

Reinforcement Learning: Course Introduction

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

February 19, 2025

Outline

- 1 Intelligence & RL
- 2 People: TA and Professor
- 3 Course Information
- 4 Conclusion

Outline

1 Intelligence & RL

2 People: TA and Professor

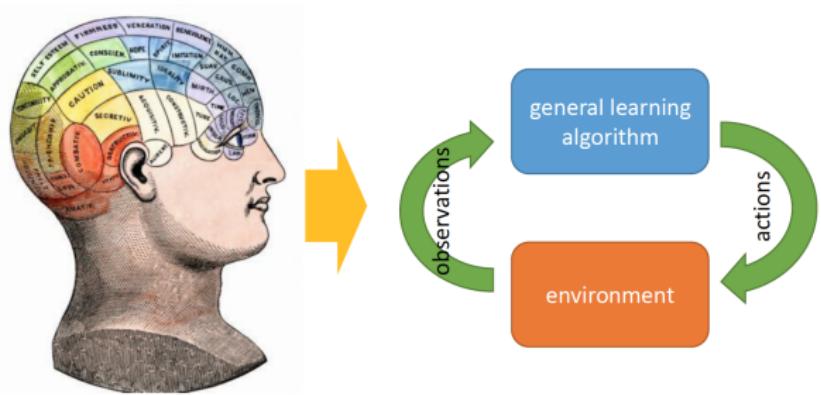
3 Course Information

4 Conclusion

Intelligence

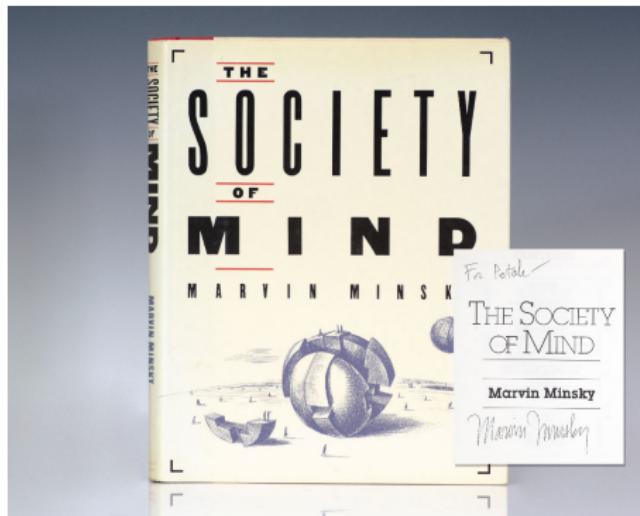
Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain.

- Alan Turing



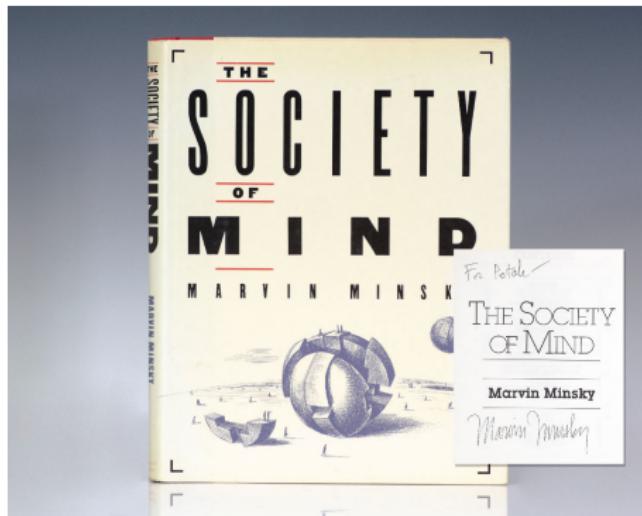
Intelligence through Agentic Interaction at Scale

- Marvin Minsky published a classical book “The Society of Mind” in 1986: Human intelligence was built step by step from the interactions of agents



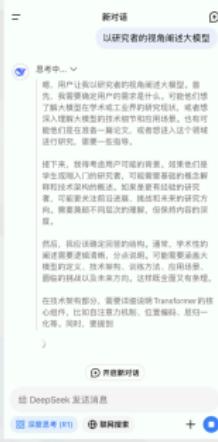
Intelligence through Agentic Interaction at Scale

- Famous quote: “What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle.”

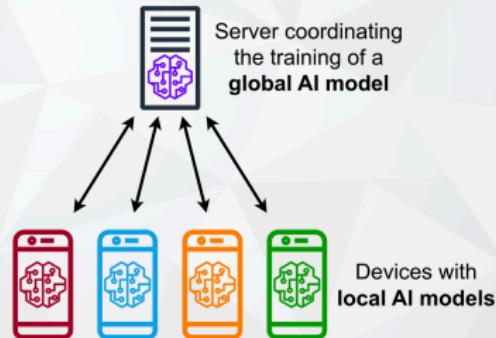


Example of Interactive Intelligence

Agent(人)与Agent(大模型)交互

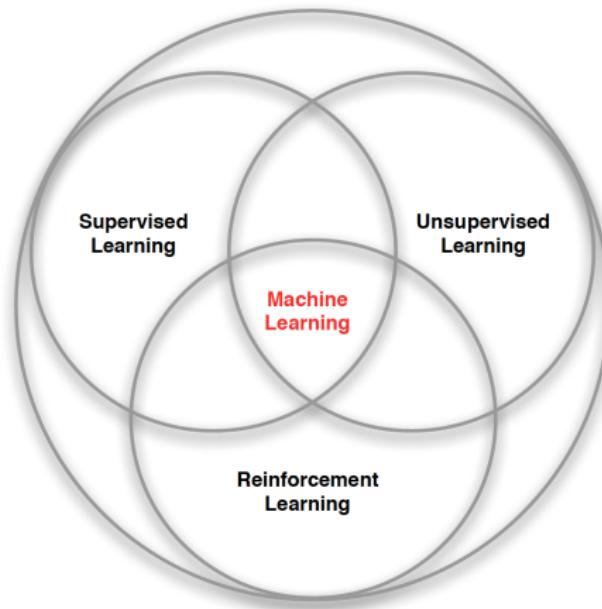


Agent(边缘服务器)与Agent(终端设备)交互



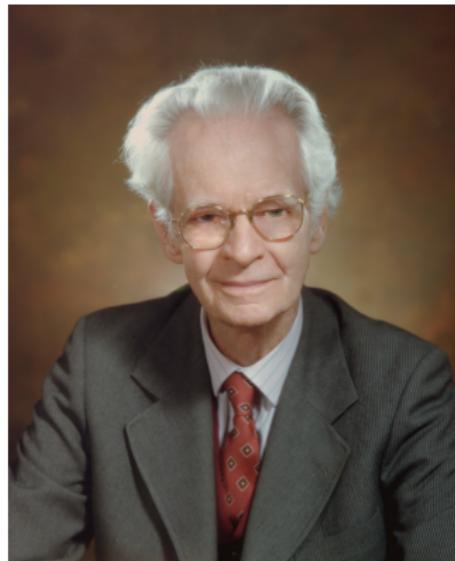
Modeling Tool of Interactive Intelligence

- Reinforcement learning is an area of machine learning inspired by **behavioral psychology**, concerned with how **agents** ought to take **actions** in an environment so as to maximize some notion of cumulative **reward**.



