DSC190: Machine Learning with Few Labels

Reinforcement Learning

Zhiting Hu

Lecture 24, November 25, 2024

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

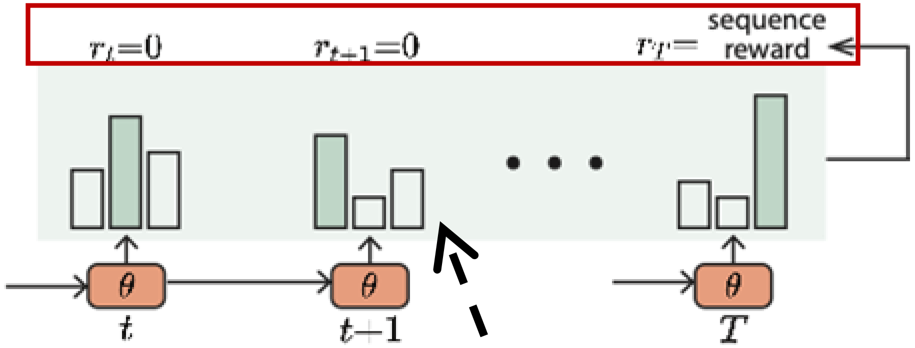
Outline

Reinforcement learning

Presentations

- **Mia Jerphagnon:** Feature Selection Strategies: A Comparative Analysis of SHAP-Value and Importance-Based Methods
- **Tongxun Hu:** Enhancing Sentiment Analysis of FOMC Minutes Using FinBERT-FOMC with Sentiment Focus
- **Yuru Feng:** Large Language Models as Commonsense Knowledge for Large-Scale Task Planning
- **Shentong Li:** ChatGPT Based Data Augmentation for Improved Parameter-Efficient Debiasing of LLMs
- **Evelyn Huang:** Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach
- **Aleck Wu:** UMAP: Uniform Manifold Approximation for Dimension Reduction

Recap: RL for LLM



- (Autoregressive) text generation model:

Sentence $\boldsymbol{y} = (y_0, \dots, y_T)$ $\pi_\theta(y_t \mid \boldsymbol{y}_{<t}) = \text{softmax}(\ f_\theta(y_t|\boldsymbol{y}_{<t})\ )$

logits

In RL terms: trajectory, τ action, a_t state, $\boldsymbol{s}_t$ policy $\pi_\theta(a_t \mid \boldsymbol{s}_t\ )$

- Reward $r_t = r(\boldsymbol{s}_t, a_t)$
  - Often **sparse**: $r_t = 0$ for $t < T$
- The general RL objective: maximize cumulative reward $J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$
- Q-function: expected *future* reward of taking action a_t in state $\boldsymbol{s}_t$

$$Q^\pi(\boldsymbol{s}_t, a_t) = \mathbb{E}_\pi\left[\ \sum_{t'=t}^{T} \gamma^{t'}\ r_{t'} \mid \ \boldsymbol{s}_t, a_t\ \right]$$

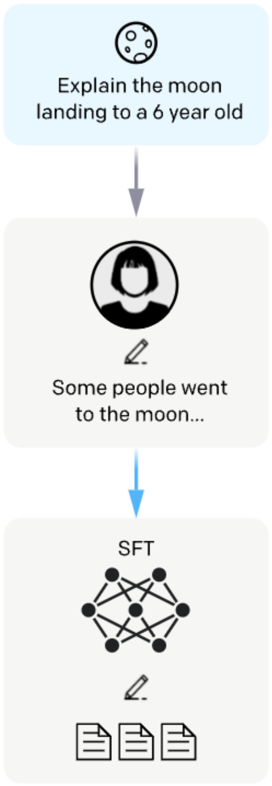
From GPT3.5 to ChatGPT: <u>Supervised Finetuning (SFT)</u> and <u>Reinforcement Learning from Human Feedback (RLHF)</u>

