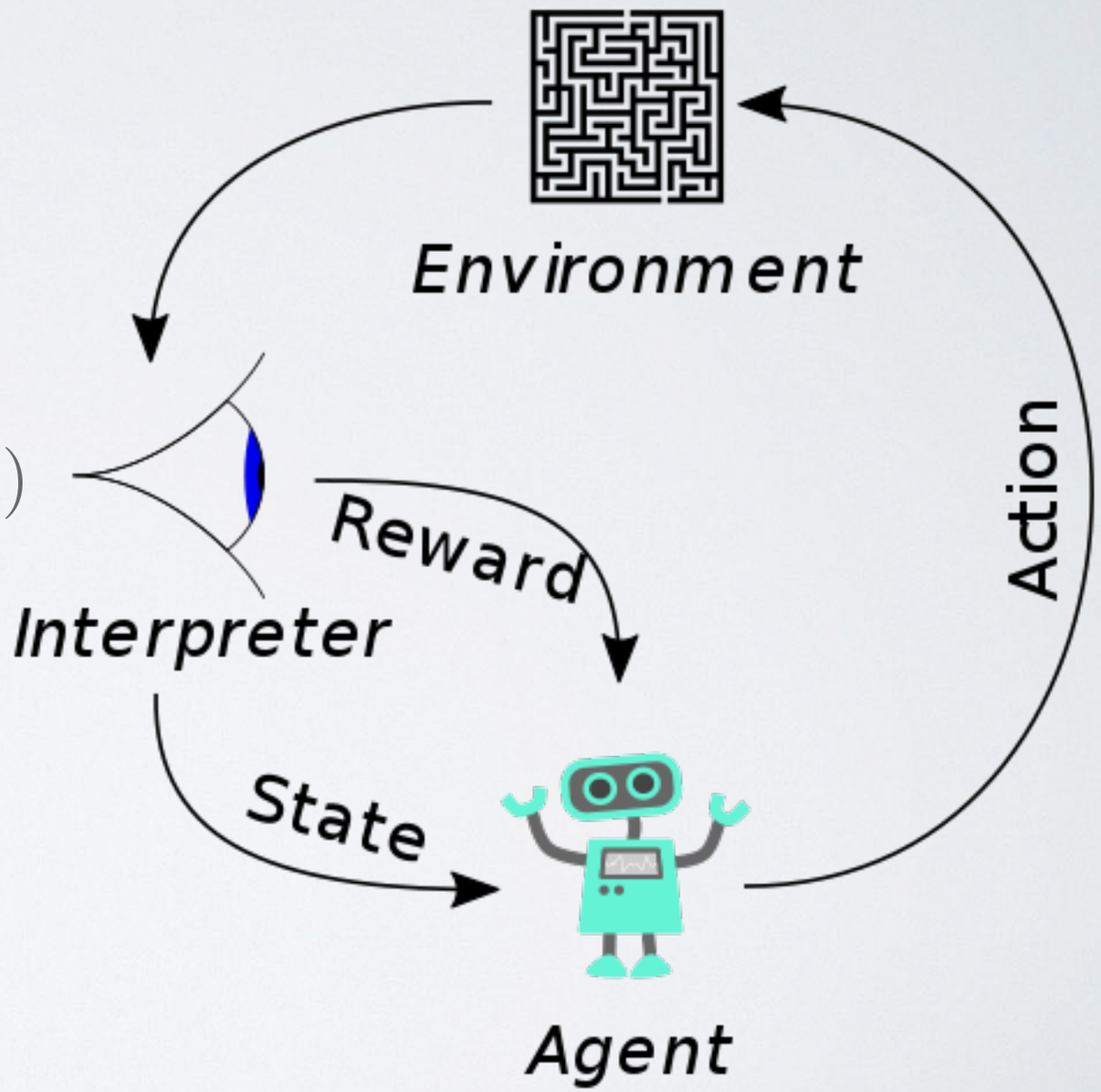


PROBLEM FORMULATION (MATH)

- A set of environment and agent states, *S*
- A set of actions, *A*, of the agent

