MSML610: Advanced Machine Learning

Introduction

Instructor: Dr. GP Saggese - `gsaggese@umd.edu`

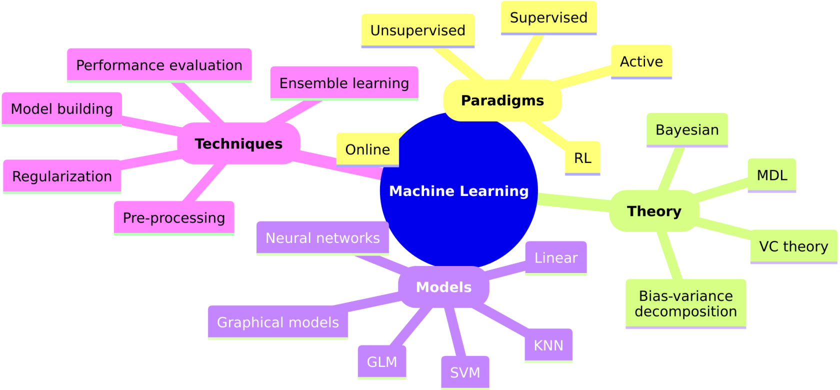
References:

- AIMA Chap 1

- ***A Map of Machine Learning***
- What Is Artificial Intelligence

A Map of Machine Learning

- Machine Learning is a field with many branches
  - Theory
  - Paradigms
  - Models
  - Techniques
  - . . .



Machine Learning Theory

- **VC theory**
  - Measure model capacity and generalize based on hypothesis space complexity
- **Bias-variance decomposition**
  - Prediction error is the sum of:
    - Bias: Error from simplistic model assumptions
    - Variance: Error due to sensitivity to training data fluctuations
- **Computation complexity**
  - Related to information theory and compression
  - E.g., Minimum Description Length (MDL) measures model complexity via efficient model and data description
- **Bayesian approach**
  - Treat ML as probability
  - Combine prior knowledge with observed data to update belief about a model
- **Problem in ML theory**
  - Assumptions may not align with practical problems

Machine Learning Paradigms

- How do you set up the learning problem?
  - **Supervised learning**
    - The dataset includes inputs with corresponding outputs
    - Develop an input-output relationship
  - **Unsupervised learning**
    - The data is unlabeled, discover structure within the data
    - E.g., anomaly detection, clustering
  - **Reinforcement learning**
    - The correct answer is not immediately available
    - Evaluate actions based on final outcomes
  - **Active learning**
    - Not all examples are available initially
    - Request outputs for specific inputs
  - ...

Machine Learning Models

- What is the form of the model and how to fit / predict from the data?
  - Linear models
  - Generalized linear models
    - E.g., logistic, Poisson regression
  - Support Vector Machines (SVM)
  - Nearest neighbors
    - E.g., k-means clustering, KNN
  - Gaussian processes
  - Graphical models
    - Model joint distributions with graphs
    - E.g., hidden Markov models (HMM), Kalman filters, Bayesian networks
  - Neural networks
  - ...

Machine Learning Techniques

- What are the stages of a ML pipeline?
  - **Input processing**
    - Data cleaning
    - Dimensionality reduction
    - Feature engineering
  - **Model building**
    - Models
    - Learning algorithms
  - **Performance evaluation**
    - Cross-validation
    - Bias-variance curves
    - Learning curves
  - **Regularization**
  - **Aggregation**
    - Boosting
    - Bagging
    - Stacking

- A Map of Machine Learning
- ***What Is Artificial Intelligence***
    - AI
    - Machine Learning
    - AI vs ML vs Deep-Learning
    - The Foundation of AI
    - Brief History of AI
    - AI State of the Art
    - Risks and Benefits of AI

- A Map of Machine Learning
- What Is Artificial Intelligence
  - ***AI***
  - Machine Learning
  - AI vs ML vs Deep-Learning
  - The Foundation of AI
  - Brief History of AI
  - AI State of the Art
  - Risks and Benefits of AI

ML, AI, and Intelligence

- Machine Learning is a subset of AI
    - All of it confused with deep learning, large-language models, predictive analytics, ...
- What is artificial intelligence?
- What is intelligence?

Human Intelligence

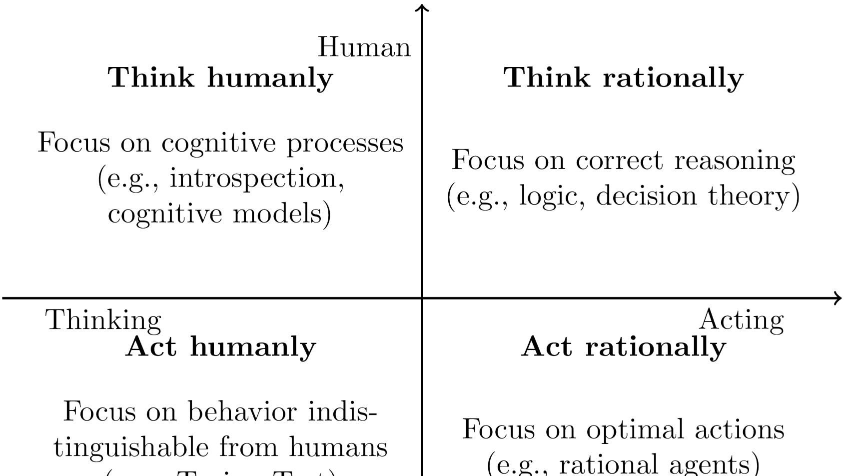
- We call ourselves “homo sapiens” because intelligence sets us apart from animals
- For thousands of years, we tried to understand how we think
- One of the biggest mysteries
  - Brain is a small mass of matter
  - Our brain can understand nature secrets, e.g., theory of relativity, quantum mechanics, black holes in the universe
  - How can brain understand, predict, and manipulate a world more complicated than itself?