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



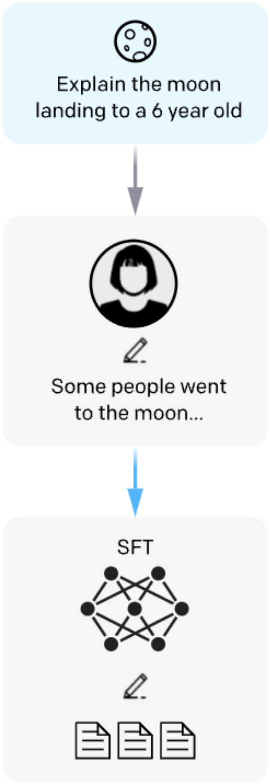
From GPT3.5 to ChatGPT: Supervised Finetuning (SFT) and Reinforcement Learning from Human Feedback (RLHF)

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



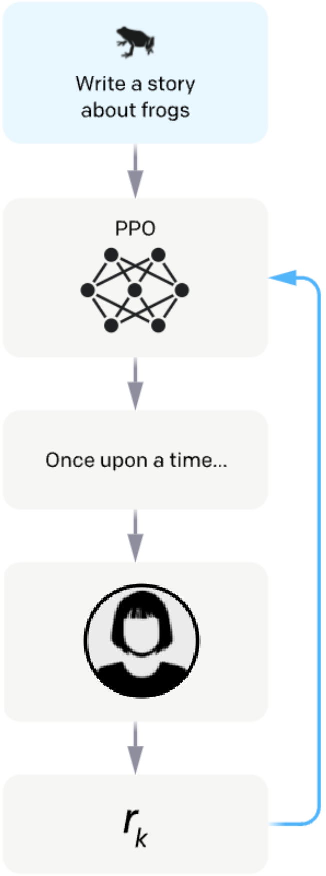
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

A labeler gives a reward for the output

The reward is used to update the policy using PPO.



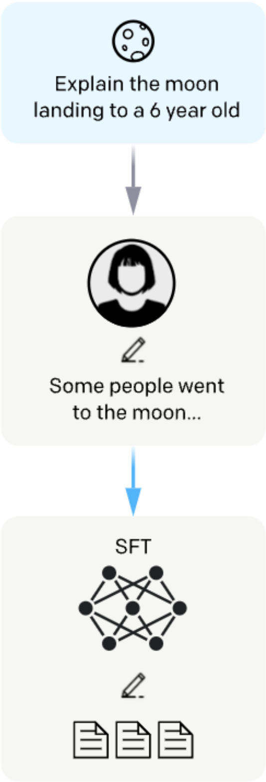
From GPT3.5 to ChatGPT: Supervised Finetuning (SFT) and Reinforcement Learning from Human Feedback (RLHF)

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



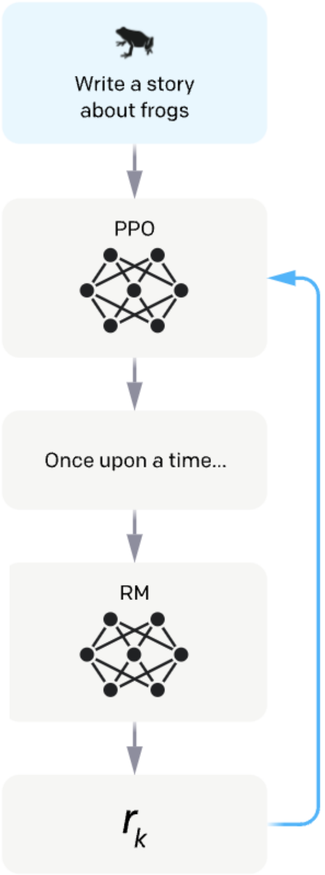
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