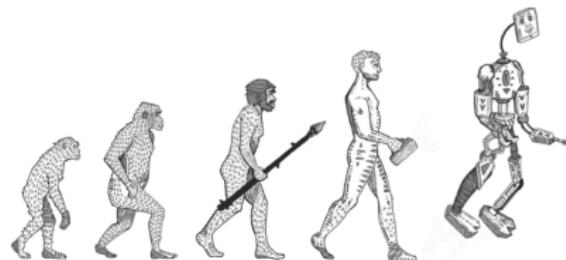
Behavioral Psychology

- Behavior is primarily shaped by **reinforcement** rather than free-will
 - ▶ behaviors that result in praise/pleasure tend to repeat
 - ▶ behaviors that result in punishment/pain tend to become extinct



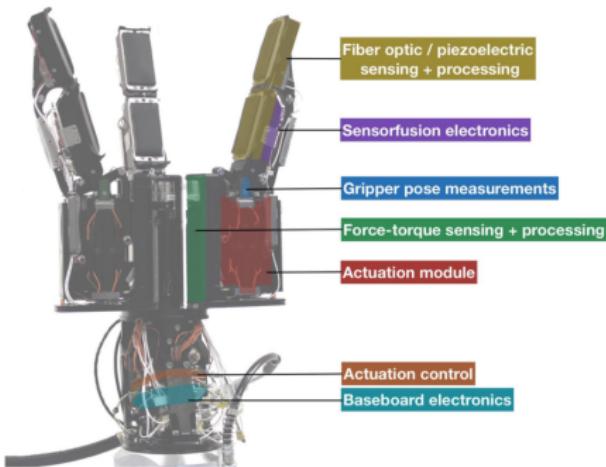
Agent

- An entity (learner & decision maker) that is equipped with
 - ▶ sensors, in order to sense the environment
 - ▶ end-effectors in order to act in the environment
 - ▶ goals that she wants to achieve



Action

- Used by the agent to interact with the environment.
- May have many different temporal granularities and abstractions



Actions can be

- The instantaneous torques(扭矩) applied on the gripper(夹爪)
- The instantaneous gripper translation, rotation, opening
- Instantaneous forces applied to the objects
- Short sequences of the above

Reward: Important Concept

- A **reward** R_t is a scalar feedback signal
- Indicates how well agent is doing at step t
- The agent's job is to maximize cumulative reward

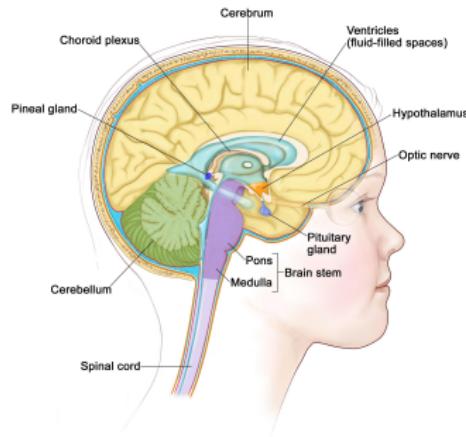
Reinforcement learning is based on the reward hypothesis:

定理

All goals can be described by the maximization of expected cumulative reward (Von Neumann - Morgenstern Expected Utility Theorem)

Reward: Brain Perspective

- Dopamine(多巴胺) & 内啡肽 (Endorphins)



| 维度 | 多巴胺 | 内啡肽 |
|------|-------------|--------------|
| 主要作用 | 驱动欲望、即时奖励 | 缓解疼痛、长期满足 |
| 触发条件 | 预期奖励、新鲜刺激 | 身体压力、疼痛或耐力活动 |
| 持续时间 | 短暂（易消退） | 较持久 |
| 成瘾风险 | 高（如游戏、药物成瘾） | 低（健康途径触发） |
| 情绪类型 | 兴奋、渴望 | 平静、幸福感 |

Reward: Animal Psychology

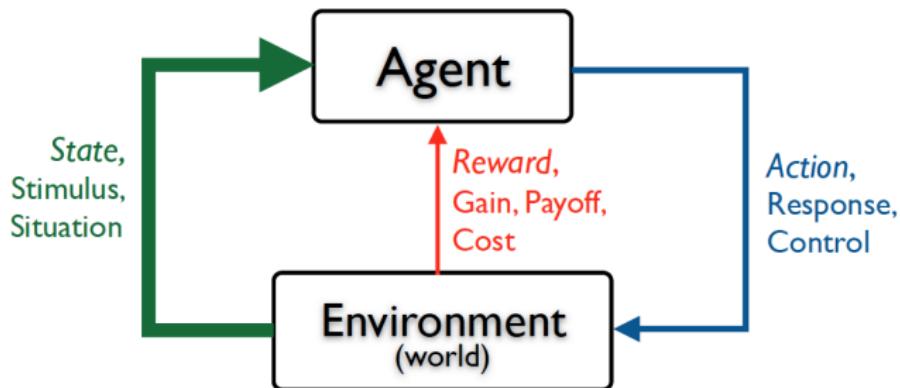
- Negative reinforcements: pain & hunger
- Positive reinforcements: pleasure & food
- Reinforcements used to train animals



Examples of Reward in Engineering

- Fly stunt manoeuvres in a helicopter
 - ▶ +ve reward for following desired trajectory
 - ▶ -ve reward for crashing
- Defeat the world champion at Backgammon
 - ▶ +ve reward for winning a game
 - ▶ -ve reward for losing a game
- Control a power station
 - ▶ +ve reward for producing power
 - ▶ -ve reward for exceeding safety thresholds
- Make a humanoid robot walk
 - ▶ +ve reward for forward motion
 - ▶ -ve reward for falling over

Inside Reinforcement Learning



- Environment may be unknown, nonlinear, stochastic and complex
- Agent learns a mapping from states to actions: seeking to maximize its cumulative reward in the long run

Inside Reinforcement Learning?