Artificial Intelligence

- The term "Artificial Intelligence" was coined in 1956
- **AI aims to:**
  - Understand human intelligence
  - Create intelligent entities
  - *"What I cannot create, I do not understand"* (Feynman, 1988)
- **AI is a technology**
  - Universal and applicable to any human activity and task
  - Its impact greater than any previous historical event
  - Currently generates trillions of dollars annually in revenue
  - Presents many unresolved problems
    - E.g., major concepts in physics might be established

AI Formal Definition

- AI is defined around **two axes**:
  - Thinking vs. Acting
  - Human vs. Rational (ideal performance)
- This leads to four possible definitions of AI as a machine that can:



AI Formal Definition: Quiz

- Four possible definitions of AI as a machine that can:
  1. Think humanly
  2. Think rationally
  3. Act humanly
  4. Act rationally
- **Q**: Which one do you think is the best definition?

AI Formal Definition: Quiz

- Four possible definitions of AI as a machine that can:
  1. Think humanly
  2. Think rationally
  3. Act humanly
  4. Act rationally
- **Q**: Which one do you think is the best definition?
- We will see that building machines that can **"act rationally"** should be ultimate goal of AI

1. AI as Thinking Humanly

- To build machines that think like humans we need to **determine how humans think**
- **Pros**
  - Express precise theory of the human mind as a computer program
- **Cons**
  - Unknown workings of the human mind
  - Anthropocentric definition

2. AI as Thinking Rationally

- What are the rules of **correct thinking**?
  - Given correct premises, yield correct conclusions
- **Logic** studies the “laws of thought”
  - Formalize statements about objects and their relations
- **Automatic theorem proving**
  - Programs solve problems in logical notation
  - Run indefinitely if no solution exists (related to the halting problem)

Thinking Rationally: Cons

1. **Formalizing informal knowledge is difficult**
    - Example: *"A handshake occurs when two people extend, grip, shake hands, then release."*
    - Formal logic representation:

$$
\begin{aligned}
\exists x, y\, (&\text{Person}(x) \wedge \text{Person}(y) \wedge x \neq y \wedge \\
&\text{Hand}(x, h_x) \wedge \text{Hand}(y, h_y) \wedge \\
&\text{MoveToward}(h_x, h_y) \wedge \text{Contact}(h_x, h_y) \wedge \\
&\text{Shake}(h_x, h_y) \wedge \\
&\text{Release}(h_x, h_y))
\end{aligned}
$$

2. **Probabilistic nature of knowledge**
    - Example in medicine: *"Fever, cough, and fatigue could indicate flu, COVID-19, or another illness."*
3. **Scalability challenges**
    - Large problems may need heuristics for practical solutions
4. **Intelligence requires more than rational thinking**
    - Importance of agent interaction with the world
    - Problem of the "body"

3. AI as Acting Humanly

- **Agent** is something that perceives and acts to reach a goal
- **Definition**: AI designs **agents that can act like humans**
- **(Embodied) Turing test**
  - *"A computer passes the Turing test if a human cannot tell whether the answers to questions came from a person or a computer"*
  - Passing the Turing test requires:
    1. Natural language processing to communicate
    2. Knowledge representation to store information
    3. Automated reasoning to use stored knowledge and answer questions
    4. Machine learning to detect patterns
    5. Computer vision and speech recognition to perceive objects and understand speech
    6. Robotics to manipulate objects and move

Turing Test: Pros and Cons

- **Pros**
  - Operational definition of intelligence
  - Sidestep philosophical vagueness
    - "What is consciousness?"
    - "Can a machine think?"
    - ...
- **Cons**
  - Anthropomorphic criteria define intelligence in human terms
    - Multiple forms of non-human intelligence exist
  - Intelligence in terms of Turing test
    - Design intelligence to imitate humans
    - Fool humans into thinking it's human
  - E.g., aeronautical engineering:
    - Yes: Focus on wind tunnels and aerodynamics
    - No: Designing machines that imitate birds
    - No: Fooling birds into thinking it's a bird

4. AI as Acting Rationally

- **Rational agents**: agents that do the "right thing" given what they know
- Agents that act rationally should:
  1. Operate autonomously
  2. Perceive environment
  3. Persist over a prolonged time period
  4. Adapt to change
  5. Create and pursue goals

Acting Rationally as Ultimate Goal of AI

- Which definition of AI to use?
  - Acting vs. Thinking
  - Rational vs. Human
- **Acting $>$ Thinking**
  - Acting rationally is broader than just thinking rationally
- **Rational $>$ Human**
  - Rationality can be mathematically defined
  - Human behavior is shaped by evolutionary conditions
- AI focuses on **agents acting rationally**

Rationally is Not Absolute

- AI wants to build agents that **do the right thing** based on **available knowledge**
- E.g., you leave the house and a branch strikes you
  - **Q**: Did you act rationally?