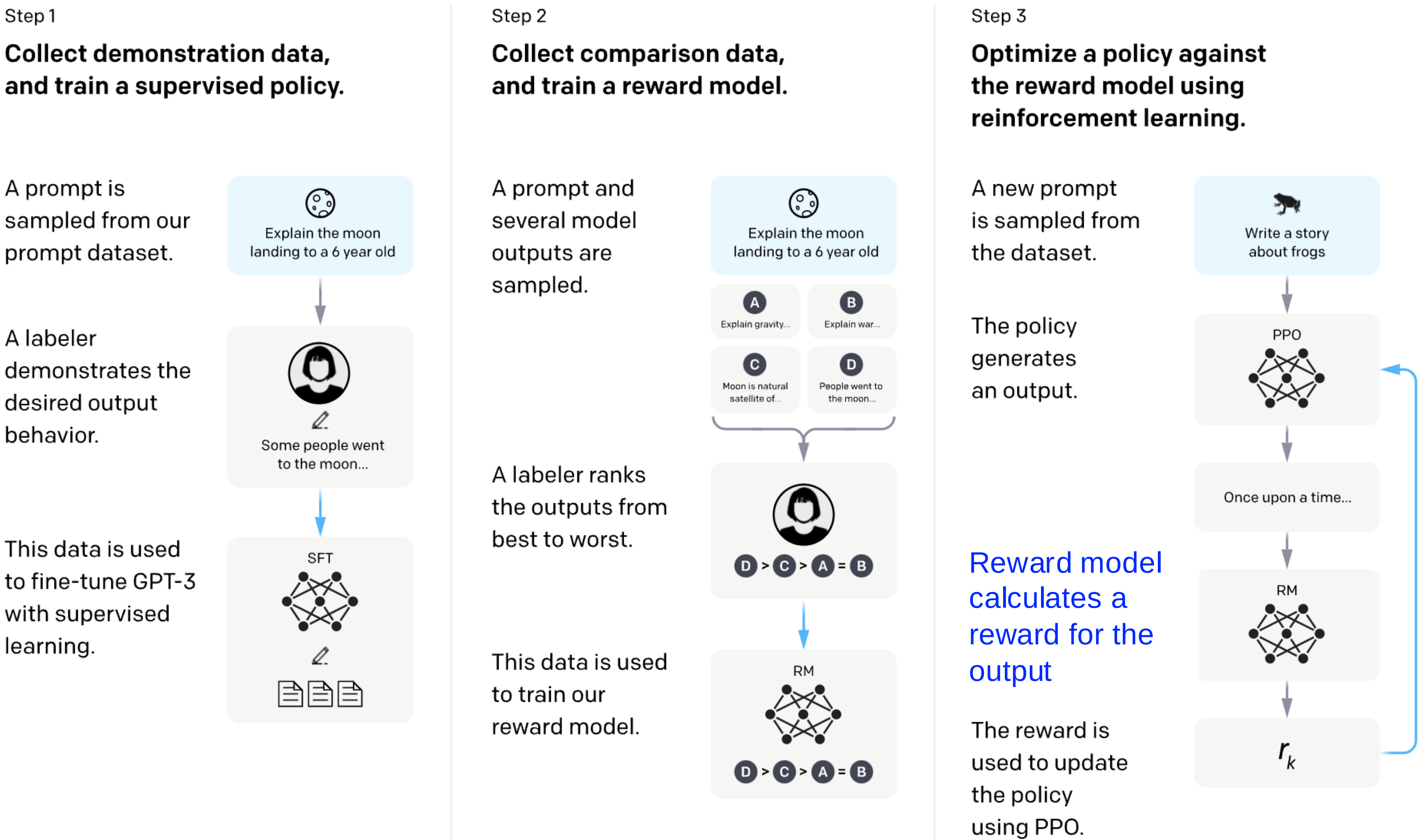
The policy generates an output.

Reward model calculates a reward for the output

The reward is used to update the policy using PPO.



From GPT3.5 to ChatGPT: Supervised Finetuning (SFT) and Reinforcement Learning from Human Feedback (RLHF)



