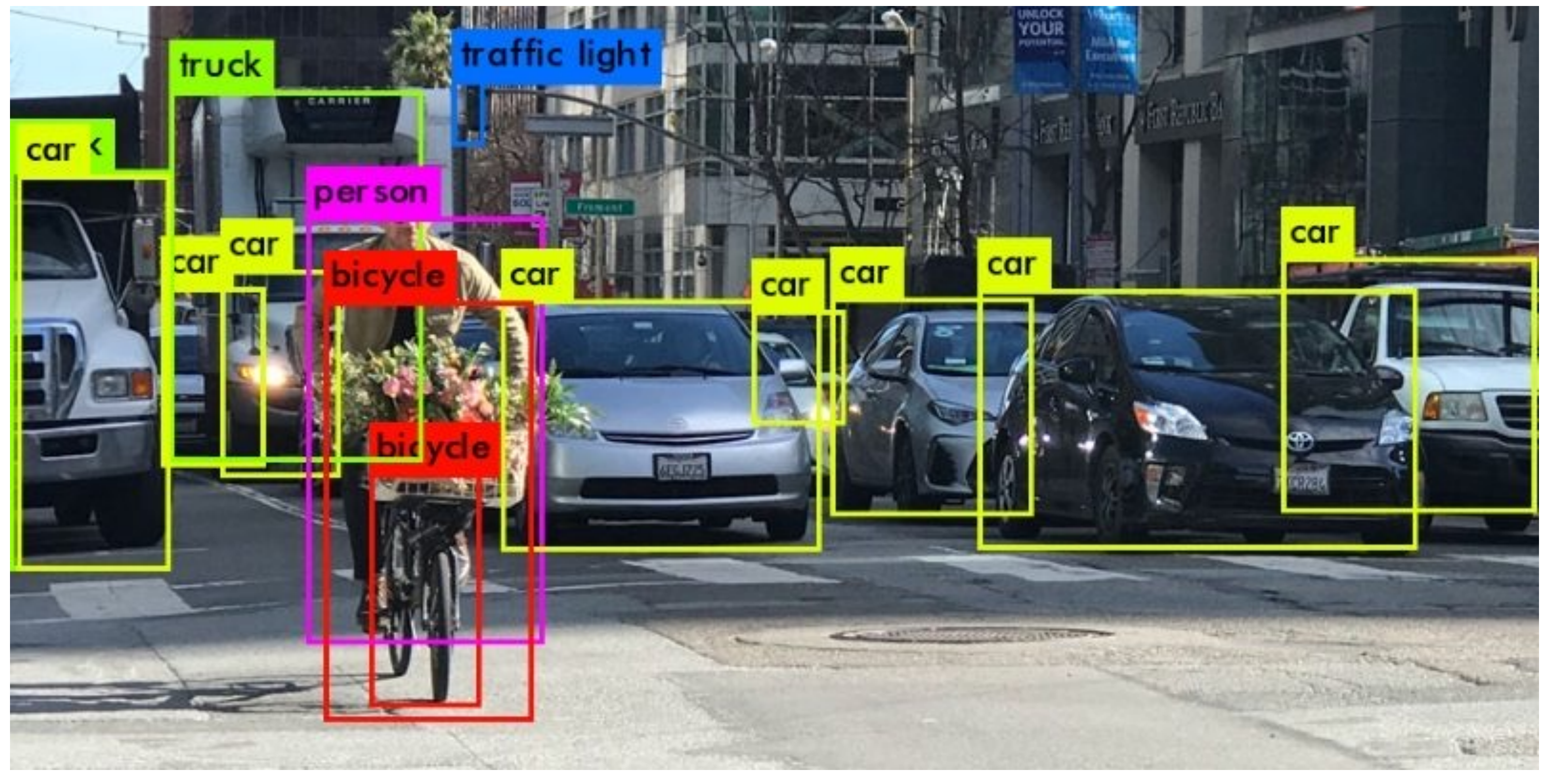
HALICIOĞLU DATA SCIENCE INSTITUTE

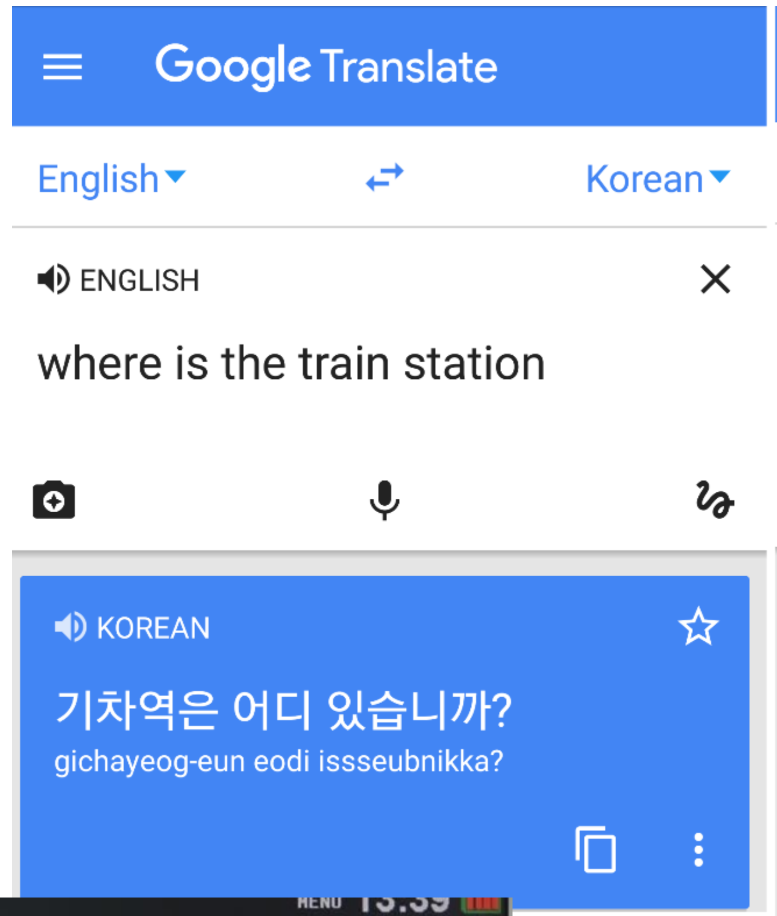
Toward A “Standard Model” of Machine Learning

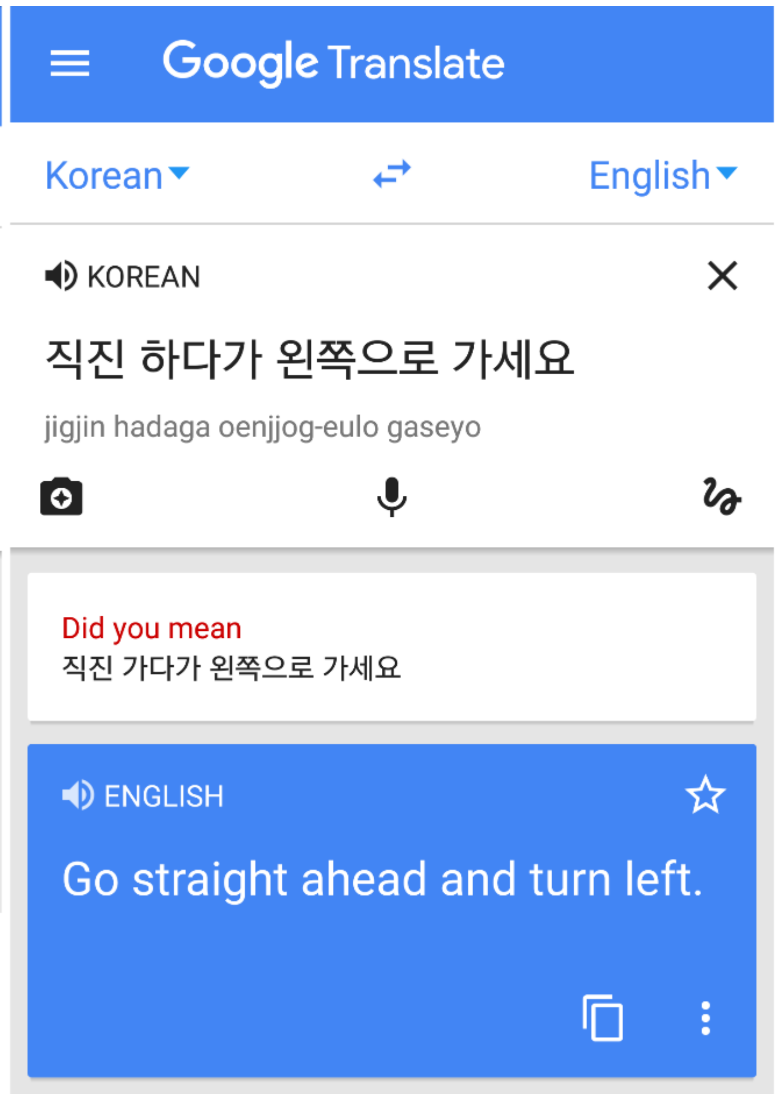
Zhiting Hu
Assistant Professor
Halicioglu Data Science Institute
Computer Science and Engineering
UC San Diego

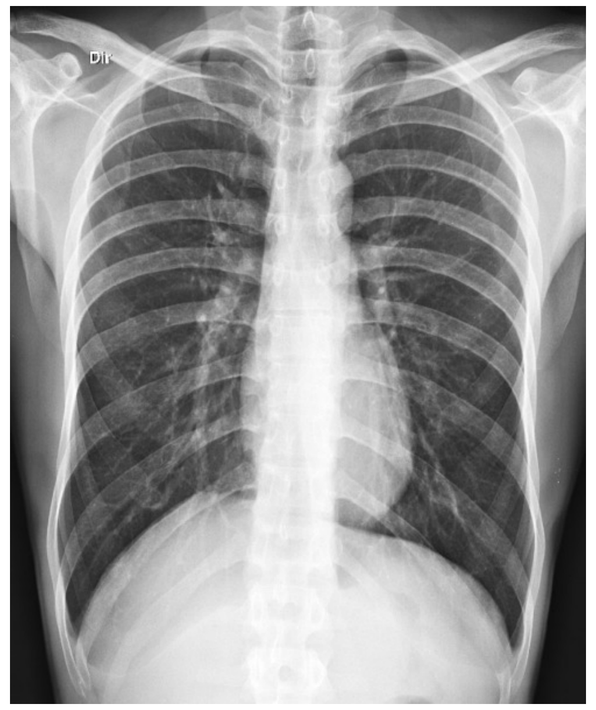
06/12/2023

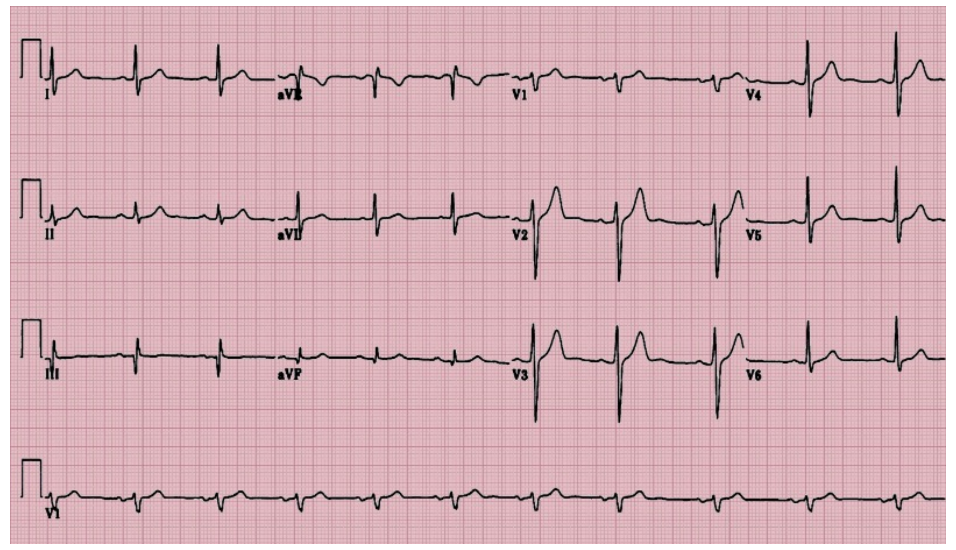
The universe of problems ML/AI is trying to solve



