Introduction

Tian-Li Yu

National Taiwan University
Department of Electrical Engineering
tianliyu@ntu.edu.tw

Readings: AIMA Chapters 1 and 2

What Is AI?

John Searle

Strong AI : A physical symbol system can have a mind and mental states.

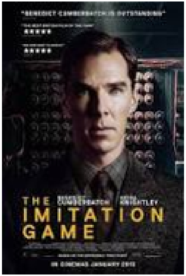
Weak AI : A physical symbol system can act intelligently.

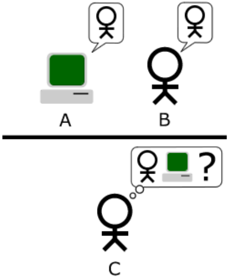
Stuart Russell

	Human	**Rational**
Think	Thinking humanly	Thinking rationally
Act	Acting humanly	Acting rationally

Turing Test (1950)

- Operational Test





Turing Test Applications

- ELIZA (1965).
- Mitsuku (2016 Loebner prize winner).
- CAPTCHA

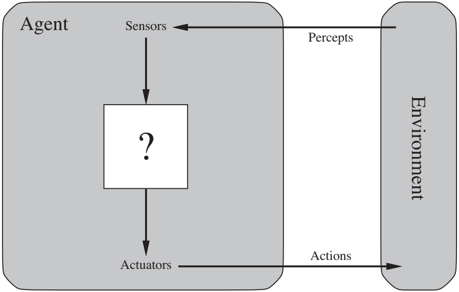
Chinese Room Argument by John Searle

- Suppose it is possible to have a program $\mathcal{P}$ that is sufficient for understanding of Chinese.
- In principle a person in the Chinese room can carry out $\mathcal{P}$.
- But such a person would not understand Chinese.
- So $\mathcal{P}$ is not sufficient for producing understanding of Chinese.
- So there is no program sufficient for producing understanding of Chinese.

Acting Rationally: Rational Agent

- Rational behavior: maximizes the expected performance, given available information.
- Doesn't necessarily involve thinking – *e.g.*, blinking reflex
- Rational $\neq$ Omniscience
  - Percepts may not supply all relevant information.
- Rational $\neq$ Clairvoyant
  - Action outcomes may not be as expected.
- High-level rationality $\Rightarrow$ information gathering, exploration, learning, autonomy.

Agents and Environments



- An agent is an entity that perceives and acts.
- The agent function maps any percept sequences to an action.

$$f : \mathcal{P}^* \rightarrow \mathcal{A}$$

- An agent program is the implementation of an agent function.
- Computational limitations $\rightarrow$ limited rationality.

PEAS

- To design a rational agent, we must specify the task environment.

Agent Type	**P**erformance Measure	**E**nvironment	**A**ctuators	**S**ensors
Taxi driver	safety, destination, profits, legality, comfort	Roads, pedestrians, customers	Steering, accelerator, brake, horn	Accelerometers, camera, engine sensors, GPS

Environment Types

	Chess	Poker	Taxi	Image analysis
Observability				

Environment Types

	Chess	Poker	Taxi	Image analysis
Observability	Fully	Partially	Partially	Fully
Agents				

Environment Types

	Chess	Poker	Taxi	Image analysis
Observability	Fully	Partially	Partially	Fully
Agents	Multi competitive	Multi competitive	Multi cooperative	Single
Deterministic /stochastic				

Environment Types

	Chess	Poker	Taxi	Image analysis
Observability	Fully	Partially	Partially	Fully
Agents	Multi competitive	Multi competitive	Multi cooperative	Single
Deterministic /stochastic	Deterministic	Stochastic	Stochastic	Deterministic
Episodic /sequential				

Environment Types

	Chess	Poker	Taxi	Image analysis
Observability	Fully	Partially	Partially	Fully
Agents	Multi competitive	Multi competitive	Multi cooperative	Single
Deterministic /stochastic	Deterministic	Stochastic	Stochastic	Deterministic
Episodic /sequential	Sequential	Sequential	Sequential	Episodic
Static /dynamic				