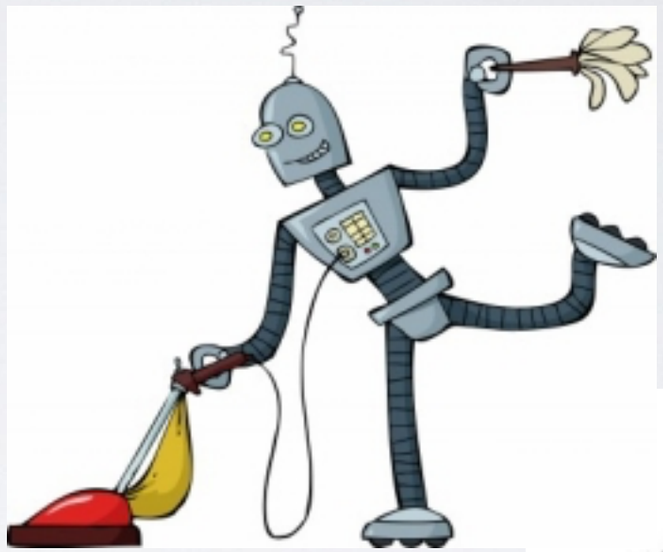
$$P_a(s, s') = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a)$$

- Ra(s,s') is the immediate reward after transition s to s' with action a
- Rules that describe what the agent observes,



WHAT HAS GONE WRONG

- Learning can take too long
- Can't always predict what will happen



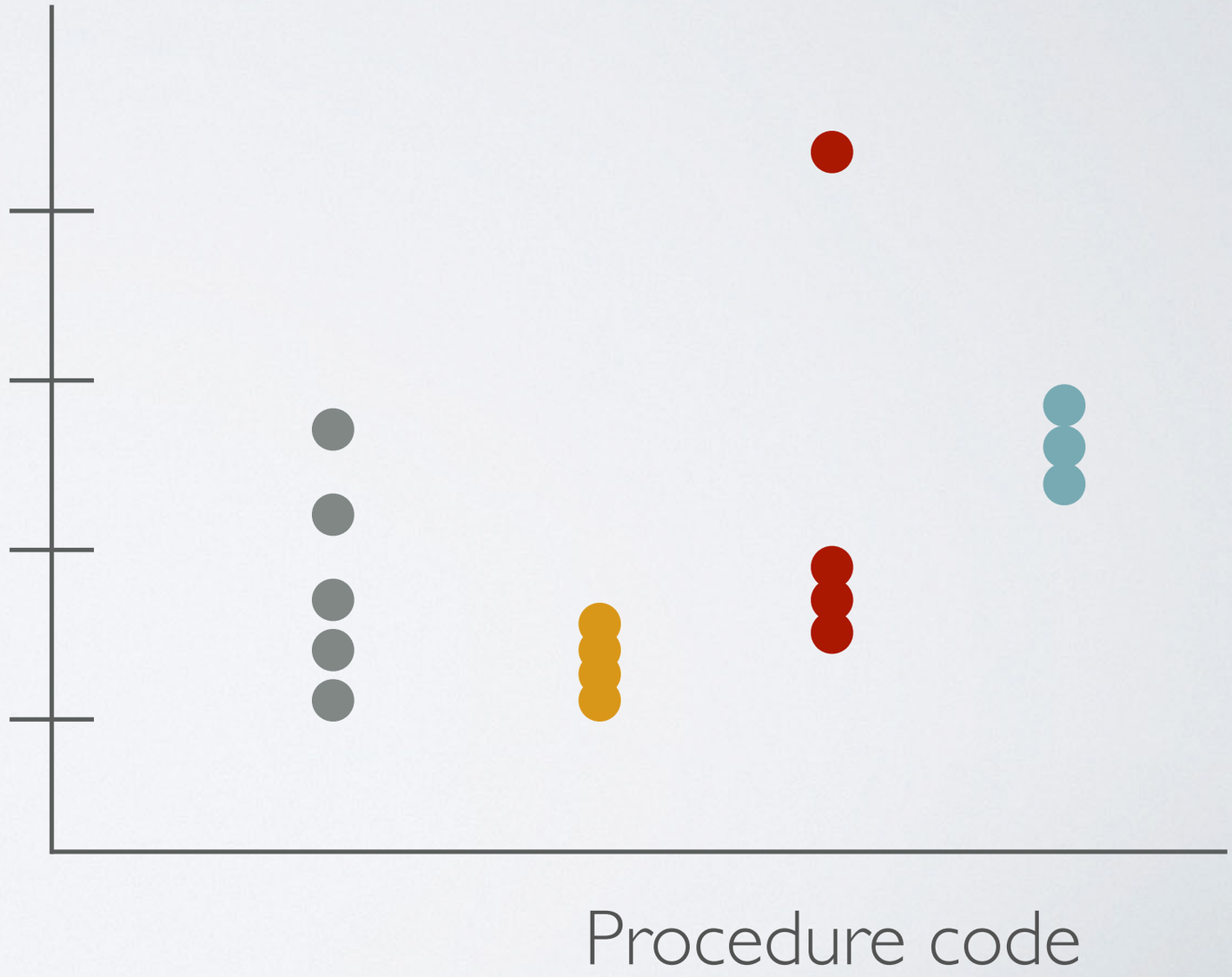
- Avoid Negative Side Effects
- Avoid Reward Hacking
- Scalable oversight
- Safe Exploration
- Robustness to Distributional Shift