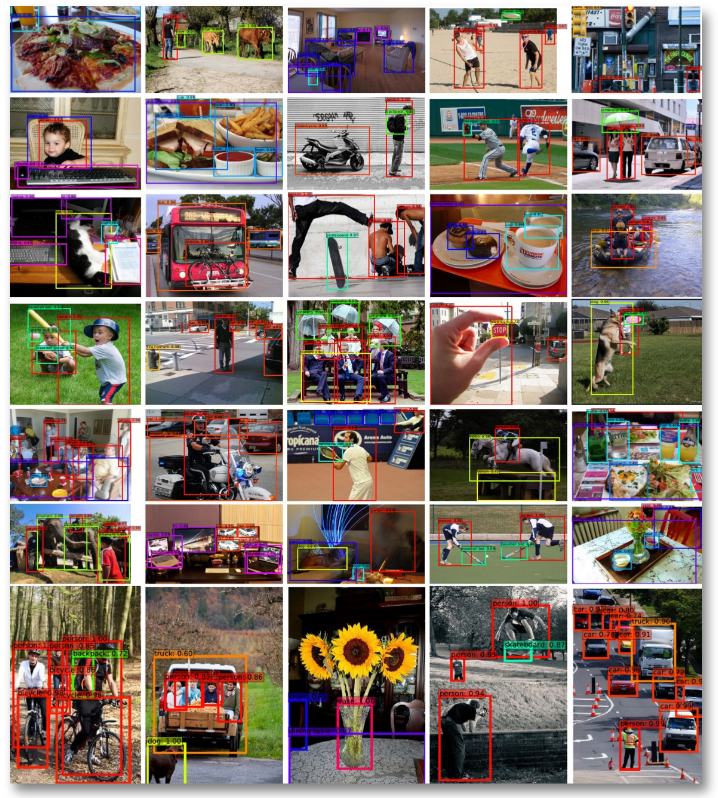








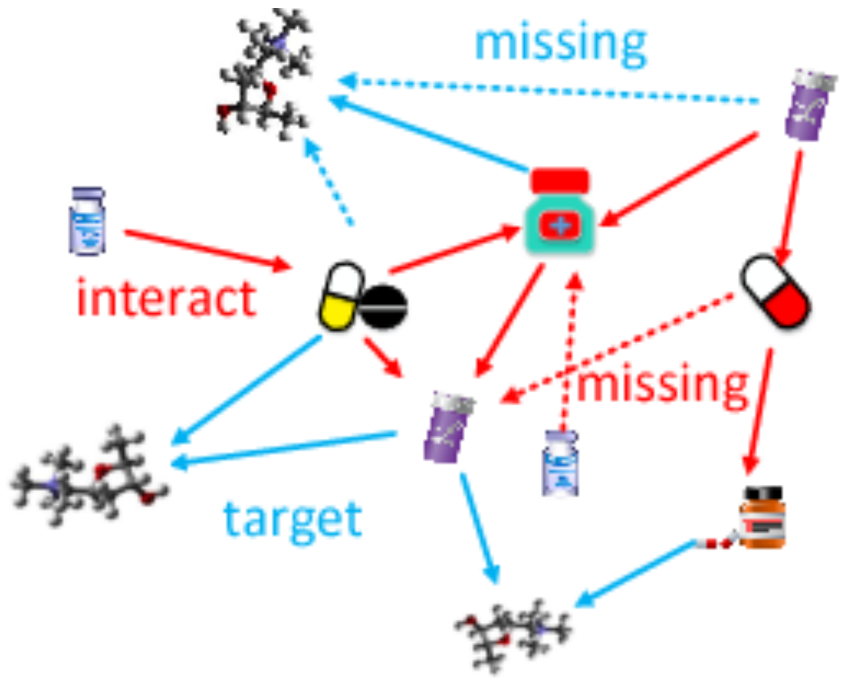
Experience of all kinds



Data examples

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

Rules/Constraints

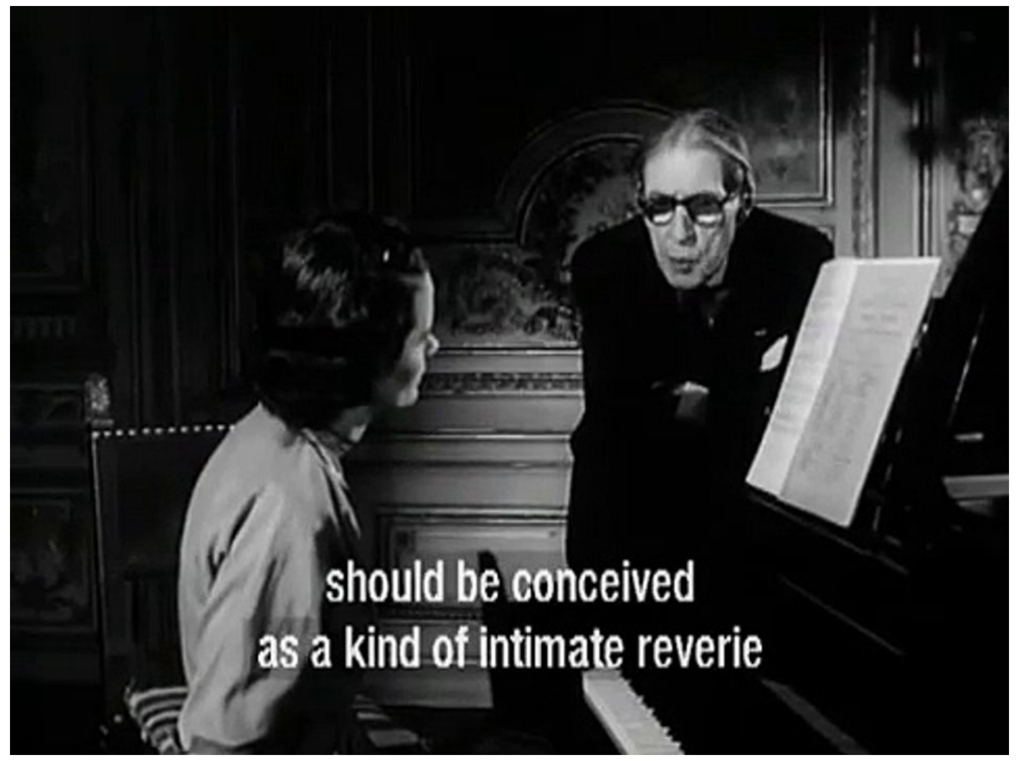


Knowledge graphs

Rewards

Auxiliary agents

Adversaries

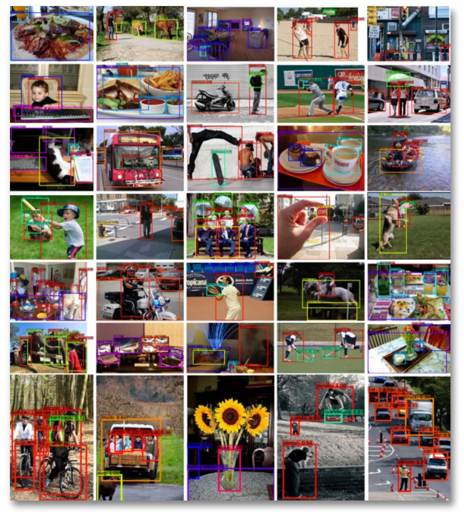


Master classes

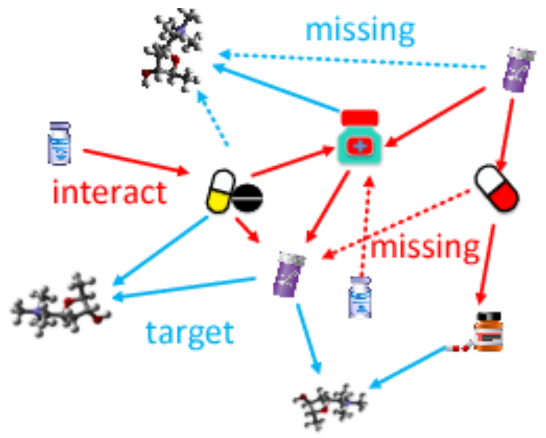
…

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

Human learning vs machine learning



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



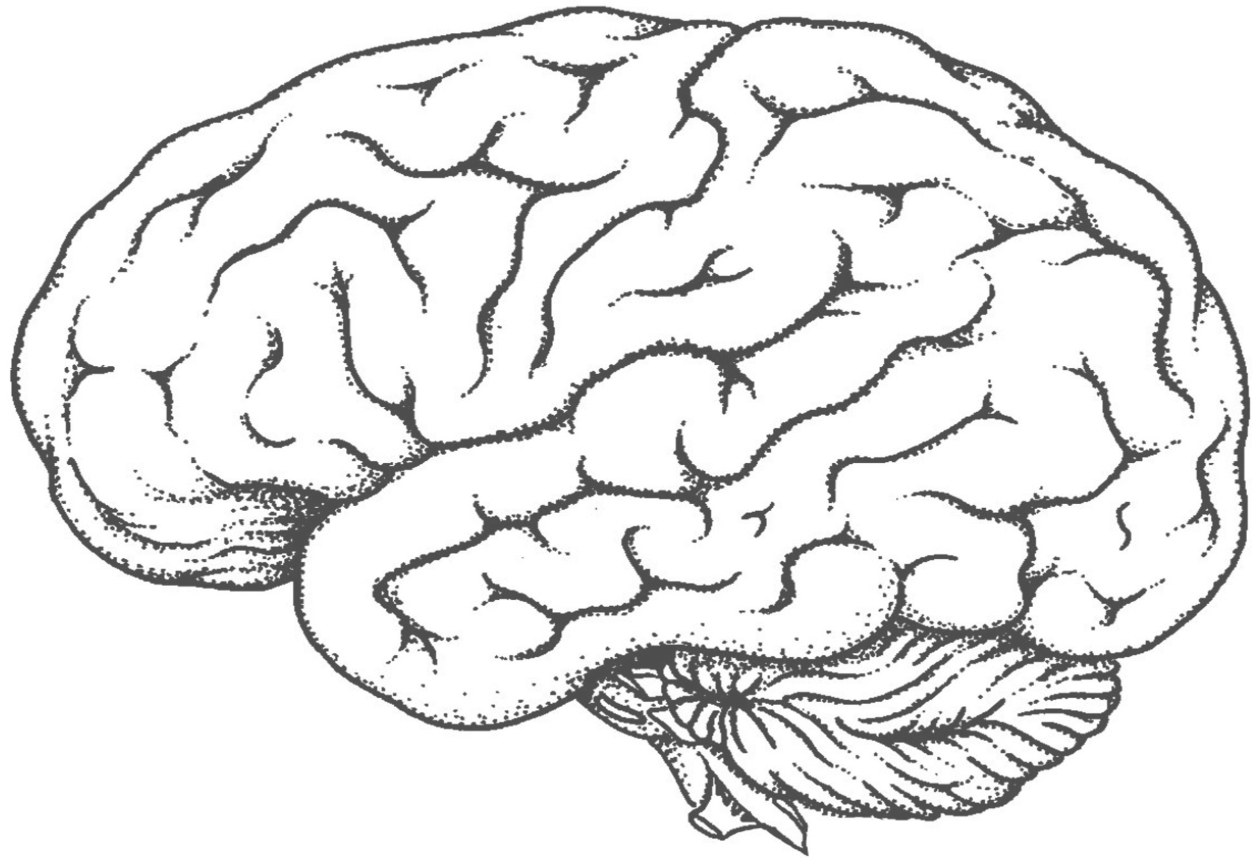
Data examples *Rules/Constraints* *Knowledge graphs* *Rewards* *Auxiliary agents*

