Experimental Design for Machine Learning on Multimedia Data

Lecture 2

Dr. Gerald Friedland,
fractor@eecs.berkeley.edu

Website: http://www.icsi.berkeley.edu/~fractor/fall2020/

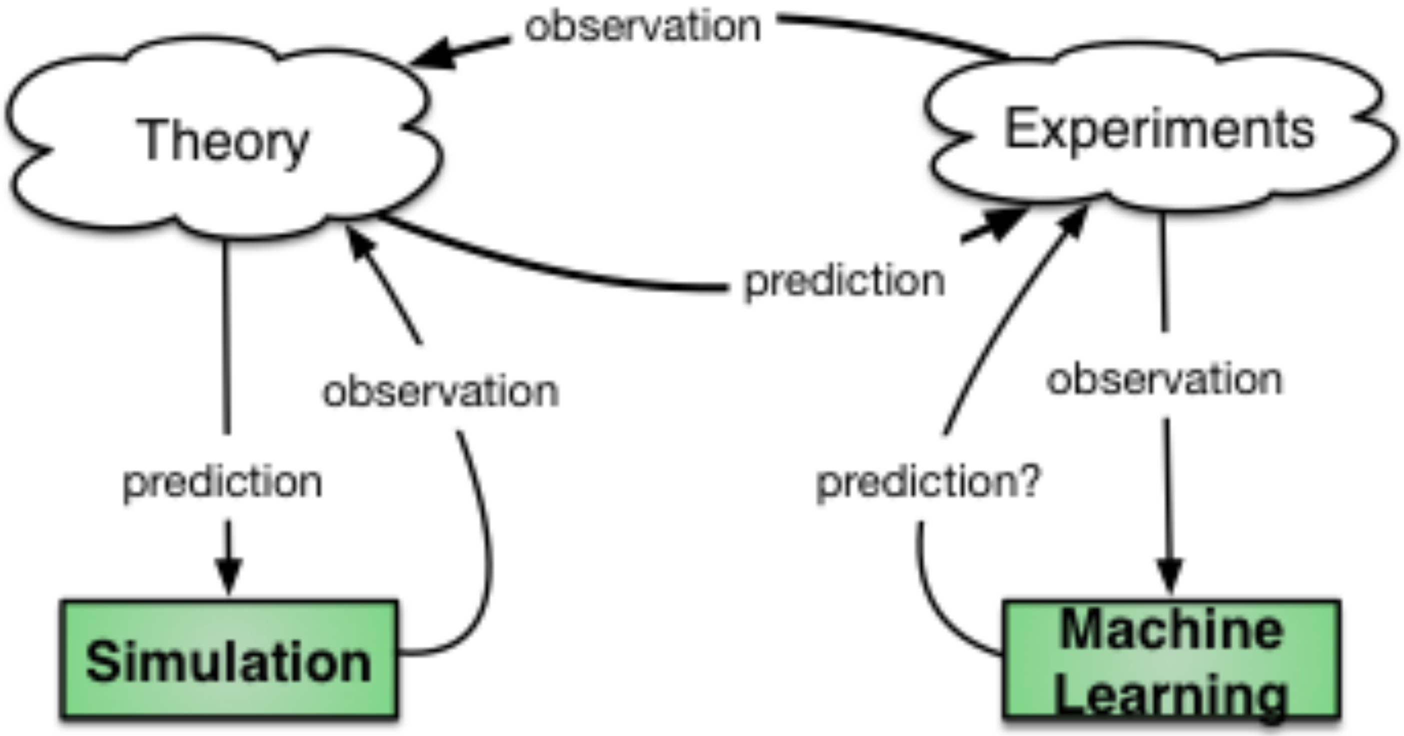
Logistics

- Office Hours:
  Gerald Friedland
  Monday 1pm-2pm
  Same Zoom link.