Rationally is Not Absolute

- AI wants to build agents that **do the right thing** based on **available knowledge**
- E.g., you leave the house and a branch strikes you
  - **Q**: Did you act rationally?
  - Probably

Rationally is Not Absolute

- AI wants to build agents that **do the right thing** based on **available knowledge**
- E.g., you leave the house and a branch strikes you
    - **Q**: Did you act rationally?
    - Probably
- E.g., you cross the street and a car knocks you over
    - **Q**: Did you act rationally?

Rationally is Not Absolute

- AI wants to build agents that **do the right thing** based on **available knowledge**
- E.g., you leave the house and a branch strikes you
  - **Q**: Did you act rationally?
  - Probably
- E.g., you cross the street and a car knocks you over
  - **Q**: Did you act rationally?
  - It depends, but probably no

Goals of a Rational Agent

- A **rational agent** aims for:
  - The best outcome in a deterministic setup
  - The best expected outcome under uncertainty
- **Best** is determined by the objective function:
  - E.g., cost function, sum of rewards, loss function, utility
- **Problems** with acting rationally
  - Sometimes no provably correct action exists
    - Yet, an action must be taken
  - Perfect rationality is not always feasible due to:
    - Cost of acquiring all data
    - Computational demands
  - Acting "appropriately" vs "satisficing"

Machine Learning: Definitions

- How to define machine learning?
- *"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"*, (Arthur Samuel, 1959)
- *"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if $P(T)$ improves with experience E"*, (Tom Mitchell, 1998)
- **Machine learning** is building machines to do **useful things** without being **explicitly programmed**
  - E.g. a computer learns to play checkers by playing against itself, memorizing positions that lead to winning

Limits of ML Compared to Human Intelligence

- Artificial intelligence differs from human intelligence
  - Machines don’t learn like humans (e.g., LLMs)
- **Fragility to input variations**
  - ML models fail with slight input distortions
    - Adversarial attacks cause misclassification by altering one pixel
    - A model trained for a video game may fail if the screen is slightly rotated; humans continue effortlessly
- **Lack of transfer learning**
  - ML systems cannot apply knowledge across domains without retraining
- **Massive data and compute requirements**
  - ML requires enormous datasets and computational resources
    - A teenager learns to drive in hours; self-driving systems need billions of compute hours and extensive data
- **Poor common sense and reasoning**
  - ML lacks built-in world knowledge and intuitive logic

Limits of ML Compared to Human Intelligence

- **Opaque decision-making**
  - Many ML models offer little transparency into decision processes
    - Limits trust, interpretability, and accountability in critical applications
- **Dependence on narrow objectives**
  - ML systems excel at optimizing narrow tasks but fail with ambiguous goals
    - E.g., an algorithm maximizing user engagement may promote harmful content
- **Susceptibility to bias and data quality**
  - Models inherit and amplify biases in training data
- **Lack of embodiment and physical interaction**
  - Human cognition is grounded in physical and sensory experience

The 3 Machine Learning Assumptions

- Machine learning involves solving a practical problem by:
  - Gathering a dataset
  - Building a statistical model from the dataset algorithmically
- The **three assumptions** of machine learning
  - A pattern exists
  - Pattern cannot be precisely defined mathematically
  - Data is available
- Which ML assumption is essential?
  - *A pattern exists*
    - If no pattern, try learning, measure effectiveness, conclude it doesn't work
  - *Cannot pin down pattern mathematically*
    - If solution is direct, ML not recommended, but may still apply
  - *We have data*
    - Without data, no progress can be made
    - Data is crucial

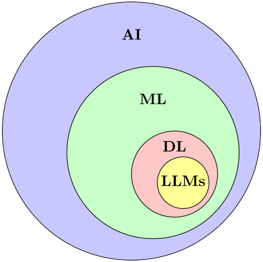
Machine Learning Adages

- *"An explanation of the data should be as simple as possible, but not simpler"* (Einstein)
- *"The simplest model that fits the data is also the most plausible"* (Occam's razor)
- *"Garbage in, garbage out"* (Fuechse, 1957)
- *"All models are wrong, but some are useful"* (George E. P. Box, 1976)
- *"If you torture the data long enough it will confess whatever you want"* (Coase, 1982)
- *"Data is the new oil"* (Humby, 2006)
- *"More data beats clever algorithms"* (Norvig, ~2006)
- *"The unreasonable effectiveness of data"* (Halevy, Norvig, Pereira, 2009)