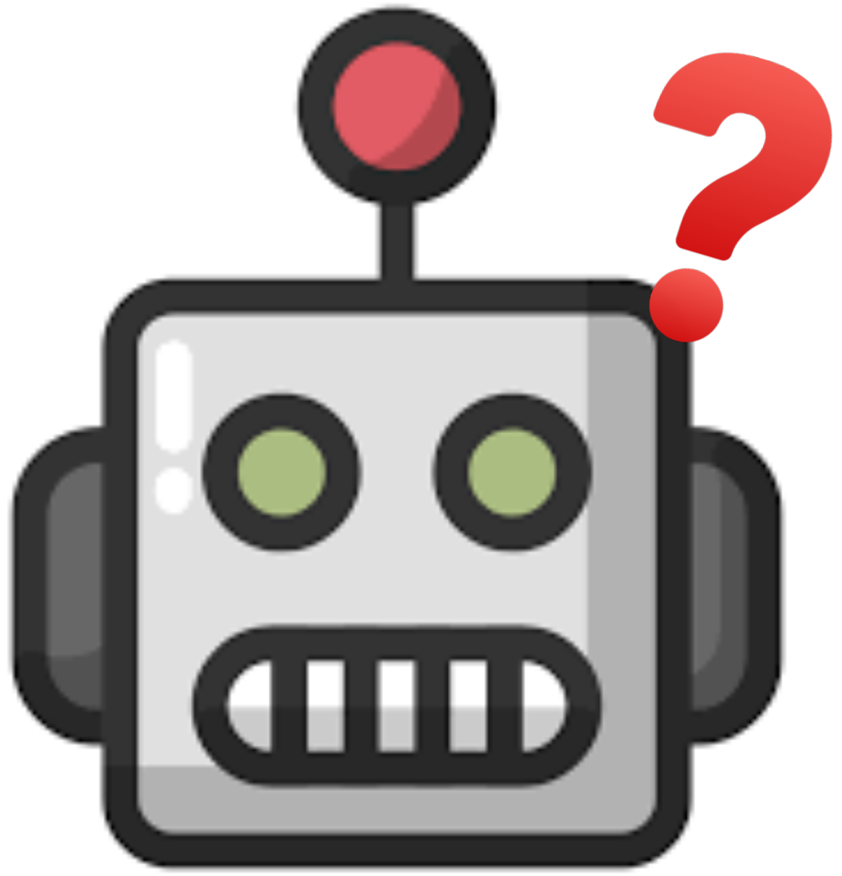
Adversaries

Master classes

…

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





The zoo of ML 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

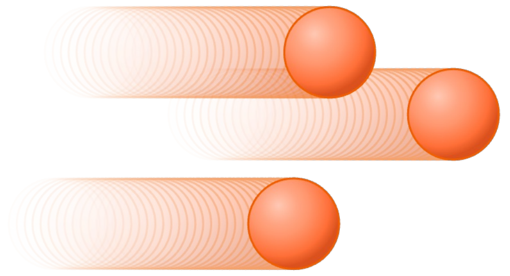
energy-based GANs

weak/distant supervision

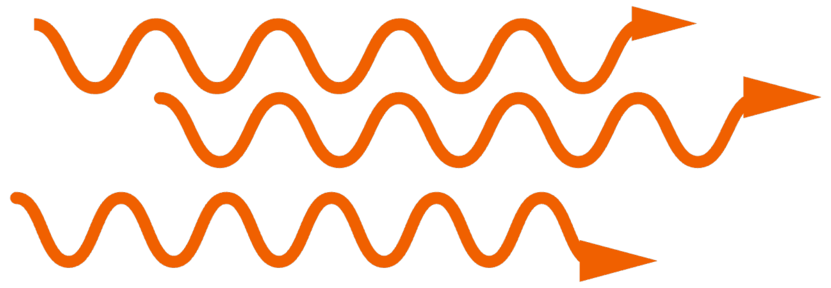
Physics in the 1800's

- Electricity & magnetism:
  - Coulomb's law, Ampère, Faraday, ...

- Theory of light beams:
  - Particle theory: Isaac Newton, Laplace, Plank



  - Wave theory: Grimaldi, Chris Huygens, Thomas Young, Maxwell



- Law of gravity
  - Aristotle, Galileo, Newton, …

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

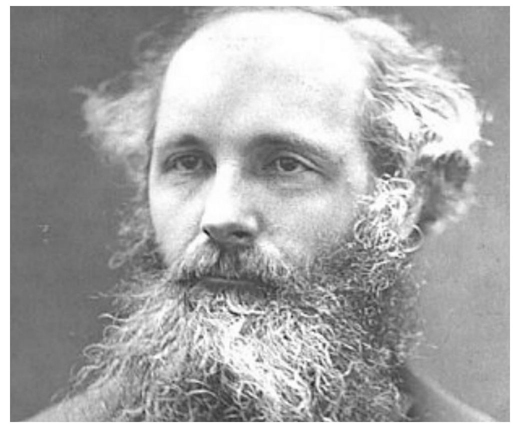
Simplified w/ rotational symmetry

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

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

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

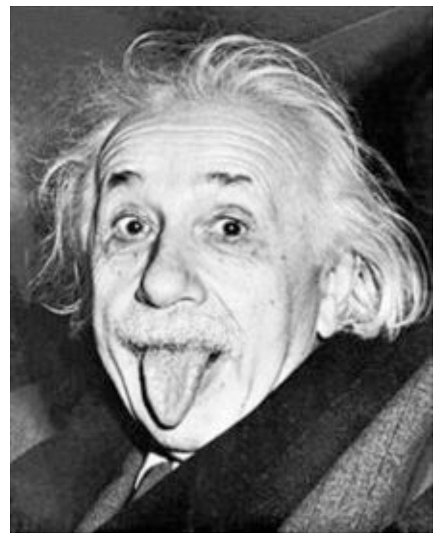
$$\nabla\times\mathbf{H}=\frac{\partial\mathbf{D}}{\partial t}+\mathbf{J}$$



Further simplified w/ symmetry of special relativity

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

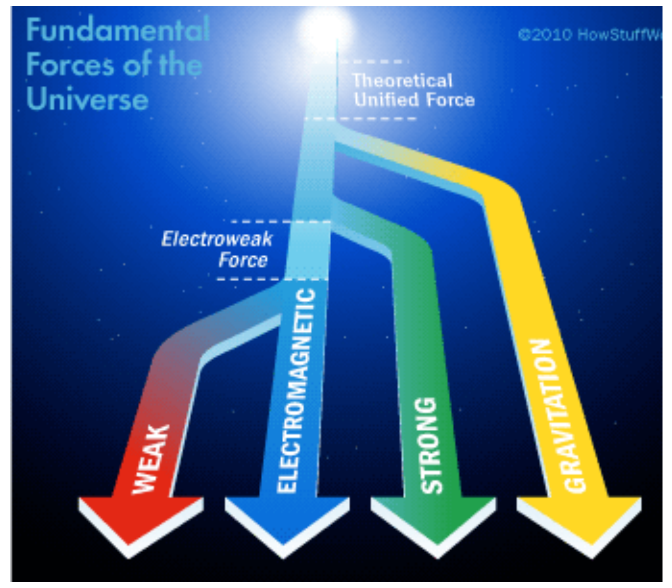
$$\partial_v F^{uV}=\frac{4\pi}{c}j^u$$



***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 — ?

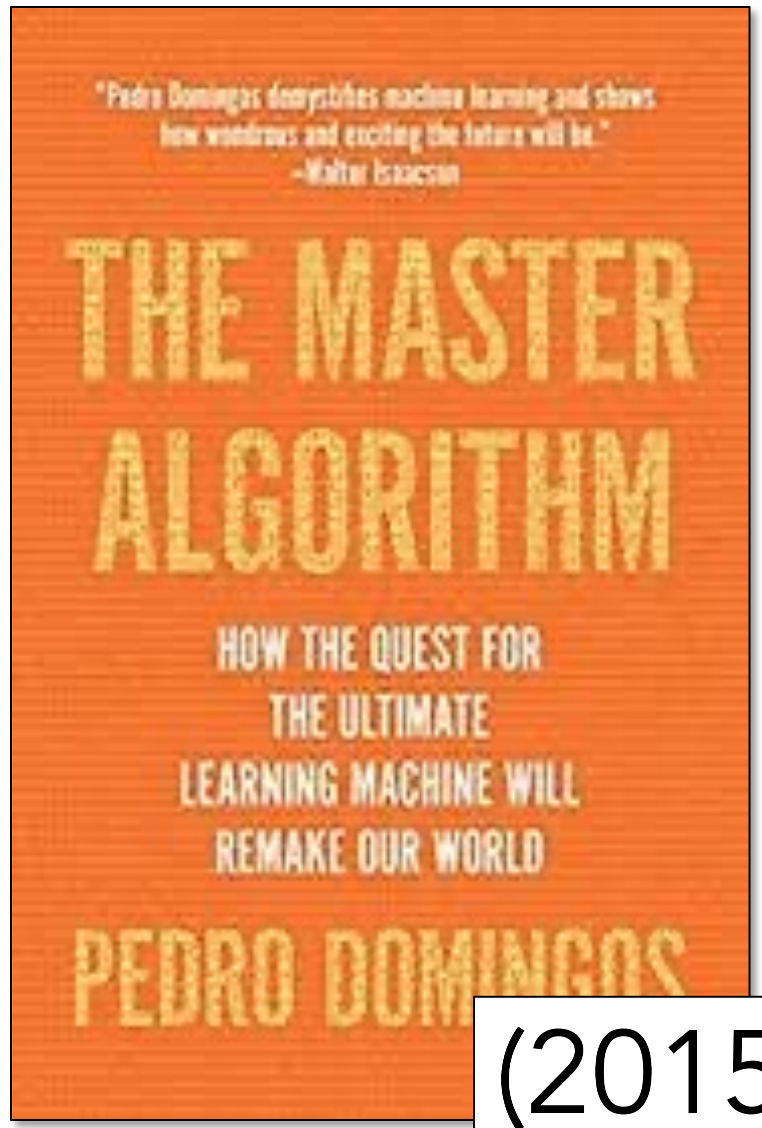
Quest for more standardized, unified ML principles

Machine Learning 3: 253–259, 1989
© 1989 Kluwer Academic Publishers – Manufactured in The Netherlands

