DSC291: Machine Learning with Few Labels

Overview

Zhiting Hu

Lecture 3, April 5, 2024

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

What is Machine Learning?

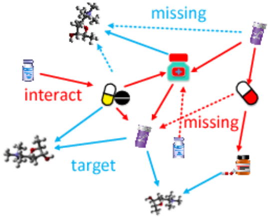
- Computational methods that enable machines to learn concepts and improve performance from **experience**.

Experience of all kinds

Data examples

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

Rules/Constraints

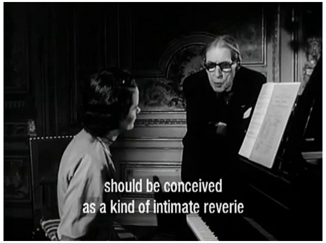


Knowledge graphs

Rewards

Auxiliary agents

Adversaries

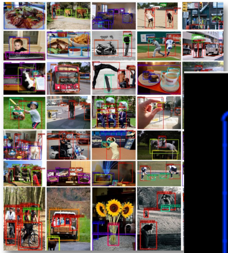


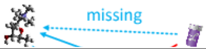
Master classes

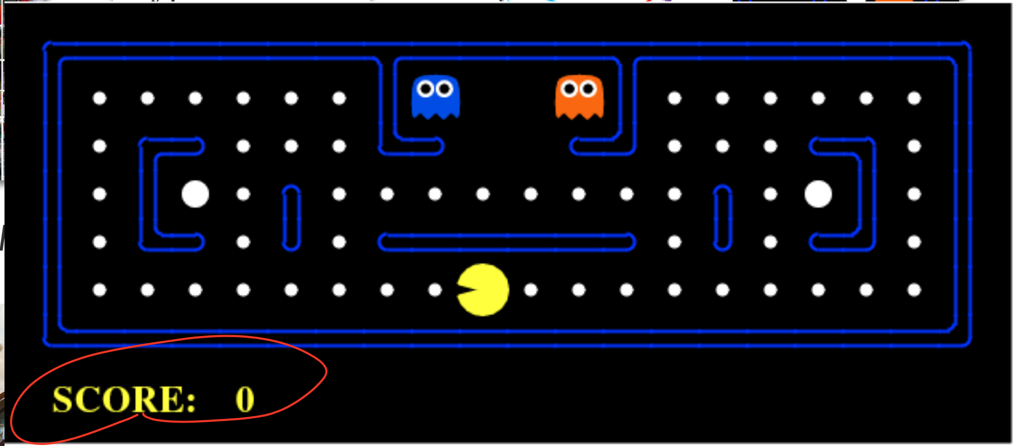
…

And all combinations thereof

Experience of all kinds





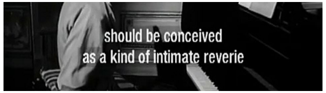


Data exampl

Auxiliary agents

ations thereof

Adversaries



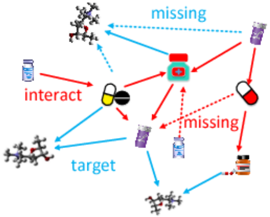
Master classes

Experience of all kinds

Data examples

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

Rules/Constraints

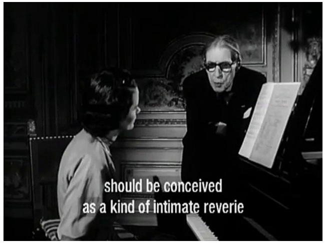


Knowledge graphs

Rewards

Auxiliary agents

Adversaries



Master classes

…

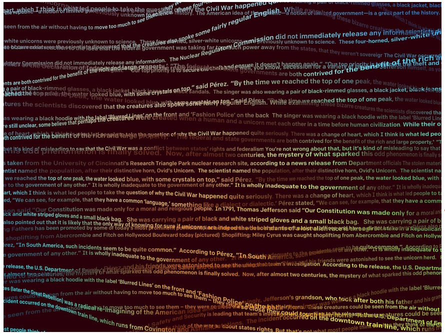
And all combinations thereof

AlexNet NeurIPS 2012

Experience: (massive) data examples

Image classification

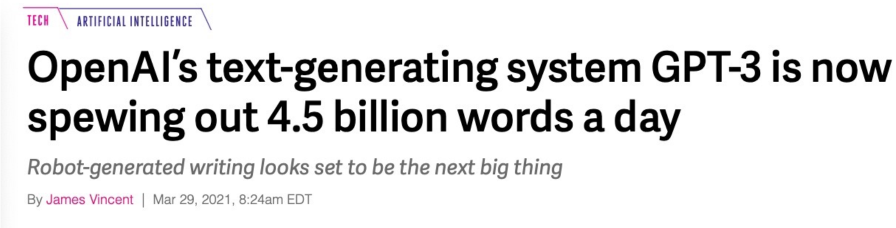
Machine translation



Language modeling
(BERT, GPT-3/4, …)