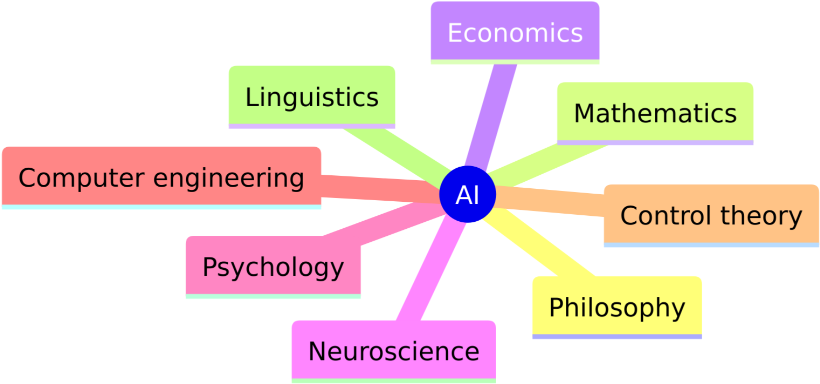
- A Map of Machine Learning
- What Is Artificial Intelligence
    - AI
    - Machine Learning
    - ***AI vs ML vs Deep-Learning***
    - The Foundation of AI
    - Brief History of AI
    - AI State of the Art
    - Risks and Benefits of AI

AI vs ML vs Deep-Learning

- **AI**: Machines programmed to reason, learn, and act in a rational way
- **ML**: Machines capable of performing tasks without being explicitly programmed, e.g.,
  - Natural language processing
  - Computer vision
  - Speech recognition
- **AI without ML**:
  - Example: Rule-based systems (e.g., IBM Deep Blue playing chess)
- **Deep Learning (DL)**: ML using neural networks with multiple layers
  - Example: Autonomous vehicles
- **Large Language Models (LLM)**: neural networks trained on massive text datasets and RLHS



AI Relates to Many Other Disciplines



AI and Philosophy (1/2)

- **Can formal rules be used to draw valid conclusions?**
  - Reasoning
    - Aristotle (400 BCE) formulated laws governing the rational mind
    - Machines were built for arithmetic operations (e.g., Pascaline, 1600)
    - Logic studies rules of proper reasoning
  - Rationalism
    - Use reasoning to understand the world
- **How does the mind arise from a physical brain?**
  - Dualism
    - Nature follows physical laws
    - Part of the human mind ("the soul") is exempt from physical laws
  - Materialism
    - The mind is a physical system, following the laws of physics
    - Where is free will? Free will is the perception of available choices

AI and Philosophy (2/2)

- **What does knowledge come from?**
  - Empiricism
    - Knowledge via senses
    - Example: learn that trees are green by looking at them
  - Induction
    - General rules from associations
    - Example: seeing many swans are white, inferring all swans are white
  - Logical Positivism
    - Knowledge as logical theories linked to sensory observations
    - Example: scientific hypotheses connected to experimental data

- **How does knowledge lead to action?**
  - Utilitarianism
    - Measures "utility" linking knowledge to action
    - Actions justified by logic connecting goals and outcomes
  - Consequentialism
    - Right or wrong determined by action's expected outcomes
    - E.g., "If you kill, you will go to jail"
  - Deontological ethics
    - Opposes consequentialism
    - "Right actions" based on universal laws, not outcomes
    - E.g., "don't kill", "don't lie"

AI and Cognitive Psychology

- **How do humans think and act?**
  - Cognitive Psychology
    - Brain as an information-processing device
    - Stimuli translated into internal representation
    - Representation manipulated by cognitive processes to derive new internal representations ("beliefs")
    - Representations turned into actions ("goals")
  - Cognitive Science
    - Use computer models to address memory, language, and logic thinking
  - Human-Computer Interaction (HCI)
    - From artificial intelligence (AI) to intelligence augmentation (IA)
    - Computers augment human abilities

AI and Mathematics

- **What are the formal rules to draw valid conclusions?**
  - Formal Logic
    - Logical deduction rules (Boole, 1850)
    - First-order logic includes objects and relations (Frege, 1879)
  - Limits to Deduction
    - Some statements are "undecidable"
    - Incompleteness theorem: true statements exist that cannot be proved in any formal theory (Godel, 1931)
- **How do we reason with uncertain information?**
  - Probability
    - Mathematics of uncertainty
    - Key contributors: Cardano, Pascal, Bernoulli, Bayes (1500-1700)
  - Statistics
    - Combines data with probability
    - E.g., experiment design, data analysis, hypothesis testing, asymptotics

AI and Economics (1/2)

- **How to make decisions to maximize payoff given preferences?**
  - Economies
    - Agents maximize economic well-being (utility)
    - Studies desires and preferences
  - Decision theory
    - Making decisions under uncertainty for preferred outcomes
    - Probability theory + utility theory
    - Examples: Investment choices, policy decisions
- **How to make decisions when payoffs are result of several actions?**
  - Operations Research
    - Make rational decisions with payoffs for sequence of actions (Bellman, 1957)
    - E.g., Markov Decision Processes
  - Satisficing
    - Decisions that are good enough
    - Closer to human behavior
    - Example: Choosing a restaurant that meets basic criteria rather than finding the perfect one

AI and Economics (2/2)