from *Concrete Problems in AI Safety*

ANOMALY DETECTION

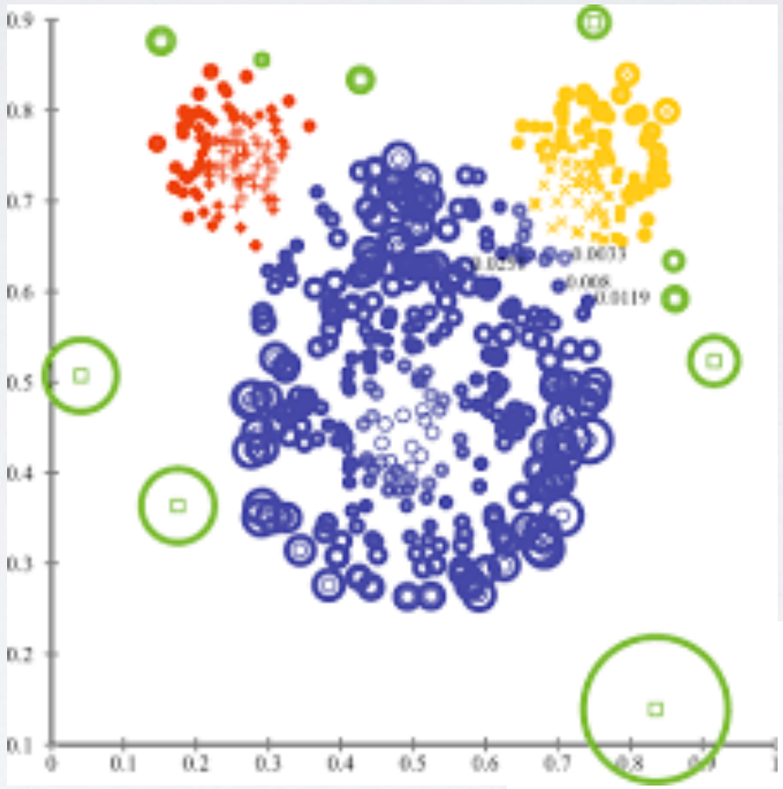
WHAT IF YOU DON'T WANT TO KNOW TRENDS

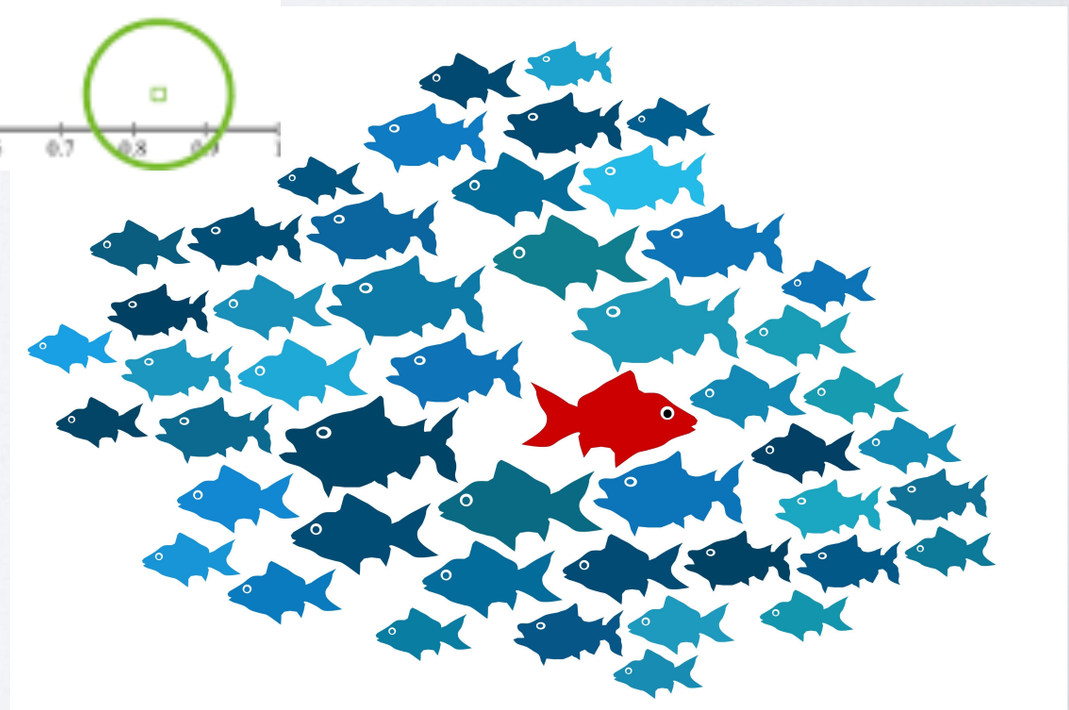
- Instead you want to find outliers
- Anti machine-learning



CLASSES OF PROBLEMS

- Fraud detection
  - Credit card
  - Tax fraud
  - Identity fraud
  - **Medicine**





WHAT HAS GONE WRONG

- An anomaly isn't always an outlier, it's a situation that cannot be explained away

- Pre-pay analytics

- Can't use payment abroad

- Errors not well-explained

EXTRA SLIDES

"Google is a religion posing as a company."

–Paul Saffo at Silicon Valley's Institute for the Future

PLAYING "GOD"

- If Google is a religion, then what is God?

- It would have to be the algorithm?

TECHNOLOGY NEUTRAL

- Intention of creator
- Possibilities and limits of its design
- Foreseen and unforeseen results of its implementation

AI NOW RECOMMENDATIONS

1. Core public agencies, such as those responsible for criminal justice, healthcare, welfare, and education (e.g "high stakes" domains) should no longer use 'black box' AI and algorithmic systems.

2. Before releasing an AI system, companies should run rigorous pre-release trials to ensure that they will not amplify biases and errors due to any issues with the training data, algorithms, or other elements of system design.

3. After releasing an AI system, companies should continue to monitor its use across different contexts and communities.

AI NOW RECOMMENDATIONS

1. Core public agencies, such as those responsible for criminal justice, healthcare, welfare, and education (e.g "high stakes" domains) should no longer use 'black box' AI and algorithmic systems.

2. Before releasing an AI system, companies should run rigorous pre-release trials to ensure that they will not amplify biases and errors due to any issues with the training data, algorithms, or other elements of system design.

3. After releasing an AI system, companies should continue to monitor its use across different contexts and communities.

TECHNOLOGY NEUTRAL

- Intention of creator
- Possibilities and limits of its design
- **Foreseen and unforeseen results of its implementation**