Environment Types

	Chess	Poker	Taxi	Image analysis
Observability	Fully	Partially	Partially	Fully
Agents	Multi competitive	Multi competitive	Multi cooperative	Single
Deterministic /stochastic	Deterministic	Stochastic	Stochastic	Deterministic
Episodic /sequential	Sequential	Sequential	Sequential	Episodic
Static /dynamic	Static/Semi	Static	Dynamic	Semi
Discrete /continuous				

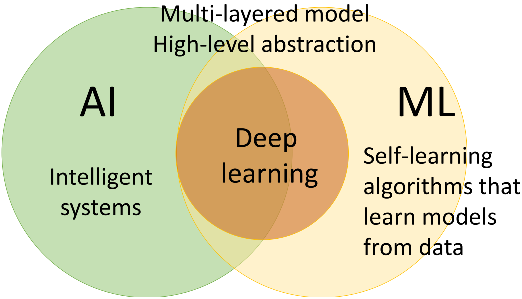
Environment Types

	Chess	Poker	Taxi	Image analysis
Observability	Fully	Partially	Partially	Fully
Agents	Multi competitive	Multi competitive	Multi cooperative	Single
Deterministic /stochastic	Deterministic	Stochastic	Stochastic	Deterministic
Episodic /sequential	Sequential	Sequential	Sequential	Episodic
Static /dynamic	Static/Semi	Static	Dynamic	Semi
Discrete /continuous	Discrete	Discrete	Continuous	Continuous

- The real world is partially observable, stochastic, sequential, dynamic, continuous, multi-agent.

Different Levels of AI

- Level 0: Simple reflex, marketing “AI”.
- Level 1: Search, planning based on some knowledge base.
- Level 2: Learning, exploration.
- Level 3: Automatic feature generation, high-level abstraction.



3 Waves of AI

1950~1970

- Search
- Neural Networks

Debut

1980~1995

- Expert Systems
- Knowledge

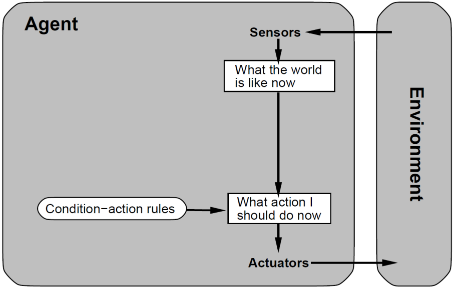
Knowledge Driven

2010~now

- Machine Learning
- Automatic Feature Generation

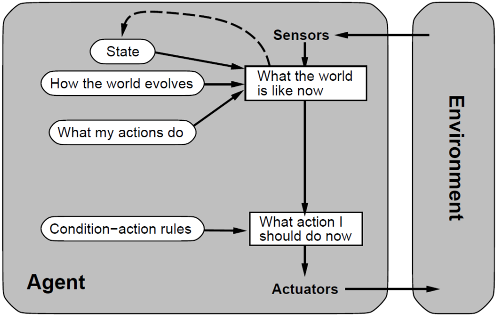
Data Driven

Simple Reflex Agents



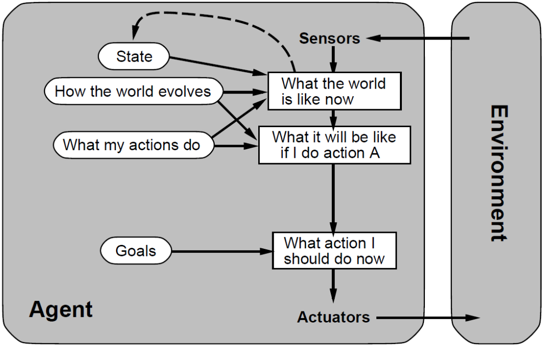
- Actions rely purely on condition-action rules.
- Also called memory-less or state-less
- Works *only if the correct decision can be made on the basis of only the current percept*.
- Works only if the environment is fully observable.
- Often trapped in infinite loops if the environment is partial observable.