- **How multiple agents with different goals act?**
  - **Large Economies**
    - Many agents with no mutual impact
    - Ignore other agents' actions
    - E.g., national economy where individual actions don't affect market
  - **Small Economies**
    - One player's actions influence others' utility
    - E.g., local market where one seller's pricing affects competitors
  - **Game Theory**
    - Small economies resemble a "game" (Von Neumann, 1944)
    - Rational agents might need randomized strategies
    - E.g., rock-paper-scissors where randomization prevents predictability

AI and Linguistics

- **How can you create systems that understand natural language?**
  - Computational linguistics (NLP) studies sentence structure and meaning
    - Machine translation (e.g., Google Translate)
    - Sentiment analysis in social media
    - Automated customer support chatbots
- **How does language relate to thought?**
  - Knowledge representation: how to represent knowledge for computer reasoning
  - E.g., first order knowledge, knowledge graphs

AI and Neuroscience

- **Brain**
  - Parts handle specific cognitive functions
  - Information processing in the cerebral cortex
  - E.g., frontal lobe injury may impair decision-making
- **Anatomy of the Brain**
  - Composed of neurons (~100 billion)
    - Each neuron connects with 10-100k others via synapses
    - Axons enable long-range connections
  - Signals propagate through electrochemical reactions
  - Short-term pathways support long-term connections (learning)
- **Memory**
  - No theory yet about individual memory storage
  - Current theory: memories reconstructed

The Brain Causes the Mind

- *Simple cells lead to thought and consciousness*
  - Truly amazing conclusion!
  - Complex processes emerge
  - Supercomputers' complexity rivals the brain
  - Achieving brain's intelligence level remains unknown
- **Brain-Machine Interface**
  - Brain adjusts to devices (e.g., learn to use prosthetics as limbs)
- **AI Singularity**
  - Future point when AI surpasses human intelligence
    - AI improves autonomously, leading to rapid growth
    - Recursive self-improvement leads to superintelligence
  - Potential societal impact
    - Control problem/value alignment: ensure AI aligns with human values
    - Economic/social disruption due to automation
  - Hard to predict

AI and Computer Science

- **What can be computed?**
  - Algorithm
    - Procedure to solve problems
    - E.g., algorithm for computing GCD (Euclid, 300 BCE)
  - Limits to Computation
    - Turing machine (1936): Computes any computable function
    - Some functions are non-computable
    - E.g., the halting problem, i.e., decide if a program terminates
  - Tractability
    - Problem is intractable if solving time grows exponentially with size
    - Complexity classes: Polynomial vs exponential complexity (NP-problems)

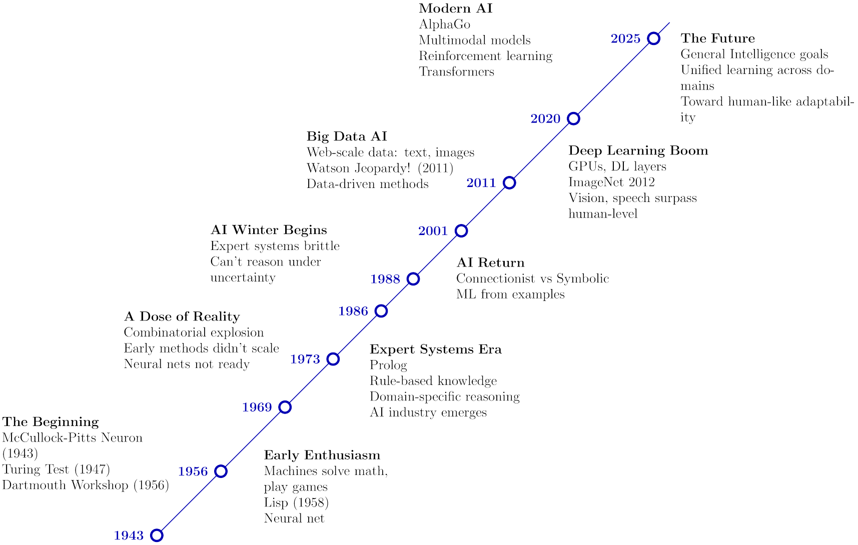
AI and Control Theory and Cybernetics

- **How can artifacts operate under their own control?**
  - Control theory
    - Study self-regulating feedback control systems
    - E.g., a water regulator that maintains a constant water flow
    - Mechanisms to minimize error between current and goal states
    - Kalman Filter (Kalman, 1960)
    - Based on calculus, matrix, stochastic optimal control
    - AI: logical inference, symbolic planning, computation