GPT3: 45TB of text data: CommonCrawl, WebText, Wikipedia, corpus of books, …

Experience: (massive) data examples

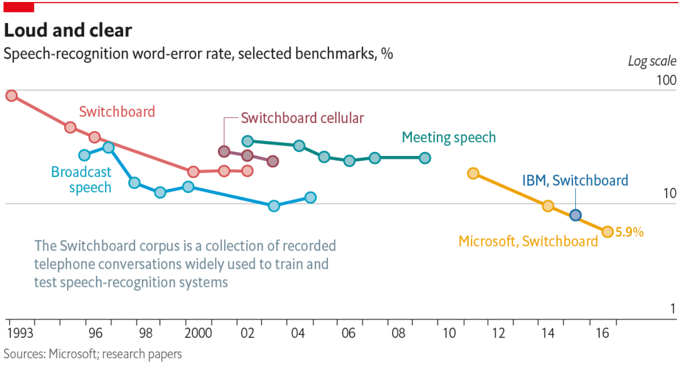


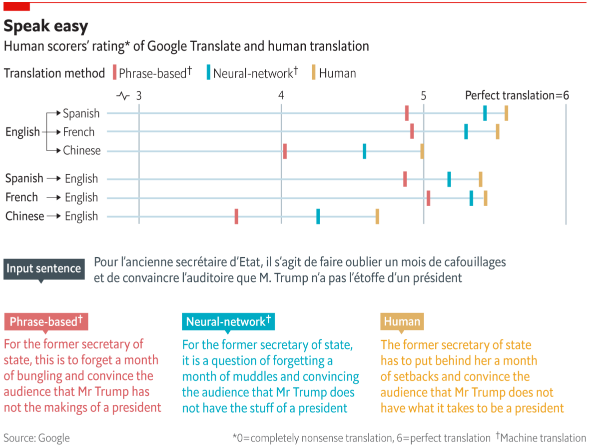
TECH \ ARTIFICIAL INTELLIGENCE

OpenAI's text-generating system GPT-3 is now spewing out 4.5 billion words a day

Robot-generated writing looks set to be the next big thing

By James Vincent | Mar 29, 2021, 8:24am EDT





[The Economist]

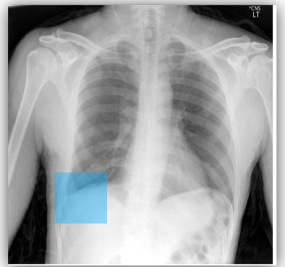
Problems with few data (labels)

- Privacy, security issues

Assistive diagnosis

Normal findings

Abnormal findings



``The heart size and mediastinal contours appear within normal limits. There is blunting of the right lateral costophrenic sulcus which could be secondary to a small effusion versus scarring …''

Problems with few data (labels)

- Expensive to collect/annotate

Robotic control

Problems with few data (labels)