Model-Based Reflex Agents



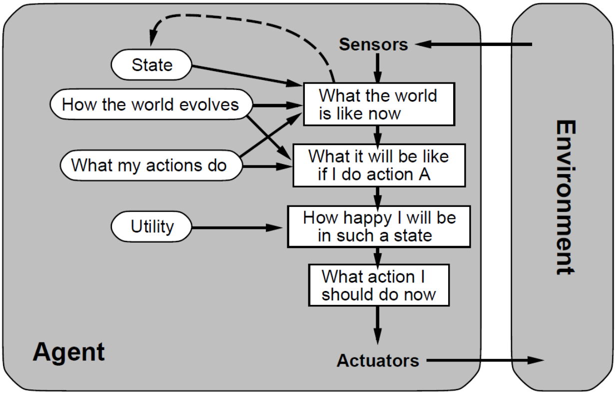
- Handle partial observability by keeping track of the part of the world it can't see now.
- Maintain internal states to model the world.
- The model of the world represents the agent's best guess(es).
- Internal states can also be used to maintain the status of the agent instead of the world.

Goal-Based Agents



- Instead of using condition-action rules, the agent uses goals to decide what action it does.
- Search (bottom-up) and planning (top-down).

Utility-Based Agents



- Utility function: Happiness of the agent.
- Maximizing the expected utility.

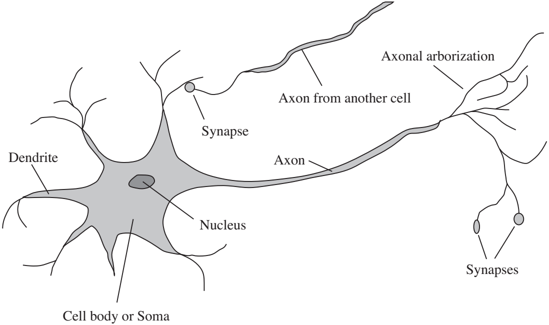
Achievement and Bottlenecks

- Milestone achievement: Deep Blue beat Kasparov in 1997.

- "Seems" mainly for "toy" problems.
- Difficult to directly apply to real problems.
  - How to treat patients?
  - Which product should we develop?

Brains

- 10^{11} neurons of > 20 types, 10^{14} synapses, $1 \sim 10$ms cycle time.

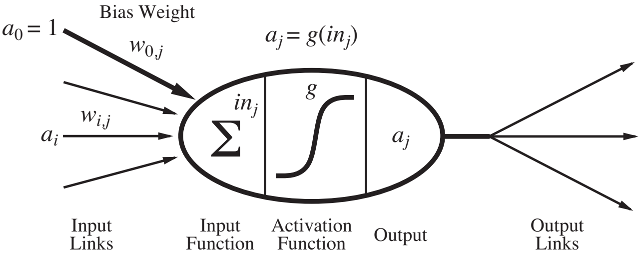


- Hypothesis: mental activity consists primarily of electrochemical activity in networks of neurons.
- Artificial neural networks, connectionism, neural computation.

McCulloch–Pitts Unit (1943)

- Output is a "squashed" linear function of the inputs:

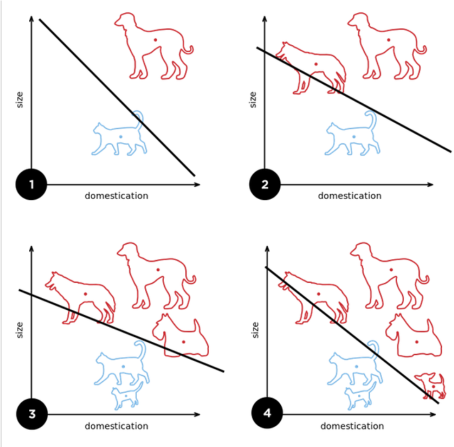
$$a_j = g(in_j) = g(\textstyle\sum_i w_{i,j} a_i)$$



- A gross oversimplification of real neurons, but its purpose is to develop understanding of what networks of simple units can do.
- Usually, the term perceptron is used to refer a neuron or a network of neurons.