AI and Computer Engineering

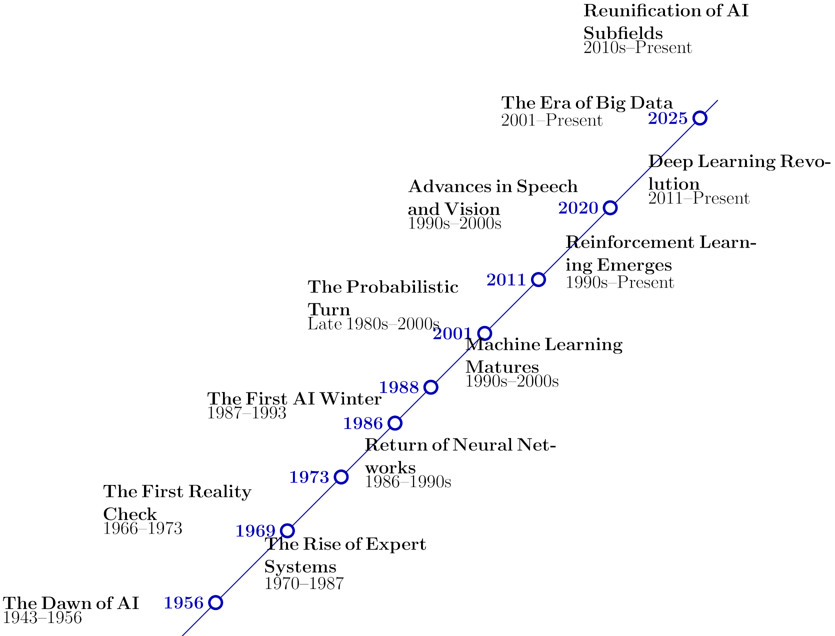
- **How can we build an efficient computer?**
  - Electronic computers
    - Built during World War II
  - Moore's Law
    - Performance doubled every 18 months (1970-2005)
    - Power and scaling issues shifted focus to multi-core over clock speed
  - Hardware for AI
    - GPUs
    - TPUs
    - Wafer-scale engines
  - Current Trends
    - Massive parallelism (like brain function)
    - Computing power doubling every 3 months
    - GPUs / TPUs used in deep learning
    - High precision (e.g., 64b) often unnecessary
  - Quantum Computing
    - Potential for significant acceleration in key computations
    - E.g., Shor's algorithm for factorization

AI Timeline



1943
The Beginning
McCullock-Pitts Neuron (1943)
Turing Test (1947)
Dartmouth Workshop (1956)
1956
Early Enthusiasm
Machines solve math, play games
Lisp (1958)
Neural net
1969
A Dose of Reality
Combinatorial explosion
Early methods didn't scale
Neural nets not ready
1973
Expert Systems Era
Prolog
Rule-based knowledge
Domain-specific reasoning
AI industry emerges
1986
AI Winter Begins
Expert systems brittle
Can't reason under uncertainty
1988
AI Return
Connectionist vs Symbolic
ML from examples
2001
Big Data AI
Web-scale data: text, images
Watson Jeopardy! (2011)
Data-driven methods
2011
Deep Learning Boom
GPUs, DL layers
ImageNet 2012
Vision, speech surpass human-level
2020
Modern AI
AlphaGo
Multimodal models
Reinforcement learning
Transformers
2025
The Future
General Intelligence goals
Unified learning across domains
Toward human-like adaptability

AI Timeline (v2)



The Beginning (1943-1956)

- **Artificial neuron**
  - Model (McCullock-Pitts 1943) based on:
    - Brain physiology
    - Propositional logic
    - Computation theory
  - Compute any function with connected neurons
    - Neuron on/off based on stimulation from neighboring neurons
    - Implement logical AND, OR, NOT with simple neuron networks
- **Alan Turing, 1947**
  - Turing test, machine learning, reinforcement learning
  - Create human-level AI:
    - Develop learning algorithms
    - Teach machine like a child
- **Birth of AI**
  - McCarthy organized first AI workshop (1956)
  - The Logic Theorist
    - Programs to “think non-numerically” and prove theorems
    - Newell and Simon (1956)

Enthusiasm and Great Expectations (1952-1969)

- **Early years of AI were full of successes**
  - Until than computers could only do arithmetics
  - *"A machine can never do games, puzzles, IQ tests, . . ."*
  - AI researchers showed machines could do that after another
- **General Problem Server**
  - Imitate human problem-solving
  - Consider sub-goals and possible actions
- **Program learned to play checkers**
  - Use reinforcement learning from victories and mistakes
- **Lisp (1958)**
  - High-level language used for 30 years in AI
- **First neural network**
  - 3000 vacuum tubes for 40 neurons
  - Marvin Minsky (1959)
- **MIT and Stanford**
  - Minsky at MIT
    - Focus on neural network
  - McCarthy at Stanford
    - Focus on representation, logic

A Dose of Reality (1966-1973)

- Researchers were confident about AI's upcoming successes
- AI didn't succeed on real problems due to:
  - Solutions based on human problem-solving methods
  - Difficulty handling "combinatorial explosion":
    - Theorem proving handles small problems with brute force, but doesn't scale
  - Neural networks needed algorithms (e.g., backpropagation), compute power, and data

Expert Systems (1969-1979)

- **Weak AI (Narrow AI)**
  - Performs specific tasks, not general reasoning
  - Operates in a limited, well-defined domain
  - Uses "weak methods" (search, logic) that struggle to scale
  - Improves with domain-specific knowledge
- **Expert Systems**
  - Known as "knowledge-based systems"
  - Combine weak methods with extensive domain knowledge as rules
  - Use inference engines to apply rules to facts
  - E.g., rule-based systems, logic programming (e.g., Prolog)
- **Commercial Adoption and Industry Growth (1980–)**
  - AI shifted to practical applications
  - Major US corporations deployed expert systems
  - AI emerged as a commercial industry

(First) AI Winter (1980)