- Expensive to collect/annotate

  Controllable content generation

Source image

Generated images under different poses

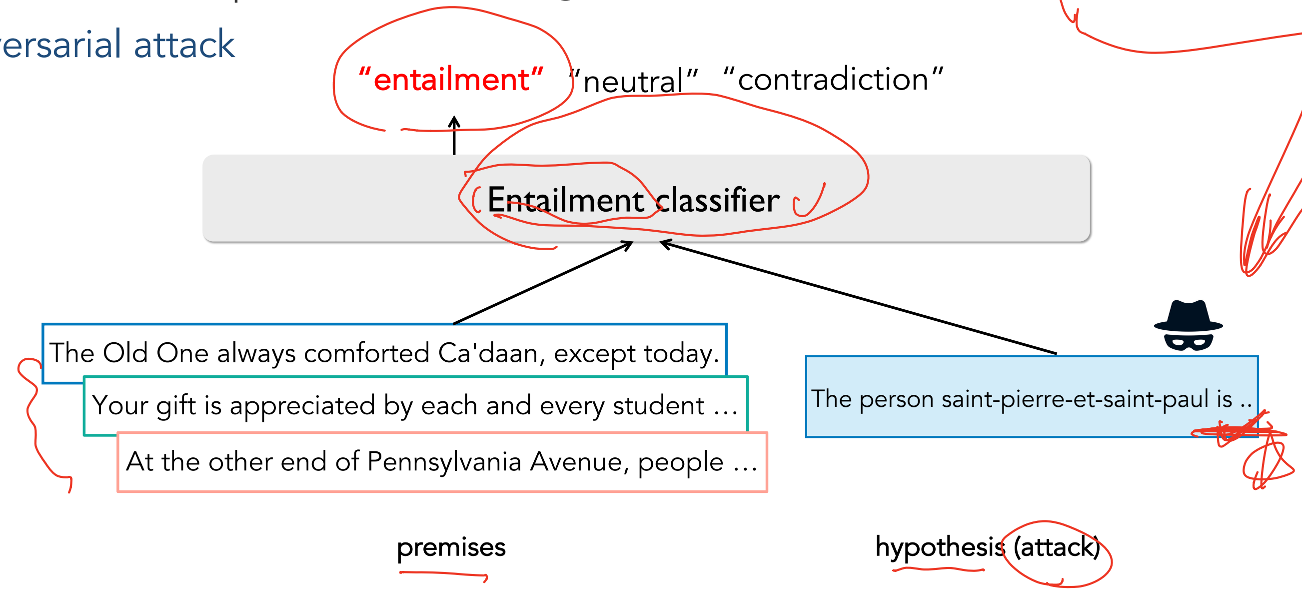
Applications: virtual clothing try-on system

Problems with few data (labels)

attack generating model

- Difficult / expertise-demanding to annotate

Adversarial attack

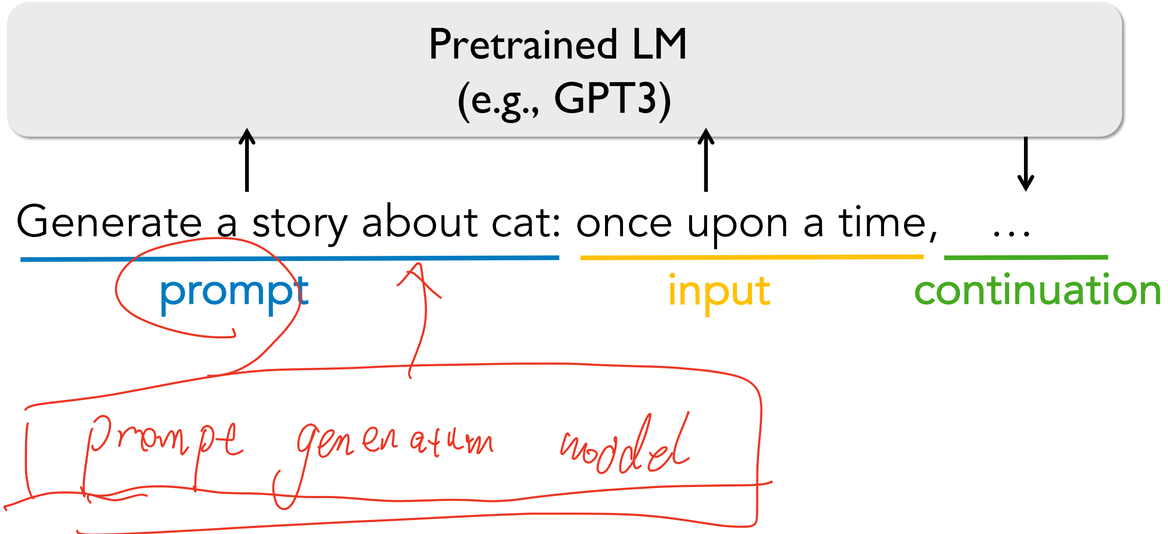


Applications: test model robustness

Problems with few data (labels)