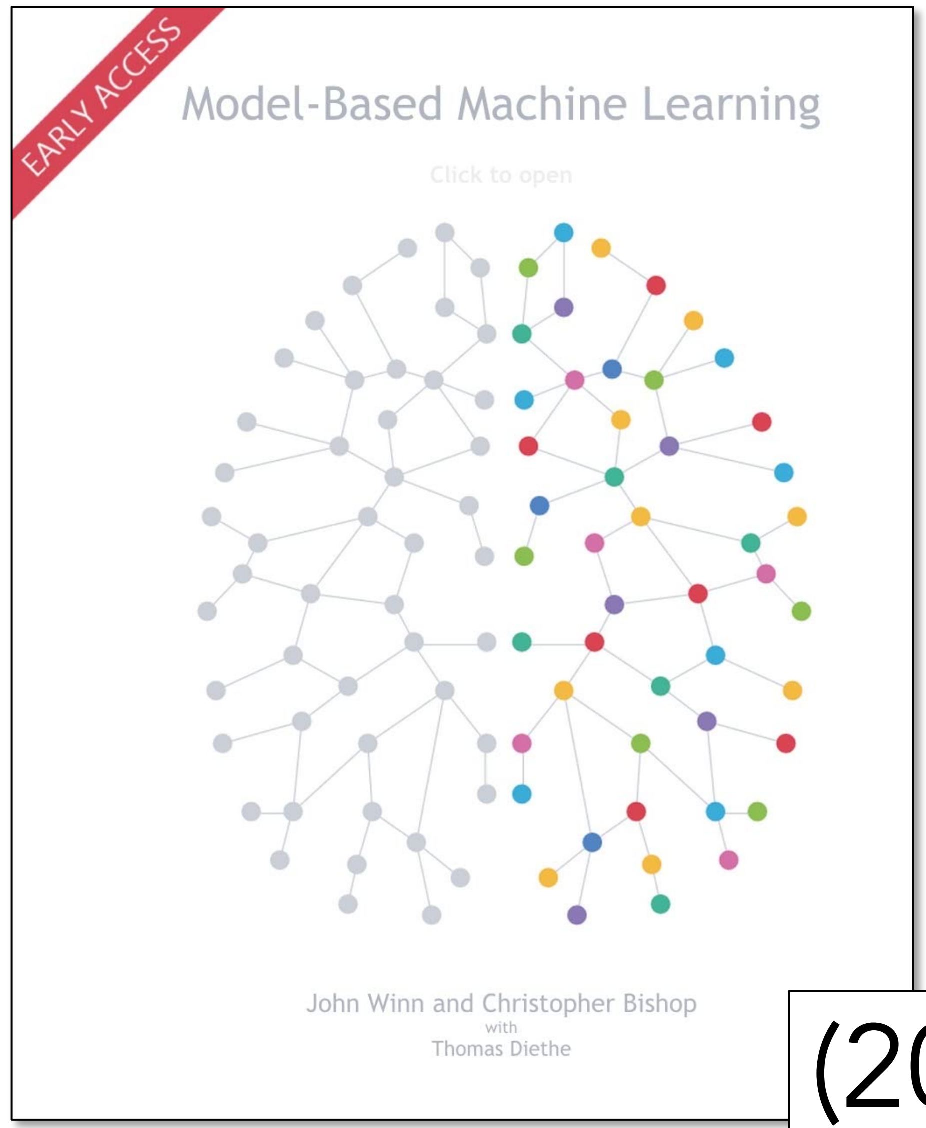
EDITORIAL

Toward a Unified Science of Machine Learning

(P. Langley, 1989)



(2015)



(2019)

REVIEW — Communicated by Steven Nowlan

A Unifying Review of Linear Gaussian Models

Sam Roweis*
Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125, U.S.A.

Zoubin Ghahramani*
Department of Computer Science, University of Toronto, Toronto, Canada

(1999)

Quest for more standardized, unified ML principles

Is Large Language Model (LLM) the answer?

"Self-supervised" learning + large (text) data

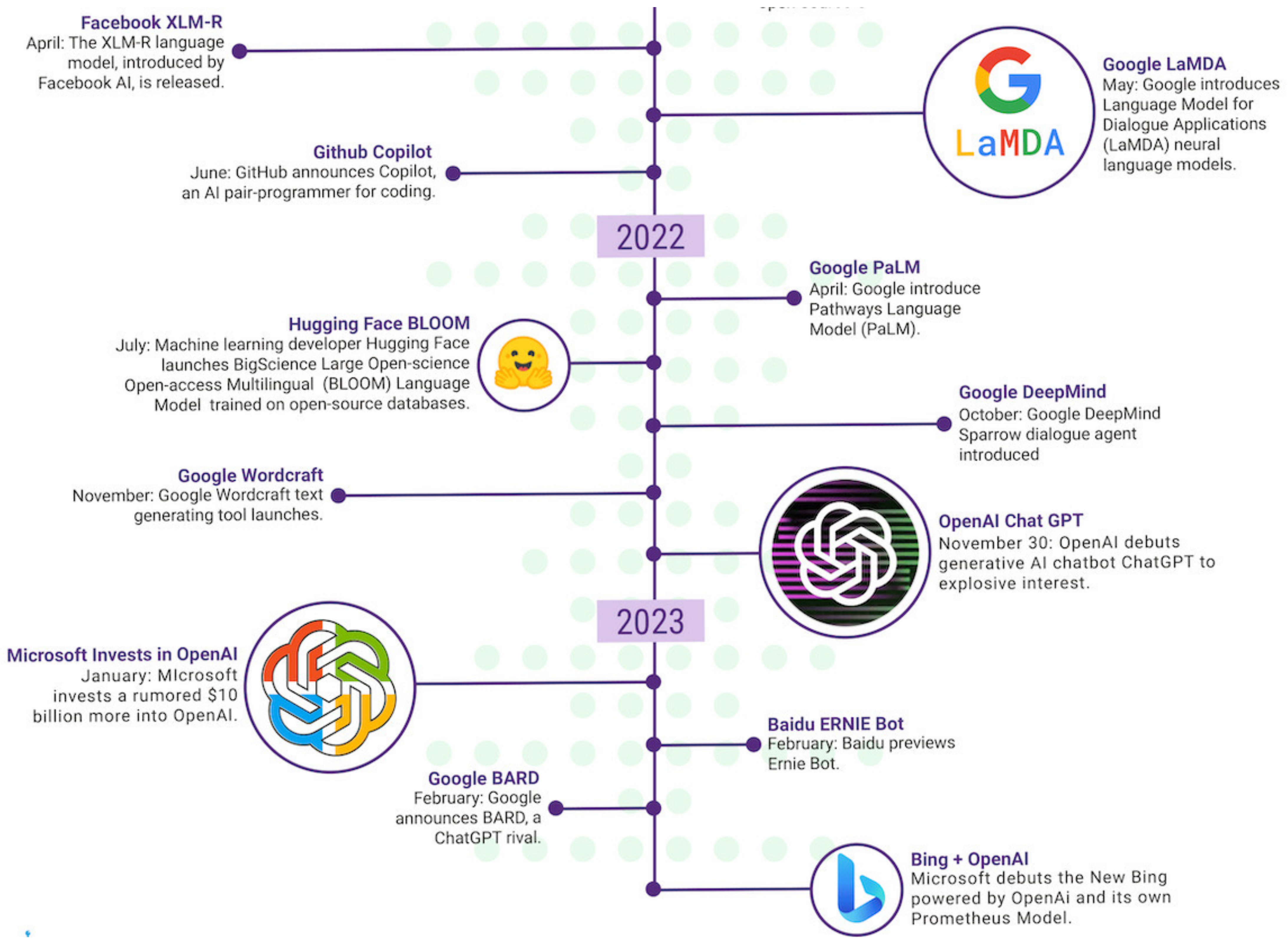


figure credit: Voicebot.ai

Limited understanding of the world

John put a **book** on the desk.
…
Mary took the **book**.
She placed it on the sofa.
…
Where was the **book**?

ChatGPT

It was on the desk.

Quest for more standardized, unified ML principles

Is Large Language Model (LLM) the answer?