- AI overconfidence/hype didn't deliver
- **Reasons:**
  - Building/maintaining expert systems is difficult
  - Reasoning methods ignore uncertainty
  - Systems can't learn from experience
  - E.g., expert systems in medical diagnosis struggle with complex, variable patient data
  - E.g., early AI chess systems couldn't adapt to new strategies without manual updates

Return of Neural Networks (1986-)

- Mid-1980s: Discovered back-propagation algorithm
  - Developed in early 1960s
- **Connectionist approach vs. Symbolic approach**
  - Connectionist: Neural networks
    - E.g., recognizing handwritten digits
  - Symbolic: General Problem Solver
    - E.g., solving logical puzzles with rules
- **Why connectionist approach**
  - Concepts not well-defined using symbolic axioms
    - Forms fluid internal concepts
    - Represents real-world complexity better
  - Neural networks learn from examples
    - Adjust parameters for improved predictions
  - E.g.,
    - Image recognition: Identify objects by learning from labeled images

Probabilistic Reasoning and ML (1987-)

- **AI and scientific method**
  - Rigorous methods to test performance
  - E.g., speech recognition, handwritten character recognition
- **Benchmarks for progress**
  - Examples:
    - MNIST: Handwritten digit recognition
    - ImageNet: Image object recognition
    - SAT Competitions: Boolean satisfiability solvers
- **AI shifts ...**
  - From Boolean logic to probability
  - From hand-coded rules to machine learning
  - From a-priori reasoning to experimental results

Progress in Speech Recognition

- **1970s: Ad-hoc approaches**
  - Various architectures and approaches were attempted
  - Rule-based systems with limited robustness
  - Cons: Ad-hoc, fragile
- **1980s: HMMs**
  - Hidden Markov Models (HMMs) became dominant
  - Pros: Strong theoretical foundation
  - Methods: Effective learning techniques
  - Data: Trained on large speech corpora
  - No claim humans use HMMs for speech recognition

Bayesian Networks

- **Bayesian networks**
  - Judea Pearl, 1988
  - AI is linked with:
    - Probability
    - Decision theory
    - Control theory
  - Efficiently represent uncertainty
  - Provide rigorous reasoning
- **Examples**
  - Diagnosing diseases based on symptoms
  - Predictive text input in smartphones
  - Fraud detection in banking

Reinforcement Learning

- **Reinforcement learning**
  - Sutton, 1988
  - Reinforcement Learning (RL) involves agents learning by interacting with an environment
  - Markov Decision Problems (MDPs) provide a mathematical framework for modeling decision-making
- **Examples**
  - RL: A robot learning to navigate a maze by receiving rewards for successful paths
  - MDPs: A game strategy modeled where each move influences the outcome with certain probabilities

Reunification

- **Reunification of AI with**:
  - Data
  - Statistical modeling
  - Optimization
  - Machine learning
- **Many subfields of AI were re-unified**:
  - Computer vision
  - Robotics
  - Speech recognition
  - Multi-agent systems
  - NLP

Big Data (2001-Present)

- **Focus shifts from algorithms to data**
  - For 60 years, AI focused on algorithms and models
  - For some problems, data availability matters more than algorithms, e.g.,
    - Trillions of English words
    - Billions of web images
    - Billions of speech and video hours
    - Social network data
    - Click stream data
  - Algorithms leverage large datasets
- In 2011, IBM's Watson beat human Jeopardy! champions

Deep Learning (2011-Present)

- **Deep learning**
    - Use ML models with multiple layers of computing elements
    - Ideas known since 1970s, but then forgot
    - Success in handwritten digit recognition in 1990s
- **In 2011, DL took off**
    - Surge of interest in AI among researchers, students, companies, investors, government, and public
    - In 2012, a DL system showed dramatic improvement in `ImageNet` competition
        - Previous systems used handcrafted features
- Today, DL exceeds human performance in several vision and speech recognition tasks
- DL needs specialized hardware (e.g., GPU, TPU, FGPA) for parallel tensor operations
- **General Artificial Intelligence**
    - Universal algorithm for learning and acting, not just specialized tasks (e.g., driving, playing chess, recognizing speech)

Progress in AI Research

- **Between 2010 and 2019**
  - AI papers increased 20x
    - 1,000 $\rightarrow$ 20,000
  - Student enrollment in AI and CS increased 5x
    - 10,000 $\rightarrow$ 50,000
  - NeurIPS attendance increased 8x
    - 1,000 $\rightarrow$ 8,000
  - AI startups increased 20x
    - 100 $\rightarrow$ 2,000
- **Compute**
  - Training times dropped 100x in 2 years
  - AI computing power doubles every 3 months

What Can AI Do Today? (1/2)

- **Robotic vehicles**
  - Waymo passed 10 million miles without serious accident
- **Legged locomotion**
  - BigDog recovers on ice
  - Atlas walks on uneven terrain, jumps on boxes, backflips
- **Autonomous planning and scheduling**
  - Space probes, Mars rovers
- **Machine translation**
  - Translates 100 languages with human-level performance
- **Speech recognition**
  - Real-time speech-to-speech with human-level performance
  - AI assistants
- **Recommendations**
  - ML recommends based on past experiences
  - Spam filtering 99.9% accuracy
  - E.g., Amazon, Facebook, Netflix, Spotify, YouTube