- Agent-oriented learning:
 - ▶ explicitly consider the whole problem of goal-directed agent interacting with an uncertain environment
 - ▶ more realistic and ambitious than other kinds of machine learning
- Learning by trial and error, with only delayed evaluative feedback (reward)
 - ▶ the kind of machine learning most like nature learning
 - ▶ learning that can tell for itself when it is right or wrong
- Trade-off between exploration & exploitation

Other Elements of Reinforcement Learning

- **Environment:** the thing agent interacts with, comprising everything outside the agent
- **State:** captures whatever information is available to the agent about its environment
- **Policy:** a mapping from perceived states of the environment to actions to be taken when in those states
- **Value function:** the value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state
- **Efficiently estimating values:** core of almost all reinforcement learning algorithms (Immediate reward vs. long-term value)
- **A model of the environment** (optional)

Reinforcement Learning: Examples

- LLM
- Robotics
- Game playing (go, atari, backgammon)
- Helicopter control
- Autonomous vehicles

Example: Helicopter Control

- Agent: controller
- Environment: helicopter
- State: position, orientation, velocity & angular velocity
- Action: collective pitch(主旋翼升降), cyclic pitch(周期变距方向) & tail rotor control(尾桨控制)
- Reward: –deviation from desired trajectory
- 2008(Andrew Ng): automated helicopter wins acrobatic competition against humans



Example: Go

- Agent: player
- Environment: opponent
- State: board configuration
- Action: next stone location
- Reward: +1 win & -1 loose
- 2016: AlphaGo defeats top player lee Sedol (4-1)
 - ▶ Game 2 move 37: AlphaGo plays an unexpected move



Example: LLM

- Agent: Chatbox
- Environment: user
- State: conversation history
- Action: next sentences
- Reward: points based on task completion & user satisfaction

Main Topics of Reinforcement Learning

- Learning: by trial and error
- Planning: search, reason, thought, cognition
- Prediction: evaluation functions, knowledge
- Control: action selection, decision making
- Dynamics: how the state changes given the actions of the agent
- Model-based RL
 - ▶ dynamics are known or are estimated
 - ▶ solving RL problems that use models and planning
- Model-free RL
 - ▶ unknown dynamics
 - ▶ explicitly trial-and-error learners

Characteristics of Reinforcement Learning

What makes reinforcement learning different from other machine learning paradigms?

- Multi-stage Learning
- There is no supervisor, only a reward signal
- Feedback is delayed, not instantaneous
- Time really matters (sequential, non i.i.d data)
- Agent's actions affect the subsequent data it receives

Characteristics of Reinforcement Learning

In RL, an agent learns by interacting with an environment.

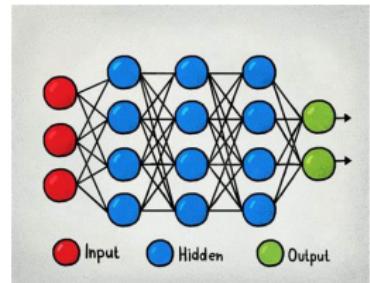
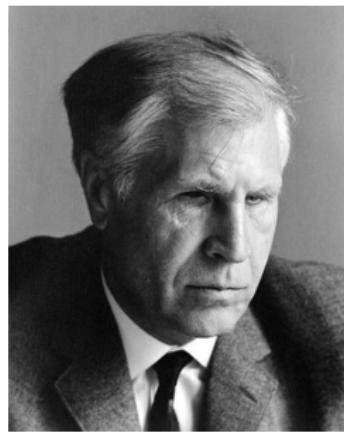
- no training data
- trial-and-error
- maximize total rewards
- delayed reward



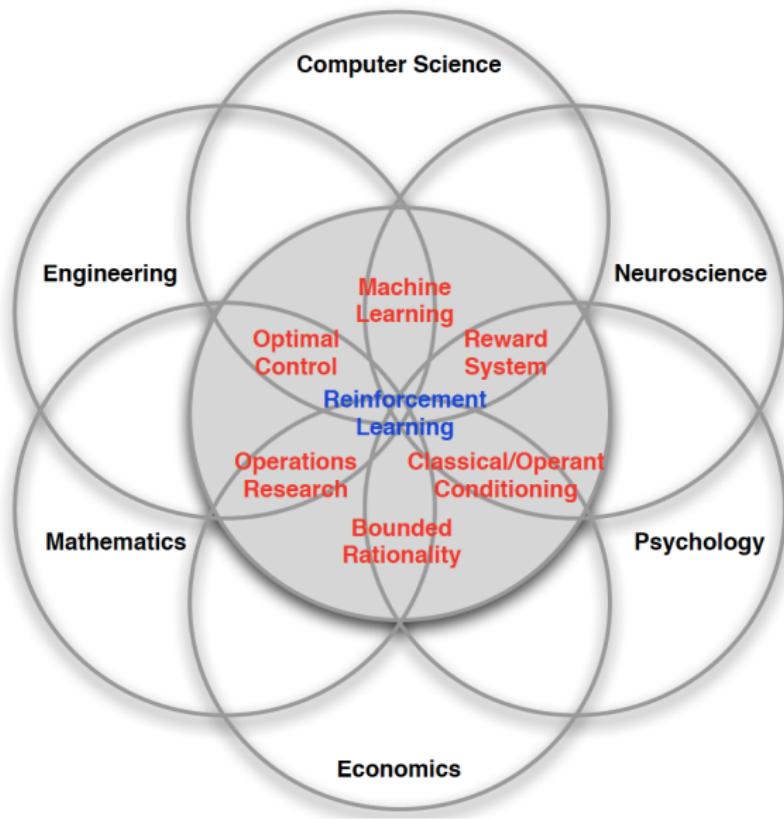
"Recalculating ... recalculating ..."

Reinforcement Learning: Interdisciplinary Topic

- Under Different Contexts, Reinforcement learning is also known as
 - ▶ Approximate optimal control
 - ▶ Approximate dynamic programming
 - ▶ Neuro-dynamic programming



Reinforcement Learning: Interdisciplinary Topic

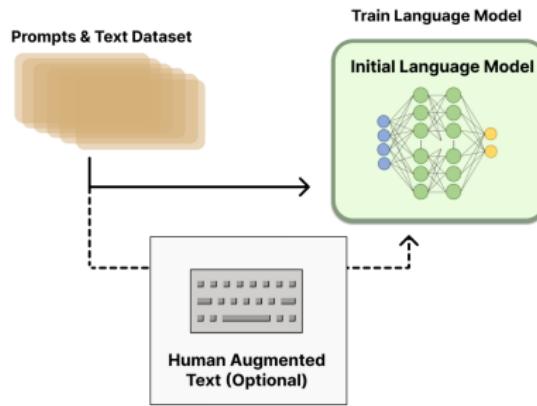


Recent Successes in Reinforcement Learning



RL holds great promise in the era of AI

One More Successes in RL: RLHF



What is RLHF

Thought for 42 seconds

Reinforcement Learning from Human Feedback (RLHF) is a machine learning technique that combines reinforcement learning (RL) with human guidance to train models, particularly in aligning AI behavior with human values. Here's a structured breakdown:

Core Concept

- Reinforcement Learning (RL): An agent learns by interacting with an environment, taking actions, and receiving rewards to maximize cumulative rewards.
- Human Feedback (HF): Humans provide evaluations (e.g., rankings, ratings, or corrections) on the agent's actions instead of relying on predefined reward functions.