"Self-supervised" learning + large (text) data

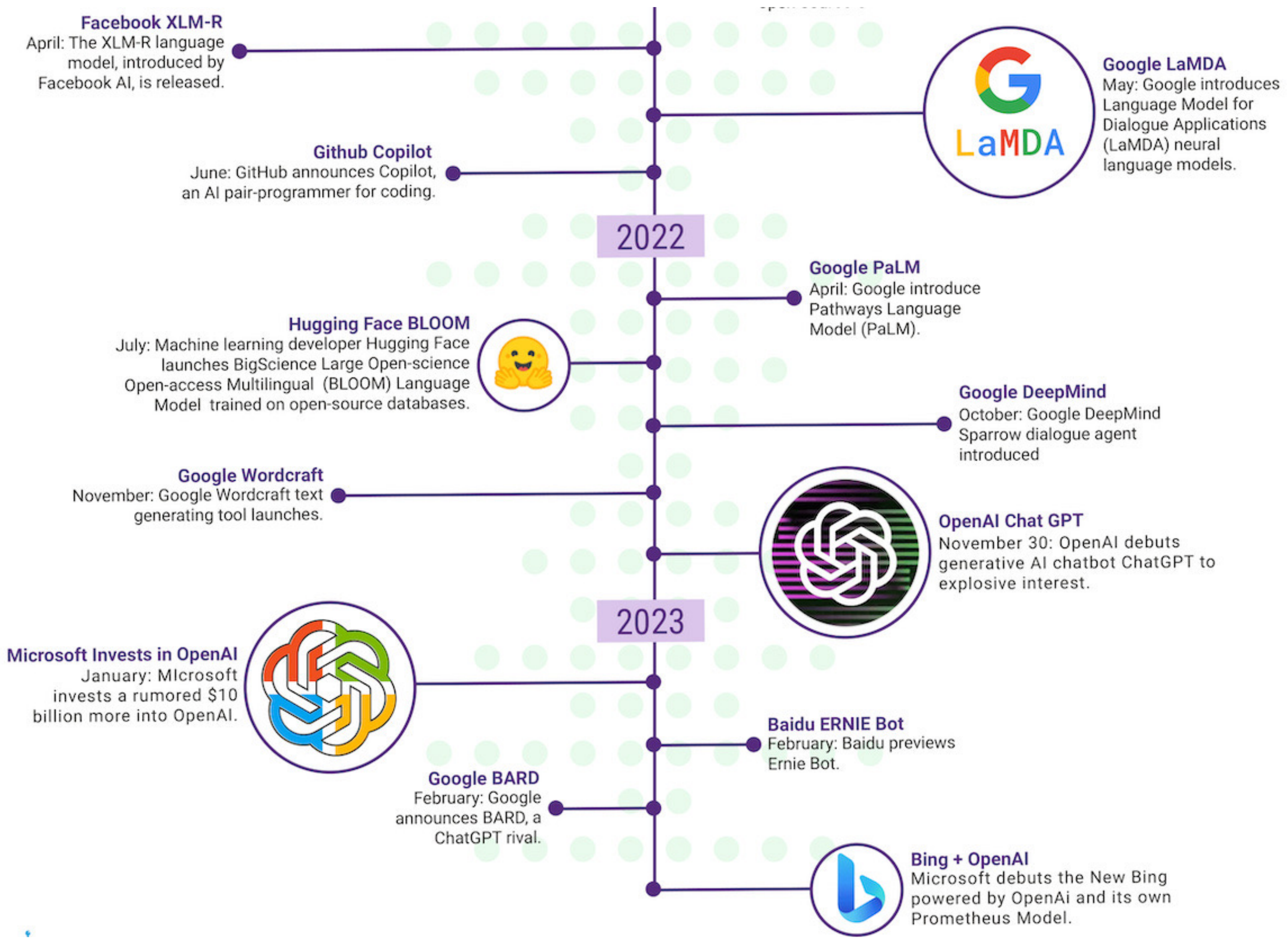
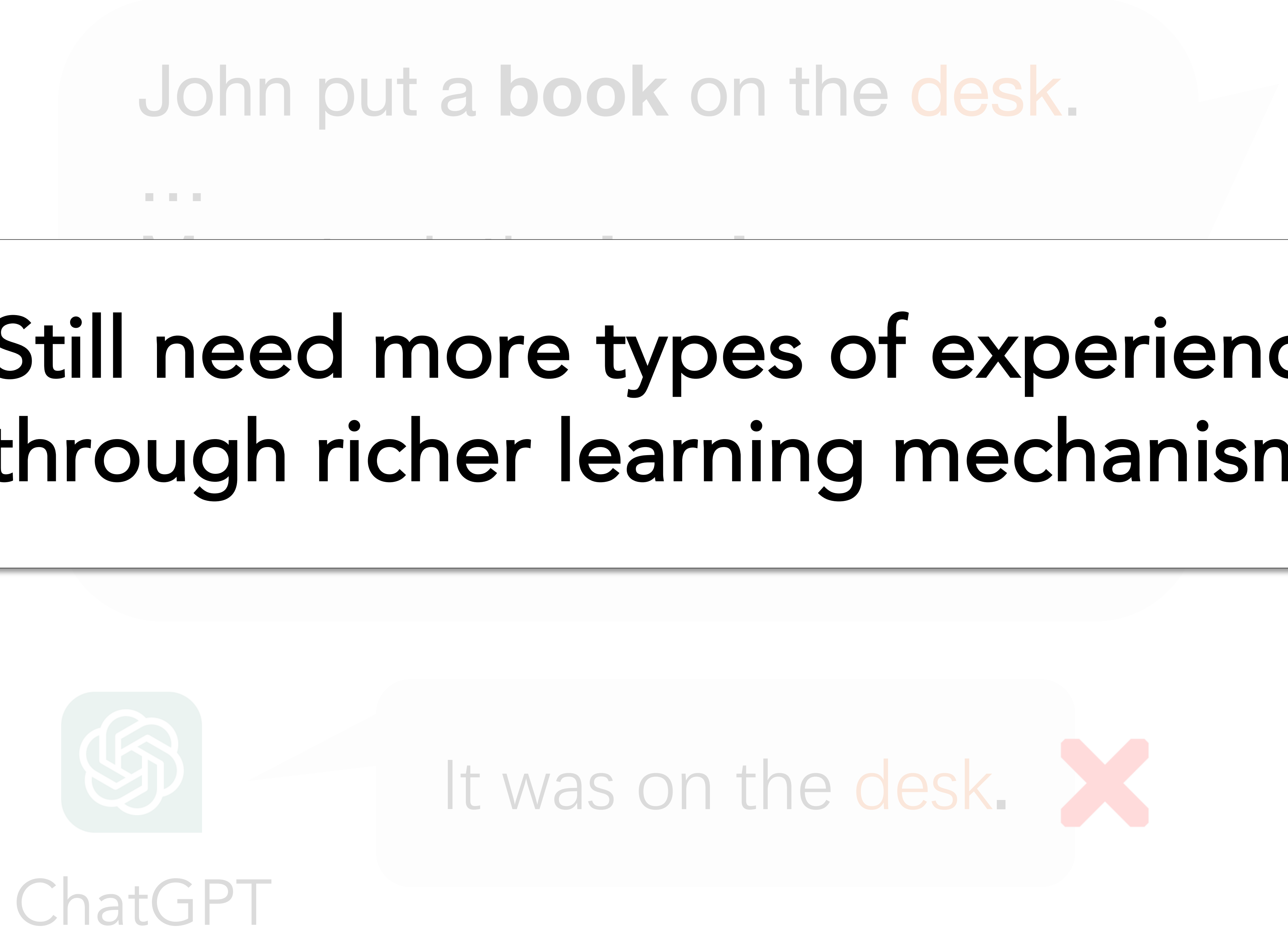


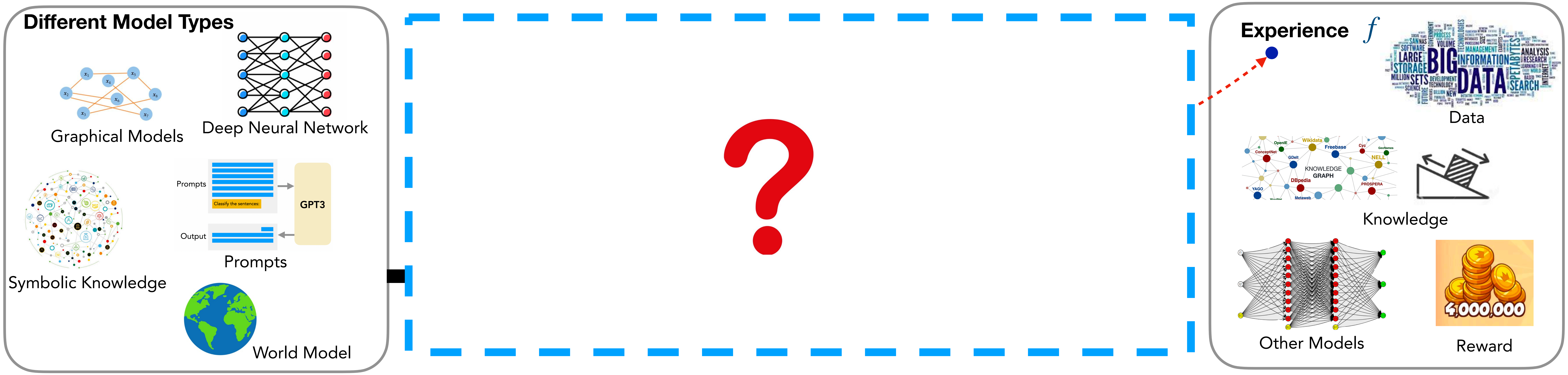
figure credit: Voicebot.ai

Limited understanding of the world

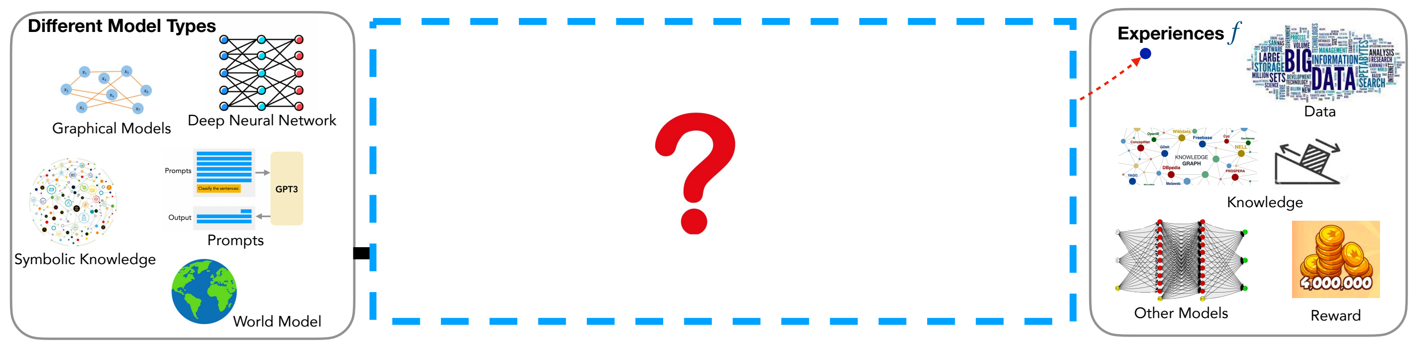


Still need more types of experience through richer learning mechanisms

A “Standard Model” of Machine Learning



Hu and Xing, **Towards A ‘Standard Model’ of Machine Learning**, Harvard Data Science Review, 2022



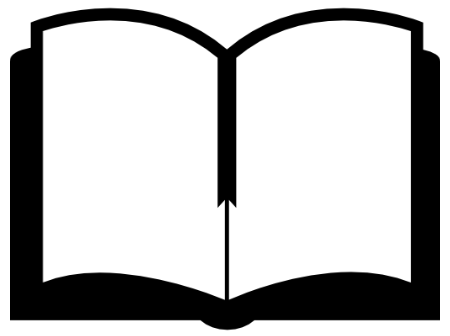
A "Standard Model" of Machine Learning

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

3 terms:

Experience	*Divergence*	*Uncertainty*
(exogenous regularizations)	*(fitness)*	*(self-regularization)*
e.g. data examples, reward	*e.g. Cross Entropy*	*e.g. Shannon entropy*

Textbook
$f(\boldsymbol{x},\boldsymbol{y} \mid .)$



Teacher
$q(\boldsymbol{x},\boldsymbol{y})$

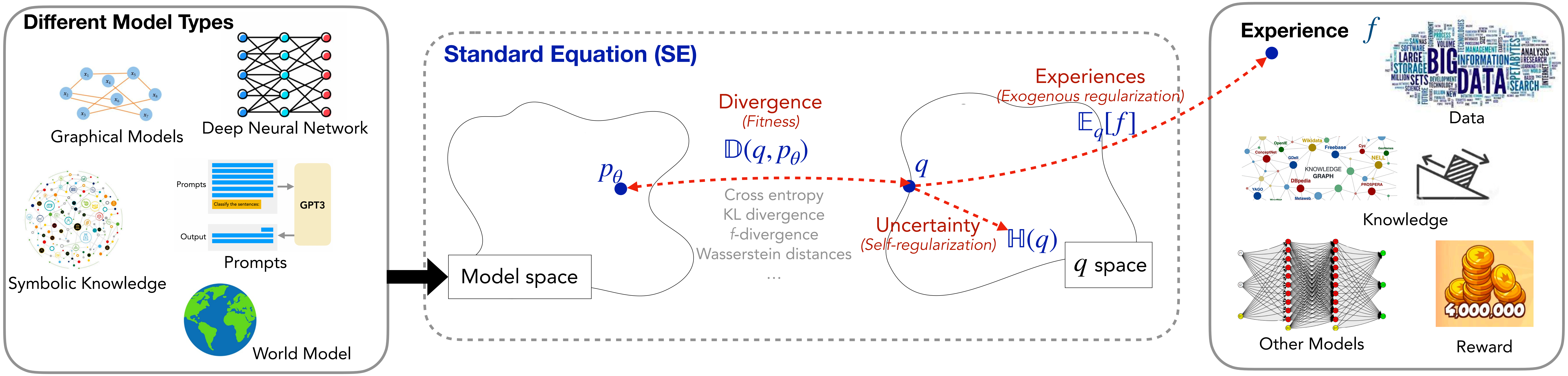
Student
$p_\theta(\boldsymbol{x},\boldsymbol{y})$

Uncertainty

Hu and Xing, **Towards A 'Standard Model' of Machine Learning**, Harvard Data Science Review, 2022

A "Standard Model" of Machine Learning

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



Hu and Xing, **Towards A 'Standard Model' of Machine Learning**, Harvard Data Science Review, 2022

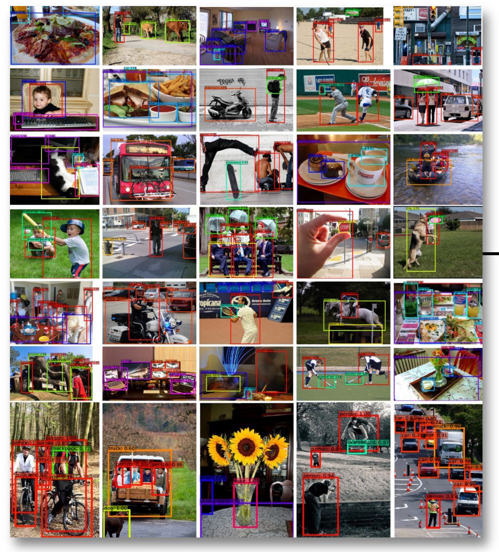
"Standard Model" encompasses well-known ML algorithms as special cases