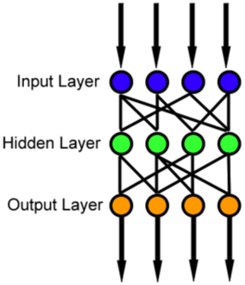
Multi-layer Perceptron (MLP)

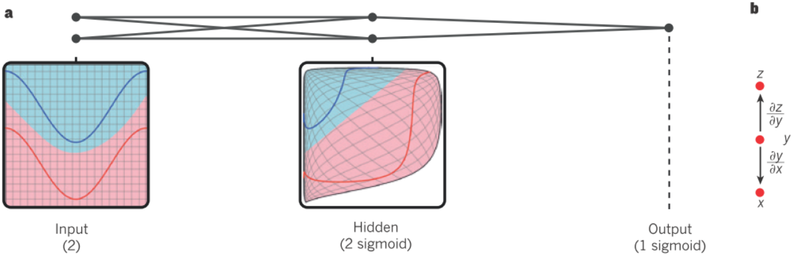
- One perceptron: Linear separation.
- Achieve basic generalization.



Multi-layer Perceptron (MLP)

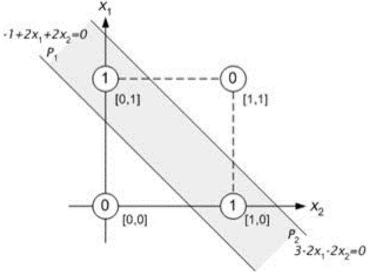
- MLP can map ANY functions.
- Usually use back propagation to train weights.

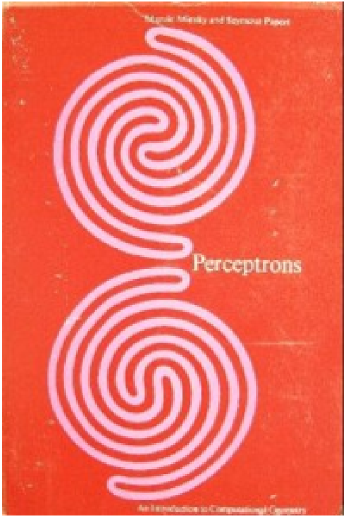




Decline of ANN

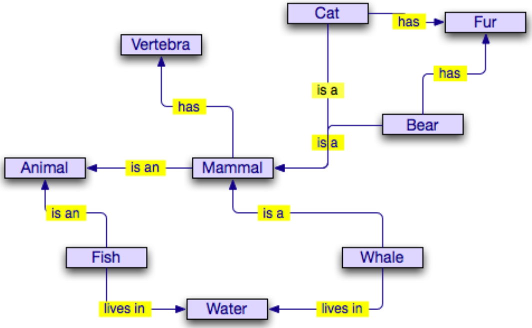
- Minsky and Papert, MIT (1969)
- XOR problem.
- Connection problem.





Expert Systems

- Standford's MYCIN (1960): Diagnosis of infections (69% vs. 80%).
- Cyc project (Lanet, 1984): Semantic network.



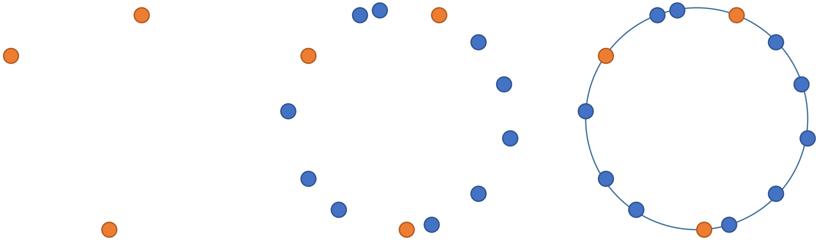
Achievement and Bottleneck

- Watson beat human champions in Jeopardy! in 2011.