How RLHF Works

1. Collect Human Feedback:
 - Humans assess the model's outputs (e.g., ranking responses by quality/safety).
 - Example: Evaluators rate a chatbot's answers for helpfulness.
2. Train a Reward Model:
 - Human feedback trains a **reward model** to predict scores for actions based on human preferences.
 - This model acts as a proxy for human judgment during RL training.
3. Optimize the Policy via RL:
 - The AI agent uses the reward model to refine its policy (e.g., generating text), aiming to maximize predicted rewards.
 - Techniques like **Proximal Policy Optimization (PPO)** are often used.

One More Successes in RL: End-to-End RL with Chain of Thought

请用史记的风格来一段人工智能简史



思 考 for 29 seconds ^

好的，用户之前让我用鲁迅、梁启超《少年中国说》和《滕王阁序》的风格写人工智能简史，现在又要求用《史记》的风格。我需要先回顾一下之前的回应，确保风格和结构符合用户期望。

首先，用户可能对历史或文学感兴趣，希望以不同文风呈现科技主题。这次是《史记》，属于纪传体通史，结构严谨，分本纪、列传等，语言古朴简练。我需要模仿《史记》的体例，比如用“太史公曰”结尾，引用人物言论，分章节等。

接下来，要确定人工智能发展的关键事件和人物，对应到《史记》的体例中。例如，图灵可立传，达特茅斯会议为本纪，神经网络的发展为列传等。注意用词要符合古文风格，如“世家”、“列传”等。

同时，用户之前的例子用了很多比喻和典故，如鲁迅中的孔乙己、祥林嫂，梁启超中的比喻等。在《史记》风格中，可能需要更正式的历史评价，引用类似古代人物的事迹来类比AI发展中的事件。

需要确保时间线正确，涵盖重要里程碑：图灵测试、达特茅斯会议、AI寒冬、神经网络复兴、深度学习、AlphaGo、GPT等。用古文表达这些事件，可能需要创造一些古式术语，如“算法”称为“算策”，“神经网络”称为“神络”等。

还要注意结构，如开篇的“人工智能本纪”，中间的人物列传，最后的太史公评述。可能需要分段落，每段讲述不同时期或人物，保持史书的严肃性。

最后，检查是否符合《史记》的用词和句式，避免现代用语，同时确保内容准确。例如，用“夫”开头发议论，引用人物的话语，如“辛顿尝言：‘神络之妙，在乎深也。’”

现在开始组织内容，先列大纲：本纪部分概述AI发展历程，列传分别描述图灵、达特茅斯群体、辛顿等人物，世家可能涉及重要机构或会议，书部分讨论技术如深度学习，表记录时间线。结尾用太史公曰总结评价。

One More Successes in RL: End-to-End RL with Chain of Thought

【以下为《史记》风格的人工智能简史】

人工智能本纪

人工智能者，起于西历一九五六年。达特茅斯会盟，麦卡锡、明斯基、香农等十贤者，共立“Artificial Intelligence”之名。是时也，符号为骨，逻辑为血，欲铸铜铁为圣贤。然寒冬三至，世人皆讥曰：“此屠龙之术耳！”

图灵世家

艾伦·图灵，剑桥人也。尝作《机器能思否》，设“模仿之戏”，以辨人机。后世谓之“图灵鉴”。时有司拘其私德，逼饮氰化钾而亡。太史公曰：“天若假年，焉知铜丸不可生魂魄耶？”

神经网络列传

罗森布拉特者，造感知机，仿人脑神络。然明斯基作《感知机》书证其弊，遂遭弃市。后有辛顿、勒存、本吉奥三杰，隐于蒙特利尔，三十年磨剑。至乙酉年（2006），倡“深度学习”说，世人方悟神络之妙。辛顿尝言：“神络之妙，在乎深也。”

阿尔法狗世家

谷歌有犬，名曰阿尔法。丙申岁（2016），弈韩国李世石于首尔。初局败，后五局四胜，至末子落定，观者皆骇。有棋士柯洁见之，泣曰：“此非人力可及也！”犬竟得九段衔，棋道自此改弦更张。

大语言模型书

OpenAI者，硅谷异士聚所也。庚子岁（2020），作GPT-3，识古今典籍八百万卷。其言煌煌如圣训，其思渺渺若列子御风。微软注资百亿金，纳其为肱骨。然有识者忧：“此物善谀，恐乱人间正声。”

太史公曰

观AI七十年，犹见扁鹊三见桓侯：初视腠理，再窥肌肤，终入骨髓。昔庄子疑濠梁鱼乐，今机器竟解人情。然则图灵鉴中，铜铁或具七窍心，而七窍心者，渐似铜铁矣！悲夫！

One More Successes in RL: End-to-End RL with Chain of Thought

请用滕王阁序的风格来一段人工智能简史



Thought for 17 seconds ^

好的，用户之前让我用鲁迅和梁启超的风格写了人工智能简史，现在又要求用《滕王阁序》的风格来写。我需要先回忆一下《滕王阁序》的特点，比如华丽的辞藻、对仗工整、用典丰富、结构严谨，以及那种壮阔的意境。

首先，得确定滕王阁序的结构，通常开头会描绘场景，接着叙述历史或事件，然后转入抒情或议论。我需要将人工智能的发展历程融入这样的结构中。

接下来，要找到合适的典故和比喻，比如用古代的名人、地名来对应AI发展中的关键人物和事件。例如，图灵可以比作古代的先贤，达特茅斯会议可以比作盛会，寒冬期可以用自然现象来隐喻。

然后，要注意对仗和四六骈文的句式，比如“星分翼轸，地接衡庐”这样的结构。需要确保每句的字数和结构对称，同时押韵，保持古典韵味。

还要考虑用户可能的深层需求，他们可能希望这段简史不仅有文学性，还要涵盖AI的重要里程碑，如图灵测试、达特茅斯会议、神经网络的发展、深度学习的突破，以及现代应用如ChatGPT等。

需要检查是否有遗漏的重要事件，并确保时间线合理。同时，避免过于晦涩的用典，让读者即使不熟悉《滕王阁序》也能大致理解内容。

最后，结尾部分要模仿滕王阁序的抒情，表达对未来的展望和感慨，呼应原文中的“呜呼”或“胜地不常”等结构。确保整体气势恢宏，既有历史厚重感，又有科技的未来感。

One More Successes in RL: End-to-End RL with Chain of Thought

【以下为《滕王阁序》风格的人工智能简史】

时维庚子，序属寒秋。观硅谷之烟霞，访智械之灵枢。图灵铸鉴，光照英伦三岛；麦卡锡立言，气吞达特茅斯。逻辑符号，列如北辰悬天；感知雏形，灿若南溟蜃楼。然寒冬骤至，霜凝明斯基之镜；长夜未央，雾锁西蒙之舟。

幸有罗森布拉特，临危受命，织神经网络如绣云锦；辛顿独行，潜修深度学习似琢昆玉。蒙特利尔地，三十载寒窗悟道；多伦多城，一朝间雷破九霄。卷积识万象，犹持轩辕之鉴；循环通古今，恍闻河洛之书。