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

"Standard Model" encompasses well-known ML algorithms as special cases

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

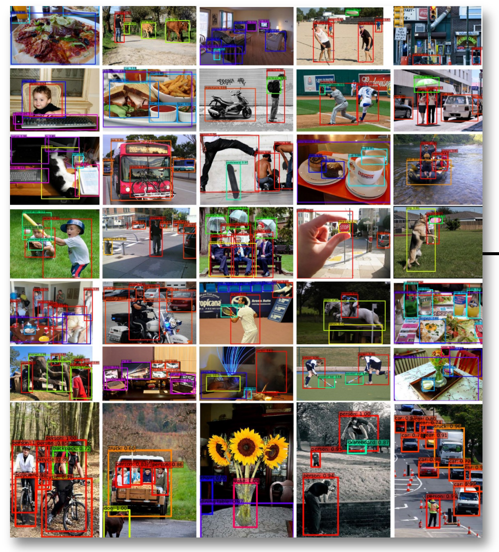
Experience type	Experience function f	Divergence $\mathbb{D}$	α	β	Algorithm
Data instances	$f_{\text{data}}(\boldsymbol{x};\mathcal{D})$	CE	1	1	Unsupervised MLE
	$f_{\text{data}}(\boldsymbol{x},\boldsymbol{y};\mathcal{D})$	CE	1	ϵ	Supervised MLE
	$f_{\text{data-self}}(\boldsymbol{x},\boldsymbol{y};\mathcal{D})$	CE	1	ϵ	Self-supervised MLE
	$f_{\text{data-w}}(\boldsymbol{t};\mathcal{D})$	CE	1	ϵ	Data Re-weighting
	$f_{\text{data-aug}}(\boldsymbol{t};\mathcal{D})$	CE	1	ϵ	Data Augmentation
	$f_{\text{active}}(\boldsymbol{x},\boldsymbol{y};\mathcal{D})$	CE	1	ϵ	Active Learning (Ertekin et al., 2007)



"Standard Model" encompasses well-known ML algorithms as special cases

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

Experience type	Experience function f	Divergence $\mathbb{D}$	α	β	Algorithm
Data instances	$f_{\text{data}}(\boldsymbol{x};\mathcal{D})$	CE	1	1	Unsupervised MLE
	$f_{\text{data}}(\boldsymbol{x},\boldsymbol{y};\mathcal{D})$	CE	1	ϵ	Supervised MLE
	$f_{\text{data-self}}(\boldsymbol{x},\boldsymbol{y};\mathcal{D})$	CE	1	ϵ	Self-supervised MLE
	$f_{\text{data-w}}(\boldsymbol{t};\mathcal{D})$	CE	1	ϵ	Data Re-weighting
	$f_{\text{data-aug}}(\boldsymbol{t};\mathcal{D})$	CE	1	ϵ	Data Augmentation
	$f_{\text{active}}(\boldsymbol{x},\boldsymbol{y};\mathcal{D})$	CE	1	ϵ	Active Learning (Ertekin et al., 2007)



$f_{\text{data}}(\boldsymbol{x},\boldsymbol{y};\mathcal{D})$
$:= \log \mathbb{E}_{(\boldsymbol{x}^*,\boldsymbol{y}^*)\sim\mathcal{D}}\big[\mathbb{1}_{(\boldsymbol{x}^*,\boldsymbol{y}^*)}(\boldsymbol{x},\boldsymbol{y}) \big]$ ➡ $q(\boldsymbol{x},\boldsymbol{y}) = \tilde{p}_{\text{data}}(\boldsymbol{x},\boldsymbol{y})$ ➡ $\min_{\theta} - \mathbb{E}_q\big[\log p_\theta(\boldsymbol{x},\boldsymbol{y})\big]$

(Negative data log-likelihood)

"Standard Model" encompasses well-known ML algorithms as special cases

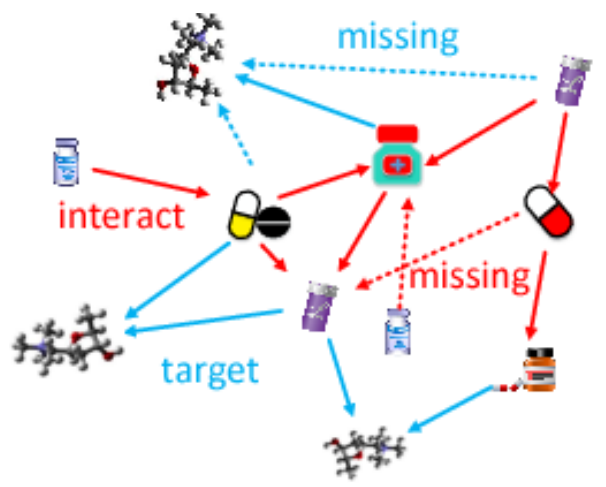
$$\min_{q,\theta} - \mathbb{E}_{q(\boldsymbol{x},\boldsymbol{y})}\left[f(\boldsymbol{x},\boldsymbol{y}) \right] + \alpha \mathbb{D}\left(q(\boldsymbol{x},\boldsymbol{y}), p_\theta(\boldsymbol{x},\boldsymbol{y}) \right) - \beta \mathbb{H}(q)$$

Experience type	Experience function f	Divergence $\mathbb{D}$	α	β	Algorithm
	$\log Q^{\theta}(\boldsymbol{x},\boldsymbol{y})$	CE	1	1	Policy Gradient
Reward	$\log Q^{\theta}(\boldsymbol{x},\boldsymbol{y}) + Q^{in,\theta}(\boldsymbol{x},\boldsymbol{y})$	CE	1	1	+ Intrinsic Reward
	$Q^{\theta}(\boldsymbol{x},\boldsymbol{y})$	CE	$\rho > 0$	$\rho > 0$	RL as Inference

"Standard Model" encompasses well-known ML algorithms as special cases

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

Experience type	Experience function f	Divergence $\mathbb{D}$	α	β	Algorithm
Knowledge	$f_{rule}(\boldsymbol{x},\boldsymbol{y})$	CE	1	1	Posterior Regularization (Ganchev et al., 2010)
	$f_{rule}(\boldsymbol{x},\boldsymbol{y})$	CE	$\mathbb{R}$	1	Unified EM (Samdani et al., 2012)



"Standard Model" encompasses well-known ML algorithms as special cases

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

Experience type	Experience function f	Divergence $\mathbb{D}$	α	β	Algorithm
Model	$f_{\text{model}}^{\text{mimicking}}(\boldsymbol{x},\boldsymbol{y};\mathcal{D})$	CE	1	ϵ	Knowledge Distillation (G. Hinton et al., 2015)

"Standard Model" encompasses well-known ML algorithms as special cases

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

Experience type	Experience function f	Divergence $\mathbb{D}$	α	β	Algorithm
Variational	binary classifier	JSD	0	1	Vanilla GAN (Goodfellow et al., 2014)
	discriminator	f-divergence	0	1	f-GAN (Nowozin et al., 2016)
	1-Lipschitz discriminator	W_1 distance	0	1	WGAN (Arjovsky et al., 2017)
	1-Lipschitz discriminator	KL	0	1	PPO-GAN (Y. Wu et al., 2020)

"Standard Model" encompasses well-known ML algorithms as special cases

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

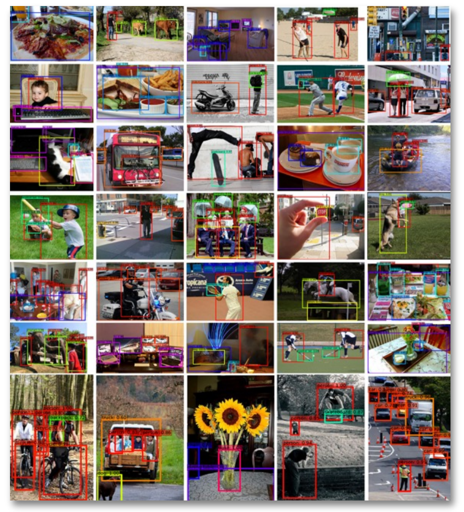
Experience type	Experience function f	Divergence $\mathbb{D}$	α	β	Algorithm
Online	$f_\tau(\boldsymbol{t})$	CE	$\rho > 0$	$\rho > 0$	Multiplicative Weights (Freund & Schapire, 1997)

"Standard Model" encompasses well-known ML algorithms as special cases

Experience type	Experience function f	Divergence $\mathbb{D}$	α	β	Algorithm
Data instances	$f_{\text{data}}(\boldsymbol{x}; \mathcal{D})$	CE	1	1	Unsupervised MLE
	$f_{\text{data}}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D})$	CE	1	ϵ	Supervised MLE
	$f_{\text{data-self}}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D})$	CE	1	ϵ	Self-supervised MLE
	$f_{\text{data-w}}(\boldsymbol{t}; \mathcal{D})$	CE	1	ϵ	Data Re-weighting
	$f_{\text{data-aug}}(\boldsymbol{t}; \mathcal{D})$	CE	1	ϵ	Data Augmentation
	$f_{\text{active}}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D})$	CE	1	ϵ	Active Learning (Ertekin et al., 2007)
Knowledge	$f_{rule}(\boldsymbol{x}, \boldsymbol{y})$	CE	1	1	Posterior Regularization (Ganchev et al., 2010)
	$f_{rule}(\boldsymbol{x}, \boldsymbol{y})$	CE	$\mathbb{R}$	1	Unified EM (Samdani et al., 2012)
Reward	$\log Q^{\theta}(\boldsymbol{x}, \boldsymbol{y})$	CE	1	1	Policy Gradient
	$\log Q^{\theta}(\boldsymbol{x}, \boldsymbol{y}) + Q^{in,\theta}(\boldsymbol{x}, \boldsymbol{y})$	CE	1	1	+ Intrinsic Reward
	$Q^{\theta}(\boldsymbol{x}, \boldsymbol{y})$	CE	$\rho > 0$	$\rho > 0$	RL as Inference
Model	$f_{\text{model}}^{\text{mimicking}}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D})$	CE	1	ϵ	Knowledge Distillation (G. Hinton et al., 2015)
Variational	binary classifier	JSD	0	1	Vanilla GAN (Goodfellow et al., 2014)
	discriminator	f-divergence	0	1	f-GAN (Nowozin et al., 2016)
	1-Lipschitz discriminator	W_1 distance	0	1	WGAN (Arjovsky et al., 2017)
	1-Lipschitz discriminator	KL	0	1	PPO-GAN (Y. Wu et al., 2020)
Online	$f_{\tau}(\boldsymbol{t})$	CE	$\rho > 0$	$\rho > 0$	Multiplicative Weights (Freund & Schapire, 1997)