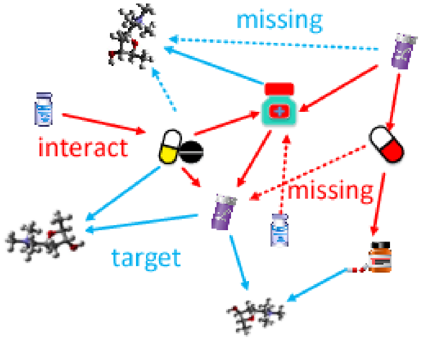
“Standard Model” of ML

Experience of all kinds

Data examples

Type-2 diabetes is 90% more common than type-1

Rules/Constraints

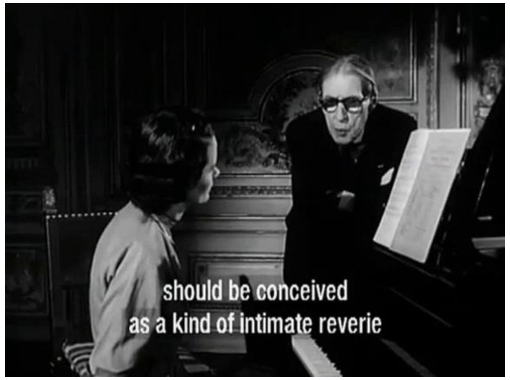


Knowledge graphs

Rewards

Auxiliary agents

Adversaries

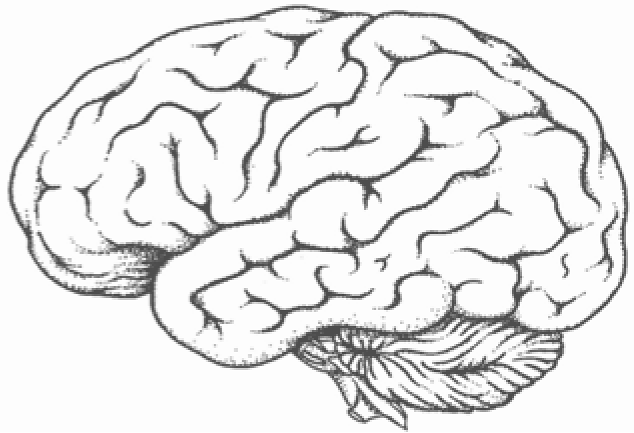


Master classes

…

- *And all combinations of such*
- *Interpolations between such*
- *…*

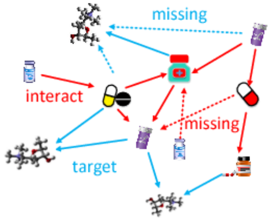
Human learning vs machine learning



Data examples

Type-2 diabetes is 90% more common than type-1

Rules/Constraints



Knowledge graphs

Rewards

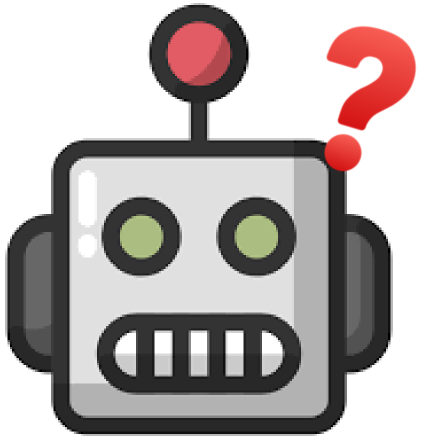
Auxiliary agents

Adversaries

Master classes

…

- *And all combinations of such*
- *Interpolations between such*
- *…*



The zoo of ML/AI algorithms

maximum likelihood estimation

reinforcement learning as inference

data re-weighting

inverse RL

policy optimization

active learning

data augmentation

actor-critic

reward-augmented maximum likelihood

label smoothing

imitation learning

softmax policy gradient

adversarial domain adaptation

posterior regularization

GANs

constraint-driven learning

knowledge distillation

intrinsic reward

generalized expectation

prediction minimization

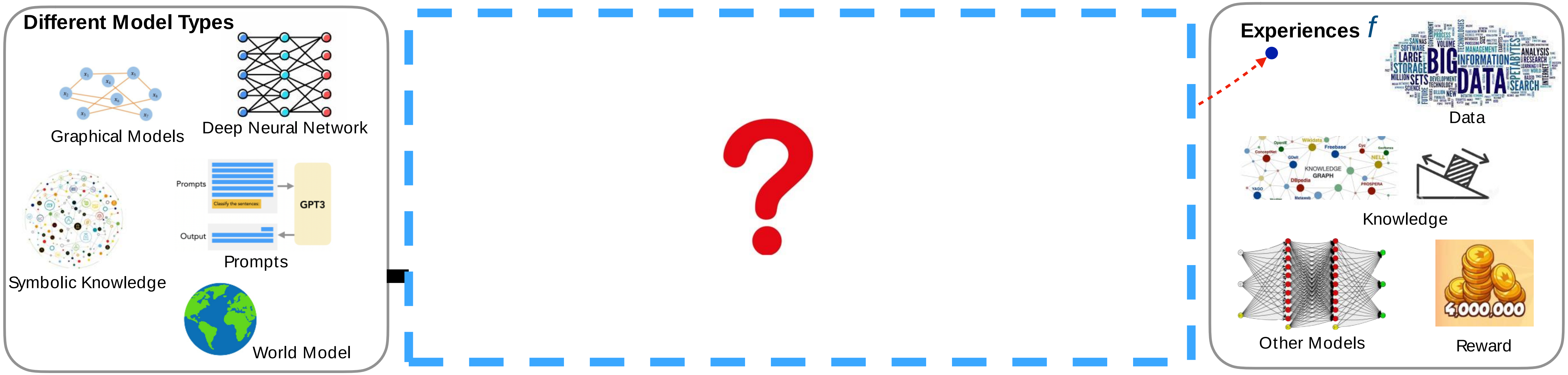
regularized Bayes

learning from measurements

energy-based GANs

weak/distant supervision

The zoo of ML/AI algorithms



Standard Model in Physics