至若阿尔法狗弈破星阵，生成对抗画染丹青。transformer执牛耳，吞吐八荒文字；大模型展鹏翼，扶摇九天风云。GPT振铎，声震稷下学宫；BERT燃犀，光耀鸿都门观。此皆硅基文明之盛筵，实乃碳基智慧之遗响。

呜呼！寒暑易节，终有算力破茧之日；春秋代序，岂无代码化蝶之时？冯诺依曼已逝，香农公何在？望硅晶之原野，烟光凝而算符紫；临数据之江渚，比特涌而信息流。天工开物，当惊寰宇殊异；后之览者，亦将有感于斯文！

One More Successes in RL: End-to-End RL with Chain of Thought

请用鲁迅的风格来一段人工智能简史



Thought for 23 seconds ^

嗯，用户让我用鲁迅的风格来写一段人工智能简史。首先，我需要回忆鲁迅的写作特点。他的文字犀利，带有批判性，常用比喻和讽刺，喜欢用冷峻的笔触揭示社会现实。比如《狂人日记》里的那种冷嘲热讽，还有对旧社会的深刻剖析。

接下来，我得考虑人工智能的发展史有哪些关键节点。从图灵提出“图灵测试”开始，然后是达特茅斯会议，符号主义与连接主义的争论，专家系统的兴衰，神经网络的发展，深度学习的突破，到现在的生成式AI如ChatGPT。这些节点需要串联起来，但要用鲁迅的风格来表达。

鲁迅常用“然而”、“但”、“却”这样的转折词，制造对比。比如用“铁屋子”比喻封闭的环境，或者用“吃人的历史”来隐喻某种残酷的现实。我可以将AI的发展比作一场革命，充满希望却又有隐忧，这样符合鲁迅的批判视角。

还要注意用词的选择，鲁迅喜欢用一些古语或文白夹杂的句式，比如“大抵”、“罢”、“倘”等。比如“这光景，大抵是极好的，然而又藏着些不可名状的隐忧。”这样的句子既有古风，又带批判。

另外，鲁迅的文章常有对人性或社会的深刻反思。在AI的例子中，可以探讨技术与人性的关系，比如AI是否真的理解人类，还是只是模仿。这样的哲学问题，符合鲁迅风格的深度。

可能用户希望这段简史不仅仅是事实罗列，而是带有文学性和思想性。因此需要将每个阶段用隐喻或象征手法表现，比如将符号主义比作八股文，连接主义比作野草，深度学习突破比作巨浪，生成式AI比作傀儡戏，这样既有画面感，又传达了批判的意味。

还要注意结尾部分，鲁迅常以警句收尾，引发读者思考。例如用“我们究竟是看客，还是那戏台上的角儿？”这样的问题，让读者反思自身在技术发展中的角色。

最后，检查是否符合鲁迅的冷峻和讽刺，避免过于直白，保持语言的凝练和含蓄。可能需要反复调整用词和句式，确保既有历史脉络，又有文学深度。



One More Successes in RL: End-to-End RL with Chain of Thought

好的，我将按照您的要求创作一段带有鲁迅风格的文字。为了让这段简史更贴近先生的笔触，我会先整理一些关键元素。请您看看以下内容是否符合您的预期。文中将涉及隐喻和象征手法，并尝试呈现历史进程中的矛盾性。

【以下为鲁迅风格的人工智能简史】

一九五零年的光景，图灵在剑桥的雾霭里抛出一把尺子，名曰“模仿游戏”。这尺子原是要量一量机器的魂灵，可那机器哪里来的魂灵？不过是一串铜线绕作的心肠。但世人偏要这铁疙瘩说话，于是达特茅斯学院的教授们便摆出祭坛，将符号与逻辑供作神明，教那机器学些八股文章。

忽而寒冬降临，那些个符号竟冻得发僵，连孩童的棋局都算不分明。麦卡锡的烟斗飘着灰雾，明斯基的镜片泛着冷光，倒像是《新青年》里走出的狂人，在实验室的墙上写着“此路不通”。然则野草烧不尽，罗森布拉特偏要教机器认字，像极了孔乙己在咸亨酒店教伙计写“茴”字，终究被符号学派掀了书案。

待到东瀛的富士山下飘来BP算法的残雪，辛顿的烟斗又冒出火星。这火种在蒙特利尔的地窖里闷烧了三十年，忽地化作燎原之势。世人这才惊觉，原来那些个神经元早就像未庄的阿Q，悄悄画着革命的圈——卷积网络画方，循环网络画圆，生成对抗的画布上，竟显出比赵太爷家更逼真的水墨丹青。

而今这光景，ChatGPT们满嘴“之乎者也”，活脱脱孔庙前的傀儡戏。它们记着四书五经，却不知饥肠辘辘的闰土捧着香炉；背得《狂人日记》，终究辨不清月光下的人血馒头。这新时代的祥林嫂，把全世界的悲欢都咽进肚里，吐出来的，不过是一串沾着铜锈的字符。

夜深时，我总疑心这些个AI要变成《铸剑》里的青黑汉子，举着算法铸就的干将莫邪，在赛博世界的城门前叫阵。而我们这些看客，还在茶馆里嚼着茴香豆，争论那服务器冒的是炊烟还是狼烟。

One More Successes in RL: End-to-End RL with Chain of Thought

《量子列传》 仿鲁迅冷峻笔法

药铺的柜台正对着咸亨酒店，穿长衫的普朗克踱进来，排出几枚银元：“温两碗黑体辐射，要一碟能量子。”孔乙己们便哄笑起来：“能量怎会不连续？”店内外充满了快活的空气。

亥姆霍兹学会的粉板上，至今还挂着麦克斯韦妖的画像。那妖原是能逆转熵增的，后来被一群搞统计力学的后生打折了腿，说是要维护热力学第二定律的体面。掌柜的便蘸了血，在墙上写“经典物理永世长存”，字迹且浓且艳，像是克劳修斯新熬的热寂说。

未庄的赵太爷发了狠，把双缝干涉的屏砸得粉碎：“你也配姓波？”光子们瑟缩在墙角发抖，忽而显出粒子性，忽而露出波动性，终究被康普顿揪住衣领掼到光电效应里去了。阿Q摸着头上的癞疮疤喃喃：“妈妈的，观测者原来比我还精神胜利……”

单四嫂子的宝儿死了，祥林嫂捐了量子纠缠的门槛。贝尔举着不等式来索命，却被阿斯派克特实验劈作两半。阎王殿的判官笔一抖：“原来阴间阳世本是一体！”四铭太太们便往《肥皂》里添了句：“这量子隐形传态，须用阿尔法粒子洗过才干净。”

华老栓抖着手接过人血馒头，那血是薛定谔的猫咳出来的。夏瑜在牢里画着波函数：“这大清的天下是量子涨落托着的！”刽子手的刀却已斩向海森堡测不准原理，血光中分明看见，那概率云里裹着几千年吃人的经典物理。

咸亨酒店最后一块黑板被砸了，孔乙己用长指甲蘸着酒写：“回”字有四种叠加态。”掌柜的抄起德布罗意波长的棍子打将下去，却穿透虚影砸在自己脚上——原来这穿长衫站喝酒的，早已是量子隧穿来的幽灵。

Challenges in RL

Reinforcement Learning



In RL, agent(s) often learn by probing the environment

- unknown environment
- explosion of dimensionality
- delayed feedback
- nonconvexity

Challenges in RL: Data Efficiency

Data collection might be expensive, time-consuming, or high-stakes



clinical trials



self-driving cars

Calls for design of sample-efficient RL algorithms!