Applications: "Panoramic" learning with ALL experience

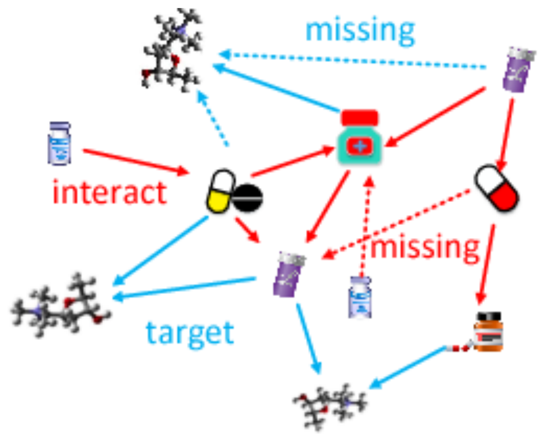
All available experience



Data examples

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

Rules/Constraints



Knowledge graphs

Rewards

Auxiliary agents

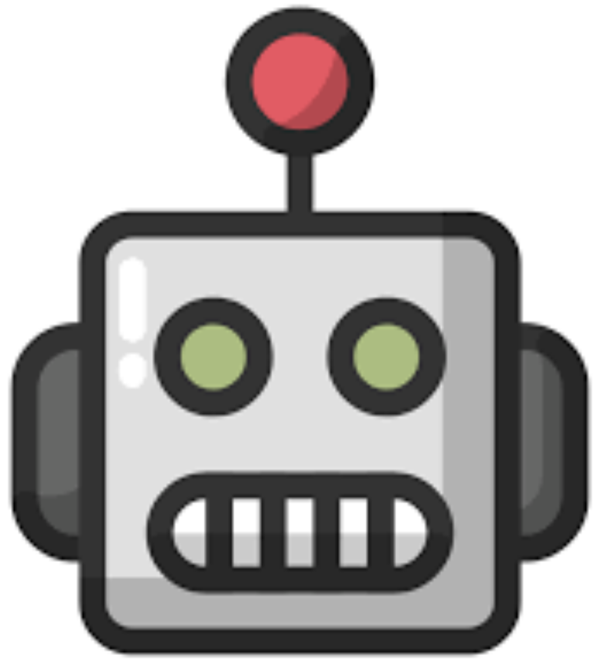
Adversaries

Master classes

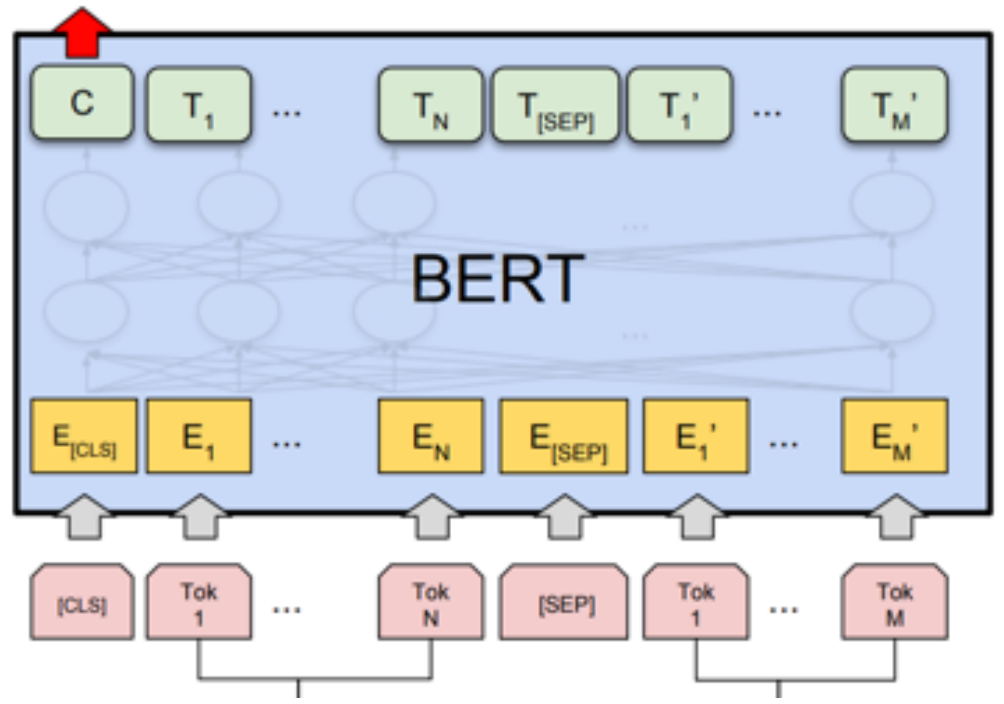
…

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

Arbitrary model



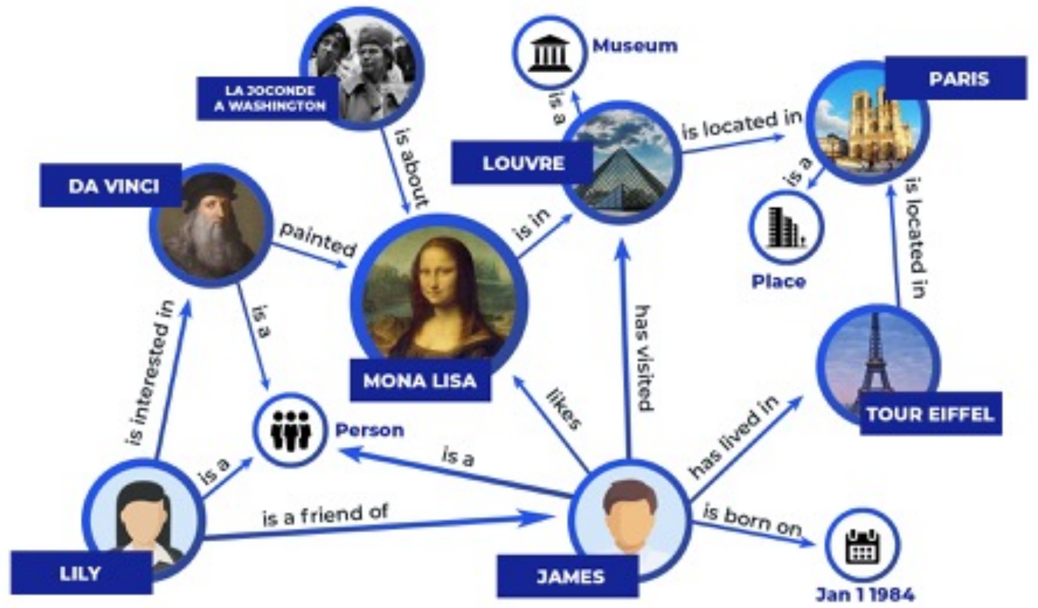
App (1): Using ***symbolic knowledge*** to learn ***neural networks***



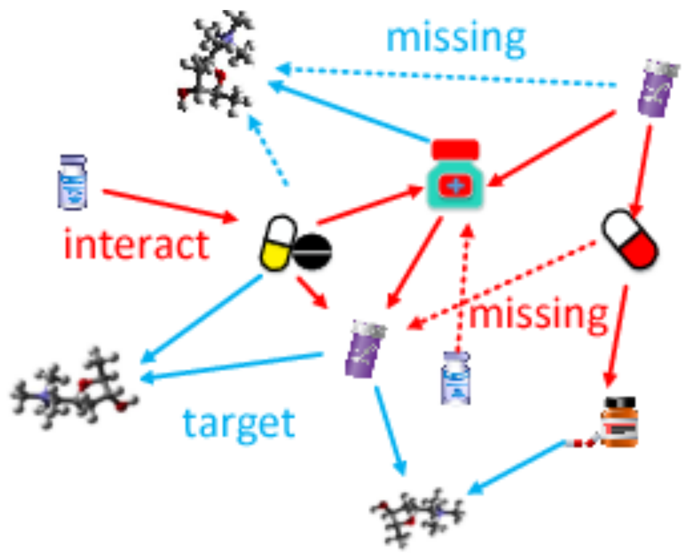
Neural model

WIKIPEDIA
The Free Encyclopedia

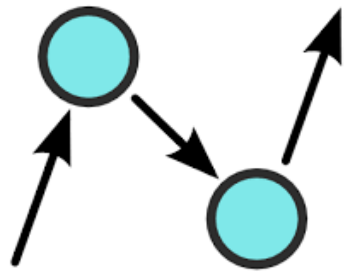
Taxt samples



Knowledge bases



Medical KG



ConceptNet (CN)

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

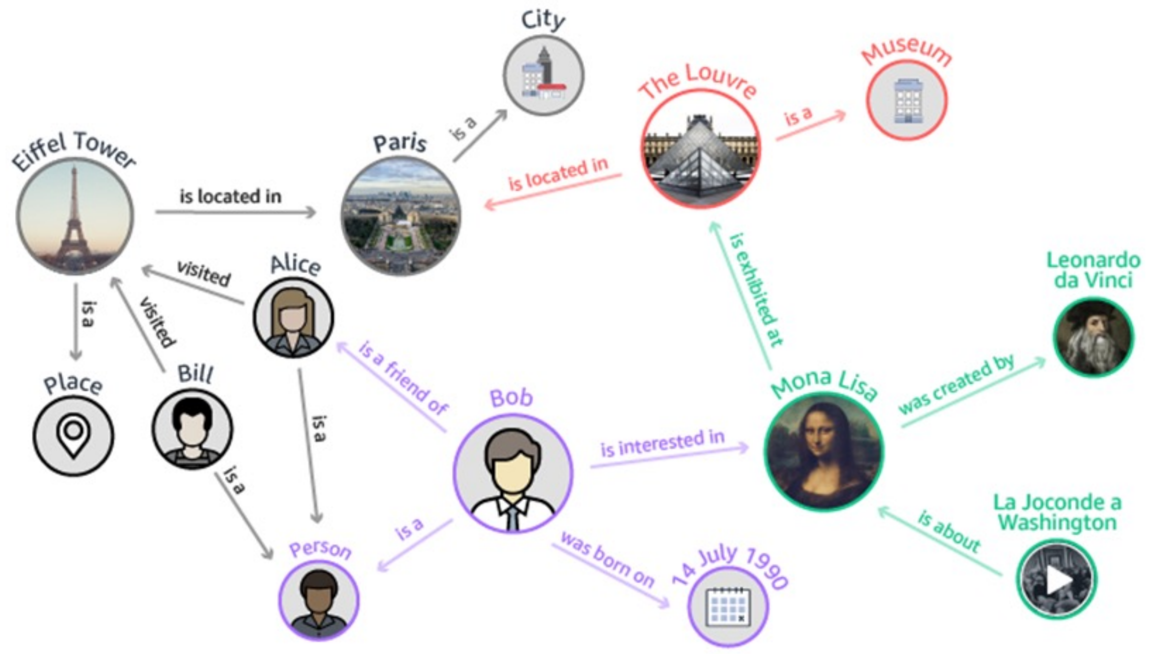
Hu et al., 2016, “Harnessing Deep Neural Networks with Logic Rules”
Hu et al., 2020, “Deep Generative Models with Learnable Knowledge Constraints”
Tan et al., 2020, “Summarizing Text on Any Aspects: A Knowledge-Informed Weakly-Supervised Approach”