Diverse electro-magnetic theories

Maxwell's Eqns: original form

$e+\frac{df}{dx}+\frac{dg}{dy}+\frac{dh}{dz}=0$	(1)	Gauss' Law
$\mu\alpha=\frac{dH}{dy}-\frac{dG}{dz}$ $\mu\beta=\frac{dF}{dz}-\frac{dH}{dx}$ $\mu\gamma=\frac{dG}{dx}-\frac{dF}{dy}$	(2)	Equivalent to Gauss' Law for magnetism
$P=\mu\left(\gamma\frac{dy}{dt}-\beta\frac{dz}{dt}\right)-\frac{dF}{dt}-\frac{d\Psi}{dx}$ $Q=\mu\left(\alpha\frac{dz}{dt}-\gamma\frac{dx}{dt}\right)-\frac{dG}{dt}-\frac{d\Psi}{dy}$ $R=\mu\left(\beta\frac{dx}{dt}-\alpha\frac{dy}{dt}\right)-\frac{dH}{dt}-\frac{d\Psi}{dz}$	(3)	Faraday's Law (with the Lorentz Force and Poisson's Law)
$\frac{d\gamma}{dy}-\frac{d\beta}{dz}=4\pi p'$ $p'=p+\frac{df}{dt}$ $\frac{d\alpha}{dz}-\frac{d\gamma}{dx}=4\pi q'$ $q'=q+\frac{dg}{dt}$ $\frac{d\beta}{dx}-\frac{d\alpha}{dy}=4\pi r'$ $r'=r+\frac{dh}{dt}$	(4)	Ampère-Maxwell Law
$P=-\xi p \quad Q=-\xi q \quad R=-\xi r$		Ohm's Law
$P=kf \quad Q=kg \quad R=kh$		The electric elasticity equation ($\mathbf{E}=\mathbf{D}/\varepsilon$)
$\frac{de}{dt}+\frac{dp}{dx}+\frac{dq}{dy}+\frac{dr}{dz}=0$		Continuity of charge

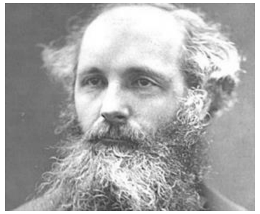
Simplified w/ rotational symmetry

$$\nabla\cdot\mathbf{D}=\rho_V$$

$$\nabla\cdot\mathbf{B}=0$$

$$\nabla\times\mathbf{E}=-\frac{\partial\mathbf{B}}{\partial t}$$

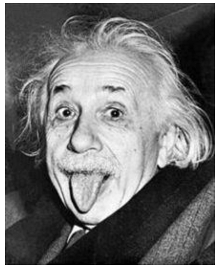
$$\nabla\times\mathbf{H}=\frac{\partial\mathbf{D}}{\partial t}+\mathbf{J}$$