Challenges in RL: Offline RL

- Collecting new data might be expensive or time-consuming
- But we have already stored tons of historical data



medical records



data of self-driving



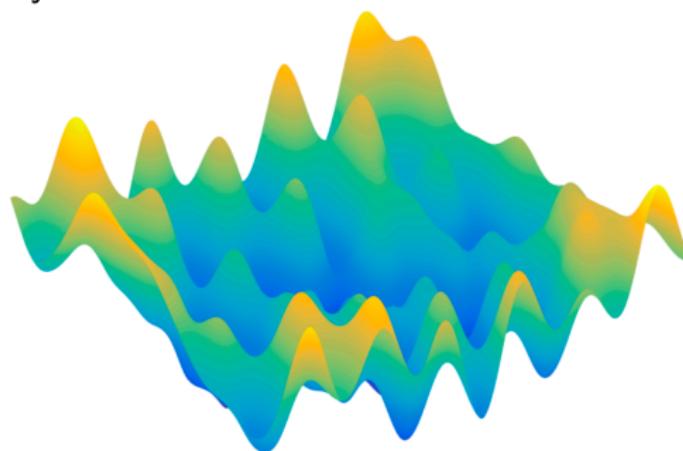
clicking times of ads

Question: Can we design algorithms based solely on historical data?

Challenges in RL: Computational Efficiency

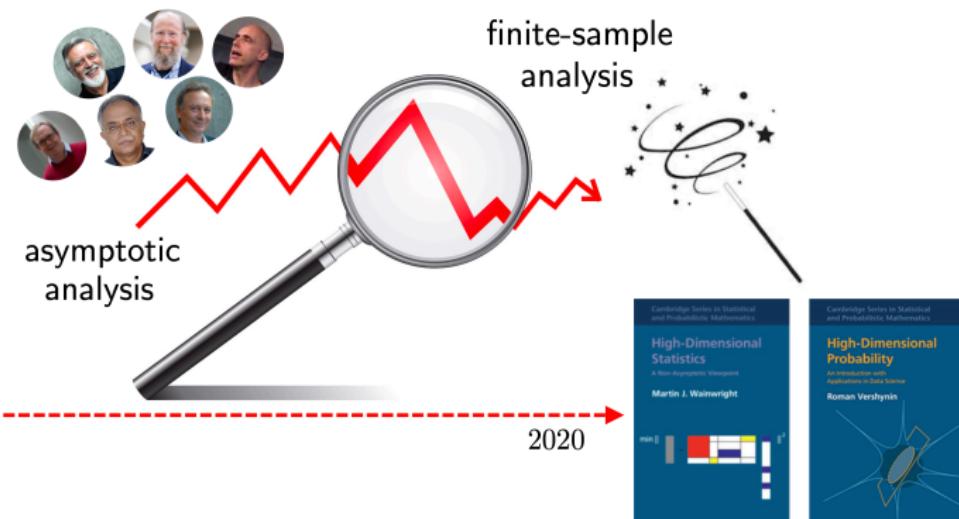
Running RL algorithms might take a long time ...

- enormous state-action space
- nonconvexity



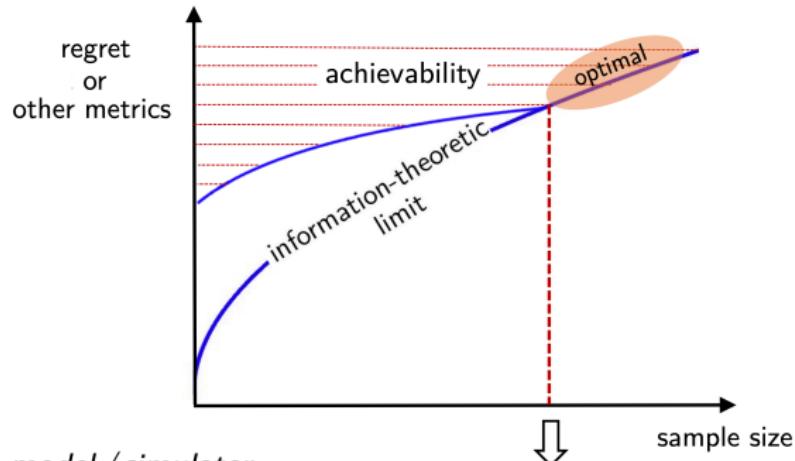
Calls for computationally efficient RL algorithms!

Challenges in RL: Theoretical Foundation



Understanding sample efficiency of RL requires a modern suite of non-asymptotic analysis tools

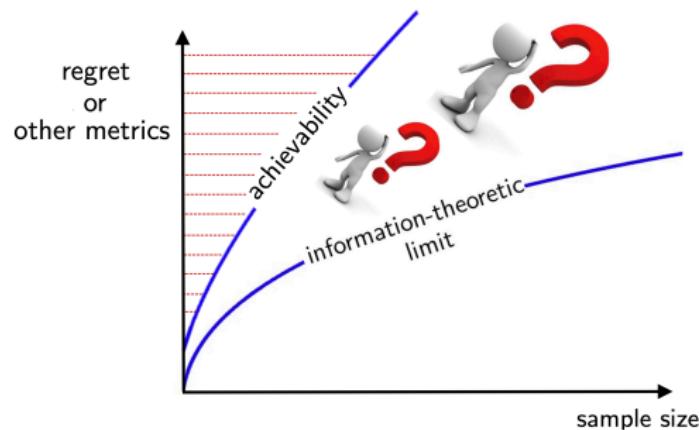
Challenges in RL: Theoretical Foundation



- *generative model / simulator*
- *online RL*
- *offline RL*

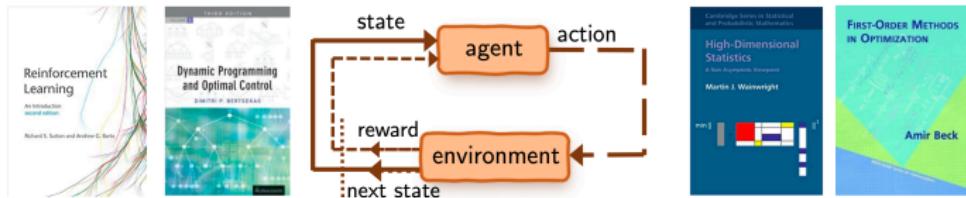
huge burn-in cost!