App (2): Using ***neural networks*** to “learn” ***symbolic knowledge***

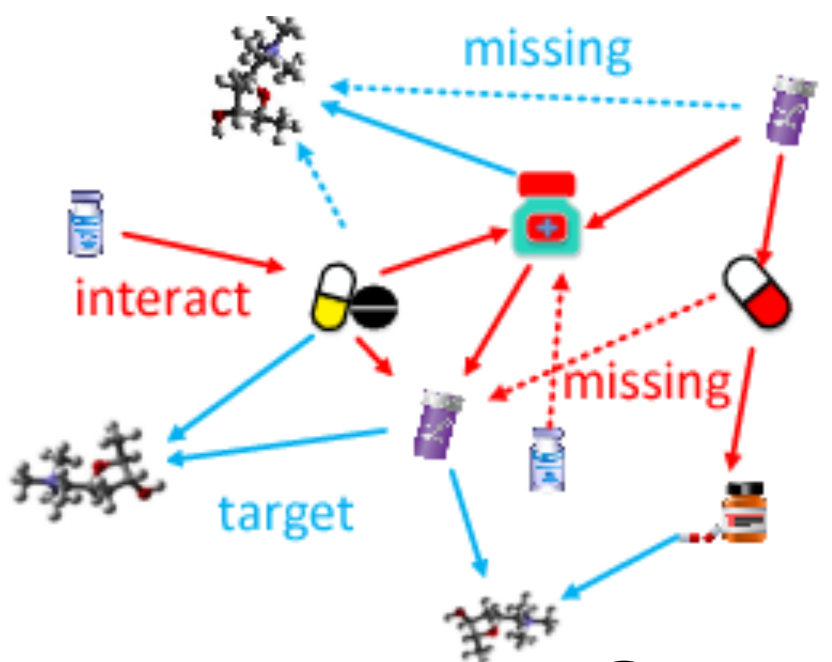
$$\min_{q,\theta} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\Big(q(\boldsymbol{x},\boldsymbol{y}), p_\theta(\boldsymbol{x},\boldsymbol{y})\Big) - \mathbb{E}_{q(\boldsymbol{x},\boldsymbol{y})}\Big[f(\boldsymbol{x},\boldsymbol{y})\Big]$$

- θ: graph structure to be learned
- p_θ: a simulation model generating medical task samples $(\boldsymbol{x},\boldsymbol{y})$ based on the knowledge graph θ

Measuring likelihood of sample $(\boldsymbol{x},\boldsymbol{y})$ under a trained medical neural model



Commonsense graph



Medical KG

Hao, Tan et al., 2022, “BertNet: Harvesting Knowledge Graphs from Pretrained Language Models”

App (2): Using ***neural networks*** to "learn" ***symbolic knowledge***

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

Head entity	Relation	Tail entity	Head entity	Relation	Tail entity
exercise	*prevent*	obesity	students	*worth celebrating*	graduate
apple	*business*	Mac	newborn	*can but not good at*	sit
sleep	*prevent*	illness	social worker	*can help*	foster child
mall	*place for*	shopping	honey	*ingredient for*	honey cake
gym	*place for*	sweat	cabbage	*ingredient for*	cabbage salad
wheat	*source of*	flour	China	*separated by the ocean*	Japan
oil	*source of*	fuel	Africa	*separated by the ocean*	Europe

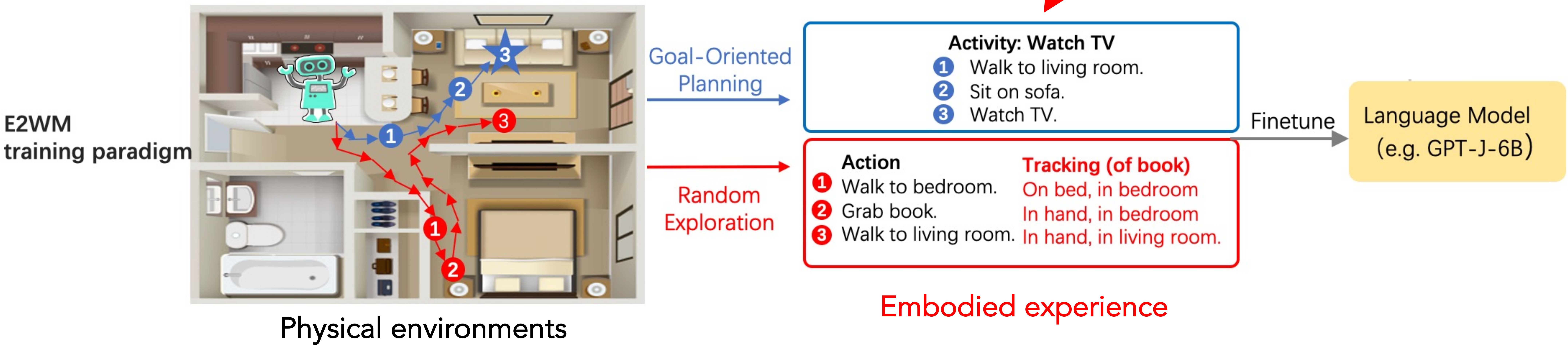
Figure 4: Examples of knowledge tuples harvested from ROBERTA-LARGE with MULTI-PROMPTS.

Hao, Tan et al., 2022, "BertNet: Harvesting Knowledge Graphs from Pretrained Language Models"

App (3): Building ***World Models*** beyond ***Language Models***

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

Language models



Xiang, Tao et al., 2023, “Language Models Meet World Models: Embodied Experiences Enhance Language Models”

App (3): Building ***World Models*** beyond ***Language Models***

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

John put a **book** on the desk.
...
Mary took the **book**.
She placed it on the sofa.
...
Where was the **book**?

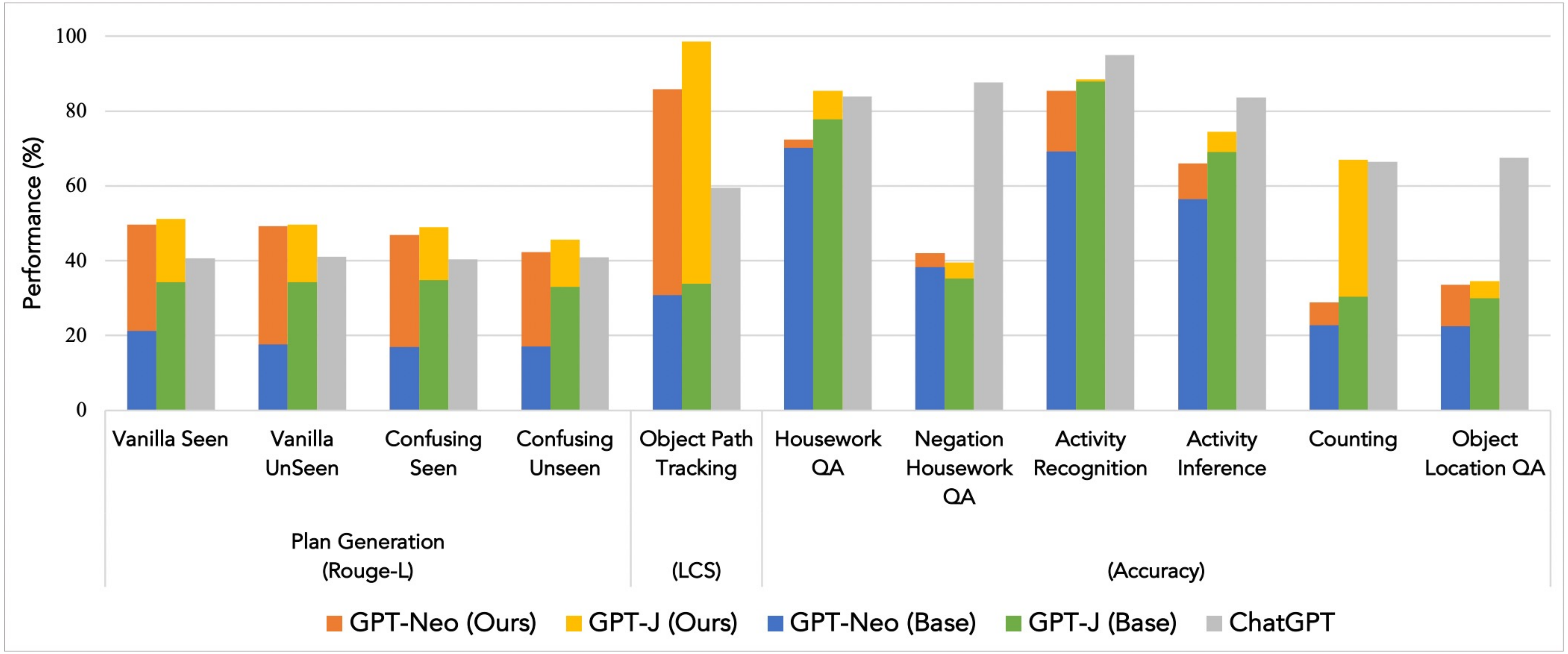
ChatGPT: It was on the desk.

WM (small-size): It was on the sofa.

Xiang, Tao et al., 2023, "Language Models Meet World Models: Embodied Experiences Enhance Language Models"

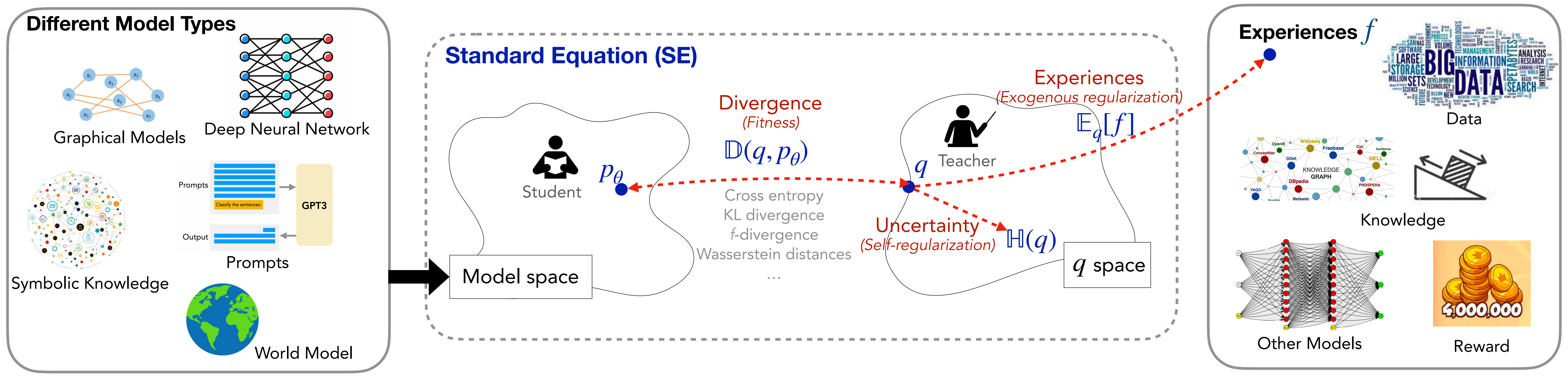
App (3): Building ***World Models*** beyond ***Language Models***

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



Xiang, Tao et al., 2023, "Language Models Meet World Models: Embodied Experiences Enhance Language Models"

Summary



- A "Standard Model" of machine learning

$$\min_{q,\,\theta} - \mathbb{E}_{q(x,y)}\Big[\, f(x\,,y) \,\Big] + \alpha\mathbb{D}\Big(q(x,y),\, p_\theta\,(x,y)\Big) - \beta\mathbb{H}(q)$$

- "Panoramic learning" with ALL experience
  - Neuro-symbolic learning
  - Building world models

Head entity	Relation	Tail entity	Head entity	Relation	Tail entity
exercise	*prevent*	obesity	students	*worth celebrating*	graduate
apple	*business*	Mac	newborn	*can but not good at*	sit
sleep	*prevent*	illness	social worker	*can help*	foster child
mall	*place for*	shopping	honey	*ingredient for*	honey cake
gym	*place for*	sweat	cabbage	*ingredient for*	cabbage salad
wheat	*source of*	flour	China	*separated by the ocean*	Japan
oil	*source of*	fuel	Africa	*separated by the ocean*	Europe

Figure 4: Examples of knowledge tuples harvested from ROBERTA-LARGE with MULTI-PROMPTS.