- Ontology problem: e.g. what is "part of"?
- Common sense
  - He saw a woman in the garden with a telescope.
- Frame problem (McCarthy).
- Symbol grounding problem (Harnad): Embodiment.

Learning & Data

- Good timing: The Internet (since 90s) provides data.
- ANN contributes to this area.



Machine Learning

- Supervised vs. Unsupervised.
- Classification.

Given:

- Instances $x \in X$: typically described by the attributes.
- Hypotheses $h \in H$.
- Training examples D: Positive and negative examples of the target function $\langle x_1, c(x_1)\rangle, \ldots, \langle x_m, c(x_m)\rangle$ for some target function c: $X \to \{0, 1, \ldots, n\}$.

Determine:

- A hypothesis $h \in H$ such that $h(x) \simeq c(x)$ for all x in D?
- A hypothesis $h \in H$ such that $h(x) \simeq c(x)$ for all x in X?

Learning Setting

- **Given:**
  - Set of instances X, set of hypotheses H, and a target concept c.
  - Set of training instances D of form $\langle x, c(x) \rangle$ generated by a **fixed, unknown probability distribution** $\mathbb{D}$ over X.
- Learner outputs a hypothesis h estimating c, with performance measures:
  - Training error:
$$\frac{|\{x|\ x \in D, c(x) \neq h(x)\}|}{|D|}$$
.
  - True error:
$$Pr_{x \sim \mathbb{D}}(c(x) \neq h(x))$$
  - Off-training set (OTS) error:
$$\frac{|\{x|\ x \notin D, c(x) \neq h(x)\}|}{|X - D|}$$

PAC Learning

- If $c \in H$, there exists a training set D, where $|D| \leq \log_2 |H|$ (more precisely, $VC(H)$), such that a learner gives an h with training error = true error = OTS error = 0.
- As long as the training set is big enough, it is highly probable that the difference between true error and training error is small.
- There exists a learner probably outputs an h that is approximately correct (small true error) (**PAC**). The time complexity is bounded by the probability and accuracy requirements, as well as the complexity of H ($VC(H) \leq \log_2(H)$).

No Free Lunch Theorem (NFL)

- Wolpert *et al.*, 1996 (for ML) & 1997 (for OPT).

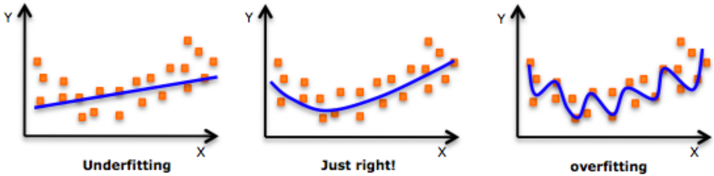
The Inductive Learning Hypothesis

Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved example.

NFL

Without prior (hypothesis), the OTS errors of any two algorithms are equal over all possible senarios.

Prior Hypothesis Bias



Occam's Razor and Minimum Description Length Principle (MDL)

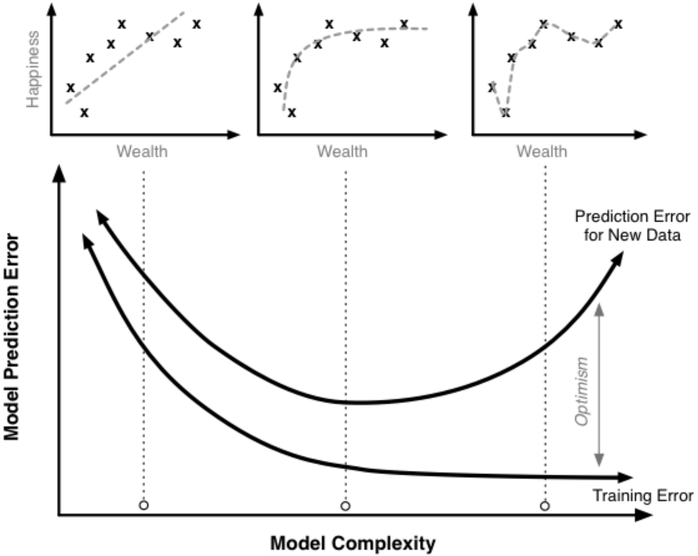
Occam's Razor (William of Occam, 13xx)

- Other things being equal, **simpler** explanations are generally better than more complex ones.
- Two models, describing the data with equal accuracy $\rightarrow$ choose the **simpler** one.