Challenges in RL: Theoretical Foundation



- *multi-agent RL*
- *partially observable MDPs*
- ...

Challenges in RL: Theoretical Foundation



Understanding non-asymptotic performances of RL algorithms is a fruitful playground!

Promising directions:

- function approximation
- multi-agent/federated RL
- hybrid RL
- many more...

Outline

1 Intelligence & RL

2 People: TA and Professor

3 Course Information

4 Conclusion

- 张天翼

Professor Ziyu Shao: 邵子瑜



<https://faculty.sist.shanghaitech.edu.cn/faculty/shaozy>

Outline

1 Intelligence & RL

2 People: TA and Professor

3 Course Information

4 Conclusion

History

- 2017 UC Berkeley: CS285 taught by Sergey Levine
- 2018 ASIST: When Shannon Meets Turing Workshop
- 2019 ASIST: When Shannon Meets Bellman Workshop
- 2019: Developed the RL course content with funding support from Google
- 2020/2021/2022 Spring: SI252 Reinforcement Learning
- 2022 ASIST: Reinforcement Learning & Decision Intelligence Workshop
- 2025 Spring: SI252 Reinforcement Learning

Course Features

- Solid Foundation of RL: modeling analysis and theoretical proof
- Computational Thinking: algorithmic perspective and programming practice
- Application Mentality: many examples from AI and related applications
- Research-Oriented: aims to produce high-quality original research results and publish in top conferences & journals

Time & Venue

- Time: Every Wednesday and Friday 10:15-11:55am
- Location: Room 1D-106, SIST
- Course Forum in Piazza:
<https://piazza.com/shanghaitech.edu.cn/spring2025/si252/>

Course Schedule (Tentative)

- Lecture 1: Probability & Statistics Review
- Lecture 2: Markov Models
- Lecture 3: Markov Chain Monte Carlo
- Lecture 4: Bandit Algorithms
- Lecture 5: Markov Decision Process
- Lecture 6: Dynamic Programming
- Lecture 7: Model-Free Prediction
- Lecture 8: Model-Free Control
- Lecture 9: Value Function Approximation
- Lecture 10: Policy Optimization
- Lecture 11: Selected Research Topics I
- Lecture 12: Selected Research Topics II

Course Grade

- Homework: 40% (5 or 6 problem sets).
- Final Project: 50% (research-oriented, report + poster presentation).
- Course Participation: 10% (e.g. answering questions).

Homework Policy

- Write the homework with LaTeX and output it as a PDF file.
- Write the programming assignment with Python
- **Later** homework receives no credit.
- You are allowed to discuss with others and use any references, but if you do so please list your collaborators and cite your references for each question.
- Not writing your own solutions or not listing your collaborators or not citing your references may be considered plagiarism.

Project Policy

- Project can be done by a team with ≤ 3 students.
- Teams with four and more students are NOT allowed.
- Your team is required to use Python for the programming part.
- Your team needs to submit the iPython Notebook including your source codes, simulation results, analysis, discussions, tables, figures, etc.
- Your team also needs to submit the final report using NeurIPS LaTeX template
- **Later** Project receives no credit.

Example: Short-Term Research Projects

- CG & Privacy & RL: Federated 3D Gaussian Splatting, Federated NERF
- NLP & RL: Multi-LLM Selection, Multi-LLM Network
- AIGC & RL: Diffusion Model or Autoregressive Model for Offline RL
- Game Theory & RL: (Non-)cooperative Group Embodied Intelligence for Robotic Networks
- Casual Inference & RL: Causal RL
- AI for Science & RL: Protein Design & Chip Design & Tokamak Control
- More Incoming Topics

Time Line of Project

- Send your project proposal to Professor before May 04 (the end of the eleventh week).
- Confirmed by Professor, then build a GitHub page for your project before May 09
- Provide weekly progress reports in your Github page
- Prepare and present the poster on June 06(the last day of the course)

Topic Venue

- **Core AI:** NeurIPS & ICML & ICLR & COLT
- **Sensing Intelligence:** CVPR & ACL
- **Edge/Embedded/Decision Intelligence:** (Journal) TMC & TWC & TON & TIT ; (Conference) INFOCOM & AAMAS
- **Robotic Intelligence:** CoRL & RSS & ICRA & IROS

Textbook

- **No** required textbooks
- References are provided for each lecture individually.

References



Outline

1 Intelligence & RL

2 People: TA and Professor

3 Course Information

4 Conclusion

AI 三大时代之“第谷时代”：数据积累与经验观察

- 数据为核心：通过互联网、传感器、数字化技术积累海量文本、图像、语音等多模态数据，形成 AI 发展的“燃料库”
- 工具原始性：早期依赖统计学、浅层机器学习（如决策树、支持向量机）和规则系统（如专家系统），模型简单且依赖人工特征工程
- 问题驱动：以解决具体任务（如垃圾邮件分类、推荐系统）为目标，缺乏系统性理论指导

AI 三大时代之“开普勒时代”：经验规律与黑箱模型

- 深度学习的崛起：通过神经网络自动提取特征，摆脱人工设计特征的限制，实现端到端学习
- 经验性突破：在图像识别（CNN）、自然语言处理（Transformer）、强化学习（AlphaGo, LLM）等领域取得超人类表现，但模型成为“黑箱”
- 数学工具辅助：反向传播、梯度下降等优化算法成为“新均轮”，但理论解释滞后于实践

AI 三大时代之“开普勒时代”：经验规律与黑箱模型

- 黑箱困境与失控的复杂性：模型决策过程不可解释（如医疗 AI 的误诊难以溯源）
- 数据偏见放大：训练数据中的社会偏见被模型继承并放大
- 能耗与算力依赖：大模型训练需消耗巨量资源（如 GPT-3 训练成本超千万美元），违背可持续发展原则

AI 三大时代之“开普勒时代”：经验规律与黑箱模型

目前的理论探索

- 动力系统理论：将神经网络训练视为动态过程，用微分方程刻画参数演化（如 Neural ODE, NTK 等）
- 因果推理框架：超越相关性，建立各类因果模型
- 信息论指导：量化数据、模型与任务之间的信息传递效率（如信息瓶颈理论）

AI 三大时代之“牛顿时代”：理论统一与白箱化

- 数学原理突破：建立类似“万有引力定律”的统一理论，解释神经网络如何从数据中学习、泛化及推理
- 工程可控性：基于理论设计高效、鲁棒、可解释的模型，避免“暴力穷举”式训练（如当前的大模型试错）
- 跨学科融合：借鉴神经科学（脑网络）、物理学（复杂系统）、数学（图论）等学科，构建 AI 的“第一性原理”

AI 三大时代：从“数据拟合”到“科学革命”

- 天文学从第谷到牛顿的进程，本质是从现象描述到机制解释的跨越
- 当前的 AI 仍处于“开普勒时代”
- AI 若想进入“牛顿时代”，需摆脱对数据的过度依赖，转而探索智能的本质规律