Today

- Repeat:
  The scientific process and machine learning
- Information flow in the scientific process
- Looking at traditional AI: Shannon and Chess
- Memory Equivalent Capacity

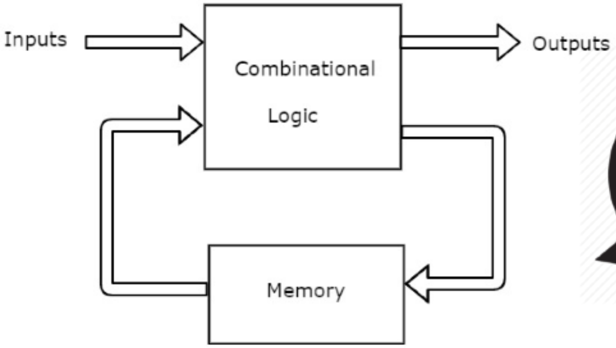
The Scientific Method



Data Science: The Science of Automating the Scientific Method

Reminder: The New Scientific Method

Title	Title	Title	Title	Title	Title	Title
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data



$$E = mc^2$$

Reminder: Thought Framework Machine Learning

- Intelligence: *The ability to adapt* (Binet and Simon, 1904)

- Machine learning *adapts a finite state machine M to an unknown function based on observations*.

- Input: *n* rows of observations (instances) in a table with header:

$$(x_1, x_2, \ldots, x_m, f(\overrightarrow{x}))$$

  where $f(\overrightarrow{x})$ is a column with labels we call target function.

- Output: State machine *M* that maps a point

$$(x_1, x_2, \ldots, x_m) \implies f(\overrightarrow{x})$$

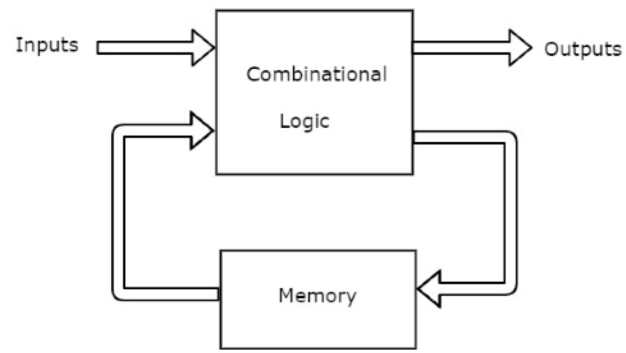
Thought Framework: Machine Learning

- Assume

$$x_i \in \mathbb{R}, f(\overrightarrow{x}) \in \{0,1\}$$

(binary classifier)

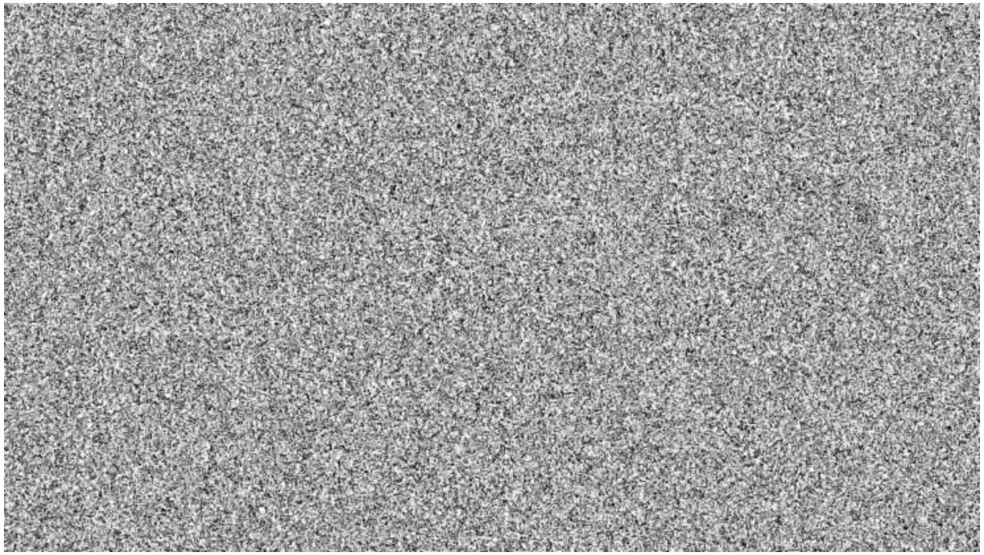
Title	Title	Title	Title	Title	Title	Title
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data



- Question:

How many states does *M* need to model the training data?

A Thought Experiment



- Which image has more information?
- Which image takes more bits of memory?