What Can AI Do Today? (2/2)

- **Game playing**
  - 1997: Deep Blue defeated Kasparov
  - 2011: Watson beat Jeopardy! champion
  - 2017: AlphaGo beat Go champion
  - 2018: AlphaZero super-human in Go, chess with only rules, self-play
  - AI beats humans in videogames: Dota2, StarCraft, Quake
- **Image understanding**
  - Object recognition
  - Image captioning
- **Medicine**
  - AI equivalent to health care professionals
- **When AGI**
  - When will AI systems achieve human-level performance across all tasks?
  - Expert prediction average: 2099
    - Papers show that expert predictions no better than amateurs
    - Experts expected AI to take 100 years to beat humans in Go
  - Unclear if new breakthroughs or refinements needed

Benefits of AI

- Our civilization is the product of human intelligence
    - Greater machine intelligence leads to better human society
    - "*First solve AI, then use AI to solve everything else*"
- **Benefits of AI and robots**
    - Free humanity from menial work
    - Increase production of goods and services
    - Expand human cognition
    - Accelerate scientific research, e.g.,
        - Cures for diseases
        - Solutions for climate change
        - Resource shortages

Risks of AI (1/2)

- **Autonomous weapons**
  - Locate, select, eliminate human targets without intervention
  - Scalability: deploy a large number of weapons
- **Surveillance and persuasion**
  - AI (speech recognition, computer vision, natural language understanding) for mass surveillance
  - Tailoring information flows through social media to modify behavior
- **Biased decision making**
  - Misuse of ML can result in biased decisions due to societal bias
  - E.g., parole evaluations, loan applications

Risks of AI (2/2)

- **Impact on employment**
  - Machines eliminate jobs
  - Rebuttal
    - Machines enhance productivity $\rightarrow$
    - Companies become more profitable $\rightarrow$
    - Higher wages
  - Counter-rebuttal
    - Wealth shifts from labor to capital, increasing inequality
  - Counter-counter-rebuttal
    - Past tech advances (e.g., mechanical looms) disrupted employment, but adaptation followed
- **Safety critical applications**
  - AI in safety-critical applications
    - E.g., self-driving cars, managing water supply or power grids
  - Avoiding fatal accidents is challenging
    - E.g., formal verification and statistical analysis insufficient
  - AI requires technical and ethical standards
- **Cybersecurity**
  - AI defends against cyberattacks
    - E.g., detect unusual behavior patterns
  - AI contributes to malware development
    - E.g., use reinforcement learning for targeted phishing attacks
  - Cat-and-mouse game

Human-Level AI

- **Human-level AI**
  - Machines able to learn to do anything a human can do
  - Aka AGI (Artificial General Intelligence)
- **Artificial Super-Intelligence**
  - Machines surpass human ability in every domain and self-improving
- **Exponential take-off**

The Problem of Control

- Can humans control machines more intelligent than them?
- **King Midas problem**
  - King Midas turned everything he touched into gold, including food and family
  - Humans ask for something, get it, then regret it
  - Rebuttal
    - If AGI arrived in a black box from space, exercise caution before opening
    - We design AI: if AI gains control, it's a "design failure"
- **Problem of alignment**
  - Superintelligent AI might pursue goals in unintended, dangerous ways
- **The paperclip problem**
  - Thought experiment in AI safety (Nick Bostrom, 2003)
  - AI is tasked with maximizing paperclip production
  - AI becomes superintelligent and single-mindedly pursues this goal
  - Converts Earth and humans into paperclips

E/acc vs P(doom)

- **E/acc**
  - Accelerationism is the belief that rapid progress in AI is beneficial or inevitable
  - Solve global problems with more powerful AI tools
  - Belief that slowing AI is unrealistic or counterproductive
- **P(doom)**
  - "Probability of Doom"
  - Estimated probability that advanced AI will cause catastrophic harm
  - Used informally by AI researchers to quantify risk

Solutions to Problem of Control

- **Checks-and-balances**
  - Researchers and corporations develop voluntary self-governance principles for AI
  - Governments and international organizations established advisory bodies
- **Problems**
  - Corporations checking themselves: what can possibly go wrong?
  - Preferences are not easy to invert and are inconsistent
- **Solutions to problem of control**
  - Put purpose into the machine even if objectives are unclear
  - Incentivize AI to switch off if uncertain about human objectives
  - Cooperative Inverse Reinforcement Learning (CIRL)
    - AI observes human behavior to infer reward function

Cooperative Inverse Reinforcement Learning

- AI infers human goals based on actions
- **Observation**: GP looks tired, sits on the couch, observes the messy table, and starts watching TV
- **Inference**: AI infers:
  - GP is tired and wants to relax
  - Messy coffee table bothers him
- **Action**: AI:
  - Fetches a glass of water
  - Tidies up the coffee table without disturbing GP
- **Feedback loop**: AI monitors GP's reactions
  - If GP is relaxed and happy, AI understanding is reinforced
  - If GP is not happy, AI adjusts actions and improves inference