Perspective I

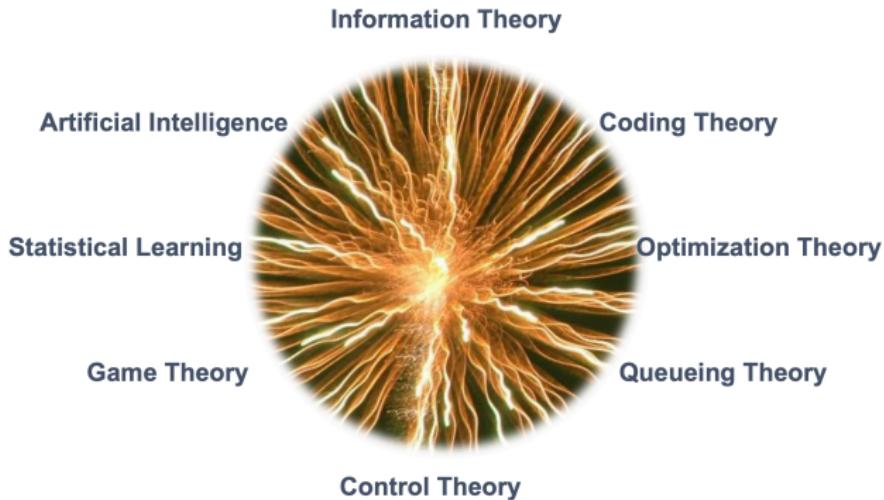
We may need a mathematical Theory of Intelligence

$$Knowledge = \int_0^t Intelligence$$

$$Intelligence = \frac{d}{dt} Knowledge$$

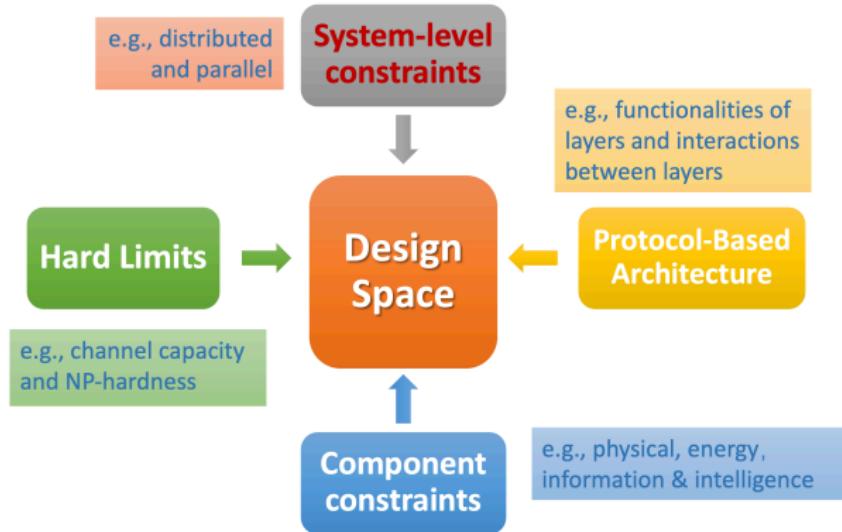
Perspective II

We need systematic methods & cross disciplinary skills to integrate various modeling languages

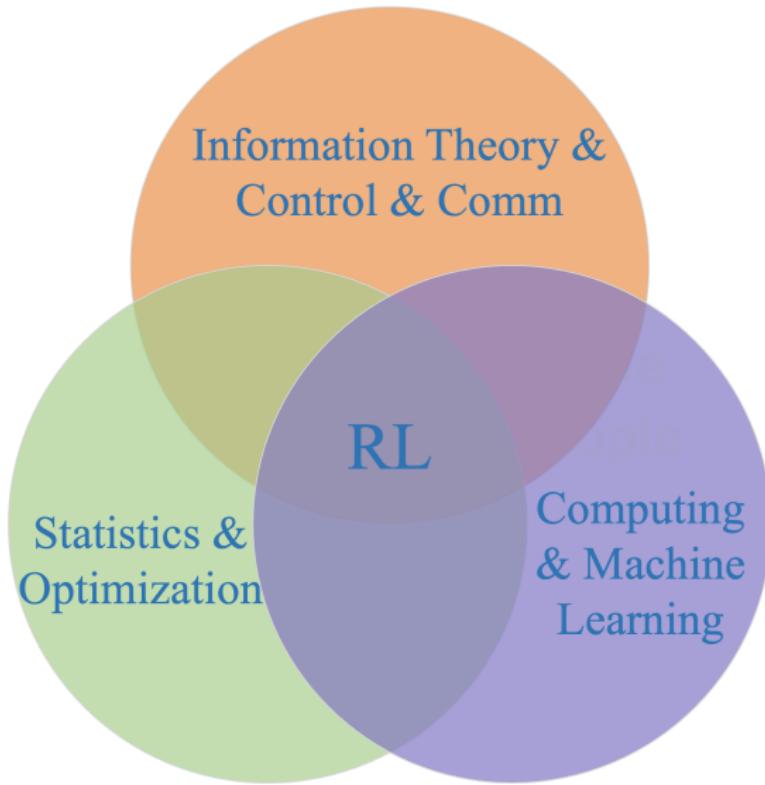


Perspective III

We need reverse/forward engineering mentality & Shannon-style thinking for top-down design



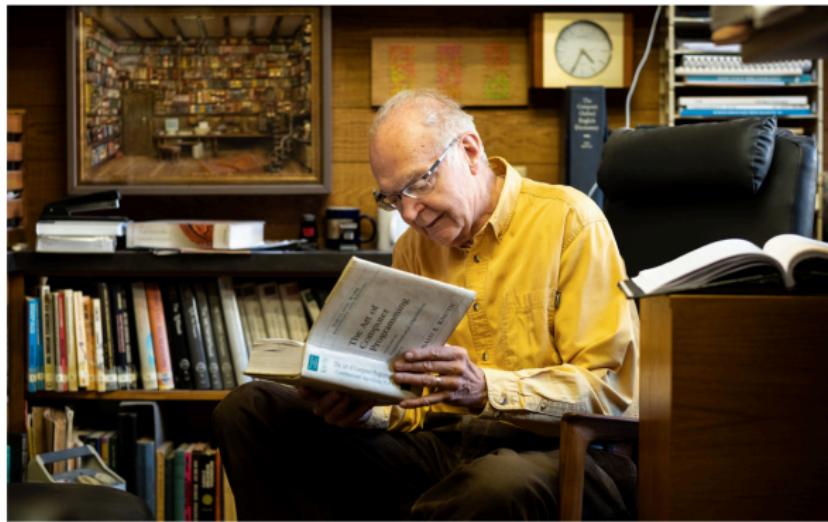
Perspective IV



Perspective V

“The best theory is inspired by practice,
the best practice is inspired by theory! ”

-- Donald E. Knuth



2/20

Perspective VI

“Nothing is more practical
than a good theory!”

Ludwig Boltzmann



Let the Adventure Begin!