Refresh: Memory Arithmetic

- ***Information is reduction of uncertainty***:
  $H = -\log_2 P = -\log_2 \frac{1}{\#states} = \log_2 \#states$
  measured in bits.

- Information: $\log_2 \#states$ (positive bits)
  Uncertainty: $\log_2 P = \log_2 \frac{1}{\#states}$ (negative bits)

- If states are not equiprobable, *Shannon Entropy* provides tighter bound.

  Important for homework!

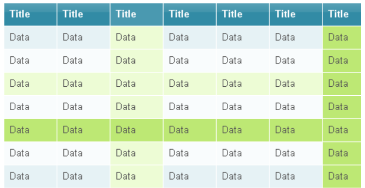
Thought Framework: Machine Learning

Assume

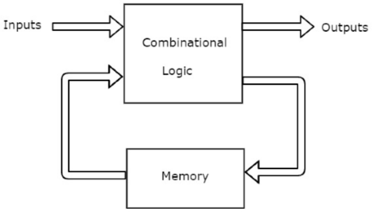
$$x_i \in \mathbb{R}, f(\overrightarrow{x}) \in \{0,1\}$$

(binary classifier)

Question:



Title	Title	Title	Title	Title	Title	Title
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data



How many state (transitions) does *M* need to model the training data?

Maximally: #rows (lookup table)
Minimally: ?

Learning from Chess Players

- C. Shannon 1950: “The game-tree complexity of chess is 10^120”. (Shannon Number)

  <=>

  the Memory Equivalent Capacity of chess using a decision tree is

  $\log_2 10^{120} = 398.63 bits \approx 400 bits$.

  <=>

  Any possible chess game fits into 400 bits of memory.

  <=>

  Starting a chess game, there are -400 bits of uncertainty that need to be reduced to determine the winner.

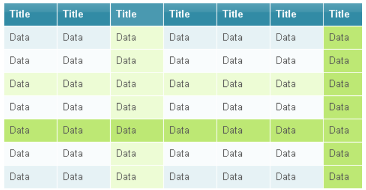
Does it make a difference if we model chess using a Neural Network that observes enough games or using a Python program by translating the human rules?

Thought Framework: Machine Learning

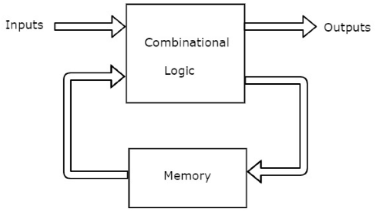
- ***Intellectual Capacity:*** *The number of unique target functions a machine learner is able to represent (as a function of the number of model parameters).*

- ***Memory Equivalent Capacity (MEC):*** *A machine learner's intellectual capacity is memory-equivalent to N bits when the machine learner is able to represent all 2^N binary labeling functions of any N inputs.*

- At MEC or higher, *M* is able to **memorize** all possible state transitions from the input to the output.



Title	Title	Title	Title	Title	Title	Title
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data
Data	Data	Data	Data	Data	Data	Data



Main Engineering Trick

Memorization is worst-case generalization

- Using more parameters than needed for memorization is a waste of resources (CPU, memory, I/O, engineer tuning time).
- Using as many parameters as needed for memorization will most likely not generalize to a held-out data set. This, the machine learner overfits.
- Reducing parameters below memorization capacity will, in the best case, make the machine learner forget what’s not relevant: generalization.

How do we calculate the Memory Equivalent Capacity?

- Binary Decision Tree: Depth of tree (if perfect).
- Neural Network (next lecture)
- Random Forrest: Count non-overlapping nodes.
- GMMs: TBD
- SVN: TBD
- k-NN: TBD

Machine Learning as Engineering Discipline

- Supervised **Machine Learners have a Memory Equivalent Capacity in bits** that is **computable** and **measurable**.
  - Artificial Neural Networks with gating functions (Sigmoid, ReLU, etc.) have
    - a capacity upper limit that can be determined *analytically* using 4 principles
    - an effective capacity that can be measured on actual implementations.
- Predicting and measuring capacity allows for task-independent optimization of a concrete network architecture, learning algorithm, convergence tricks, etc…
- Capacity requirement can be approximately predicted given the input data and ground truth.