Further simplified w/ symmetry of special relativity

$$\varepsilon^{uvk\lambda}\partial_v F_{k\lambda}=0$$

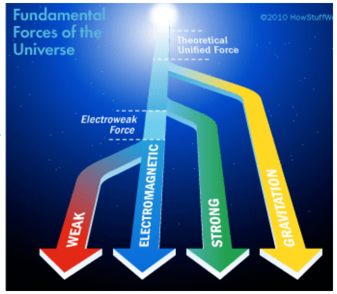
$$\partial_v F^{uV}=\frac{4\pi}{c}j^u$$



Standard Model w/ Yang-Mills theory and US(3) symmetry

$$\mathcal{L}_{\mathrm{gf}}=-\frac{1}{2}\mathrm{Tr}(F^2)=-\frac{1}{4}F^{a\mu\nu}F^a_{\mu\nu}$$

Unification of fundamental forces?



1861   1910s   1970s   ?

Toward a 'Standard Model' of Machine Learning

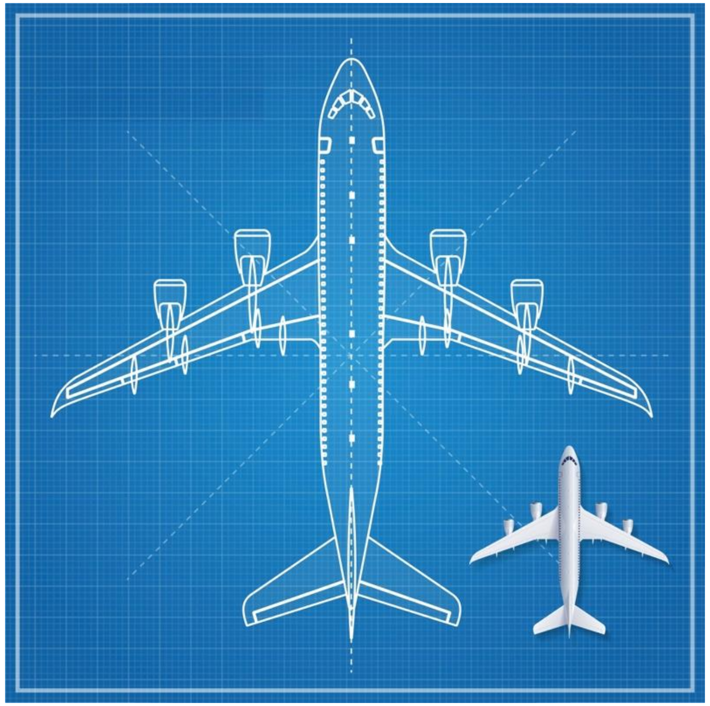
Zhiting Hu[†,*], Eric P. Xing[‡,⋄,♮,**]

[†] Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, USA

[‡] Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA

[♮] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

[⋄] Petuum Inc., Pittsburgh, USA



[Hu & Xing, Harvard Data Science Review, 2022]: *https://arxiv.org/abs/2108.07783*

$$\min_{q,\theta} \; -\mathbb{E} + \mathbb{D} - \mathbb{H}$$

Experience Divergence Uncertainty

A "Standard Model" of ML

$$\min_{q,\theta} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\left(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\right) - \mathbb{E}_{q(\boldsymbol{t})}\left[f(\boldsymbol{t})\right]$$

3 terms:

Uncertainty
(self-regularization)
e.g., Shannon entropy

Divergence
(fitness)
e.g., Cross Entropy

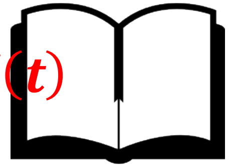
Experiences
(exogenous regularizations)
e.g., data examples, rules

Uncertainty

Teacher
$q(\boldsymbol{t})$

Student
$p_\theta(\boldsymbol{t})$

Textbook $f(\boldsymbol{t})$



Presentations

Questions?