MDL (Rissanen, 1978)

- Extension of Occam's razor.
- Model description & data description given the model.
- Choose the model that overall description length is minimum.

Model Complexity vs. Prediction Error



Cross-Validation

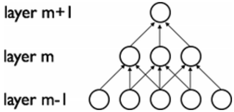
- The idea of having training and testing sets is called cross-validation.
- Holdout cross-validation
  - Randomly split the available data into a training set and a testing set.
  - Simple, fast, but not able to use all available data.
- k-fold cross-validation
  - Randomly split the data into k equal-sized subsets.
  - Perform k rounds of learning using $k-1$ subsets as training and the rest as testing.
  - Popular choice of k is 5 to 10.
  - Accurate statistics, but longer computation.

Deep Learning

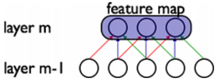
- Shallow
  - Gaussian mixed model (GMM), max entropy, support vector machine (SVM), naïve Bayes, kernel methods.
- Deep
  - Supervised: MLP, DCNN
  - Unsupervised/Semi-supervised: Autoencoder, Deep Belief Networks.
- Why go deep?

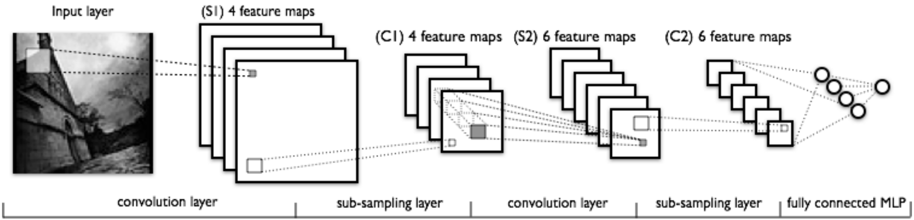
Convolutional Neural Networks (CNN)

- Variant of MLPs inspired from visual cortex.
- Sparse connectivity



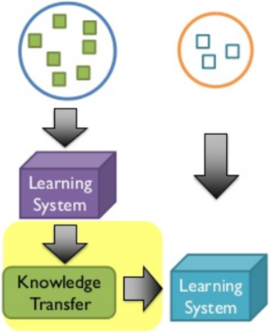
- Shared weights

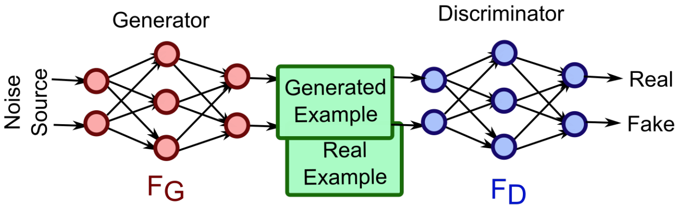




Ongoing Researches

- Reinforcement learning (delayed reward)
- Learning to learn
  - Multi-class/label/task learning
  - Transfer learning
  - One/zero-shot learning
- Generative learning
  - Variational autoencoder (VAE)
  - Generative adversarial network (GAN)





What AI/ML Can Do

- Learn some hidden patterns
  - If there exists one.
  - Given sufficiently many training data.
  - With reasonable prior hypotheses (human's best understanding of the problem).
  - With high probability, output an accurate enough hypothesis.
- Automatically construct high-level features from low-level ones
  - Still, with reasonable prior.
  - Finding suitable representations.
- Current 3rd-wave techniques usually lack of theoretical supports.

What AI/ML Can NOT Do

- Can NOT generate new information out of nothing.
- Can NOT learn useful things without a prior hypothesis (bias).
  - No bias, no learning.
- Can NOT arbitrarily increase SNR.
  - Garbages in, garbages out.