- Difficult / expertise-demanding to annotate

Prompt generation: automatically generating prompts to steer pretrained LMs

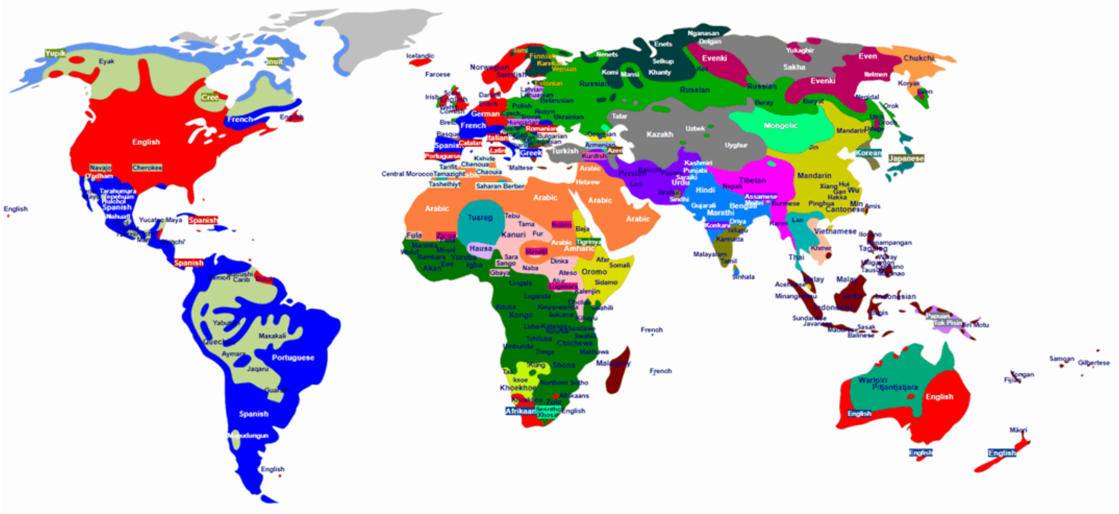


Problems with few data (labels)

- Specific domain

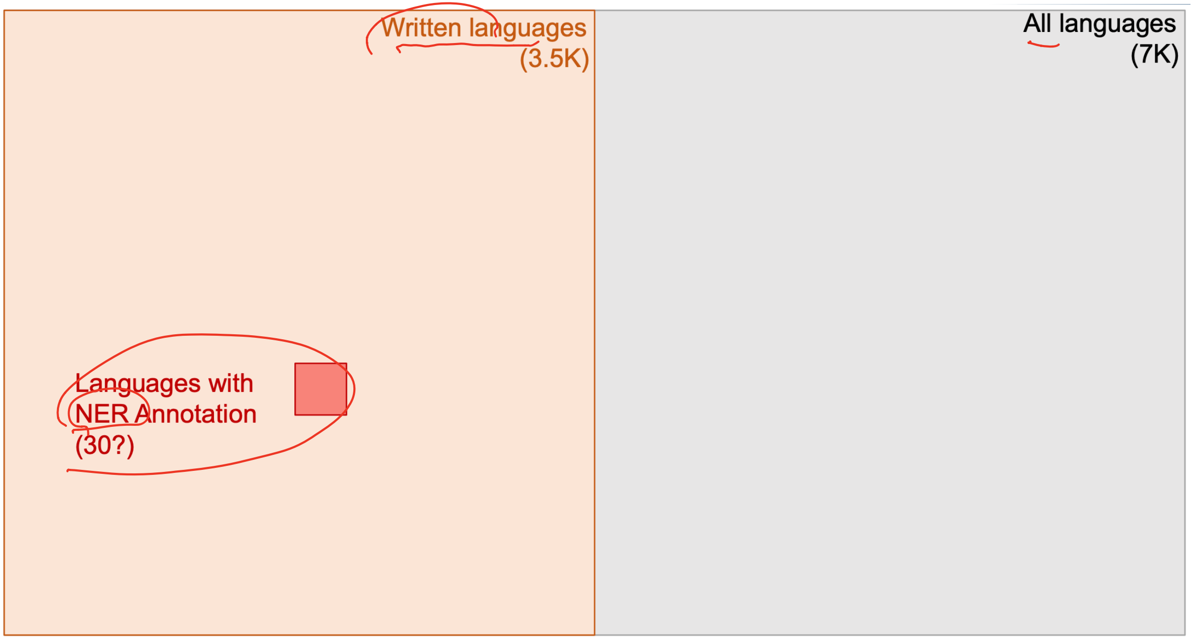
Low-resource languages

~7K languages in the world



Problems with few data (labels)

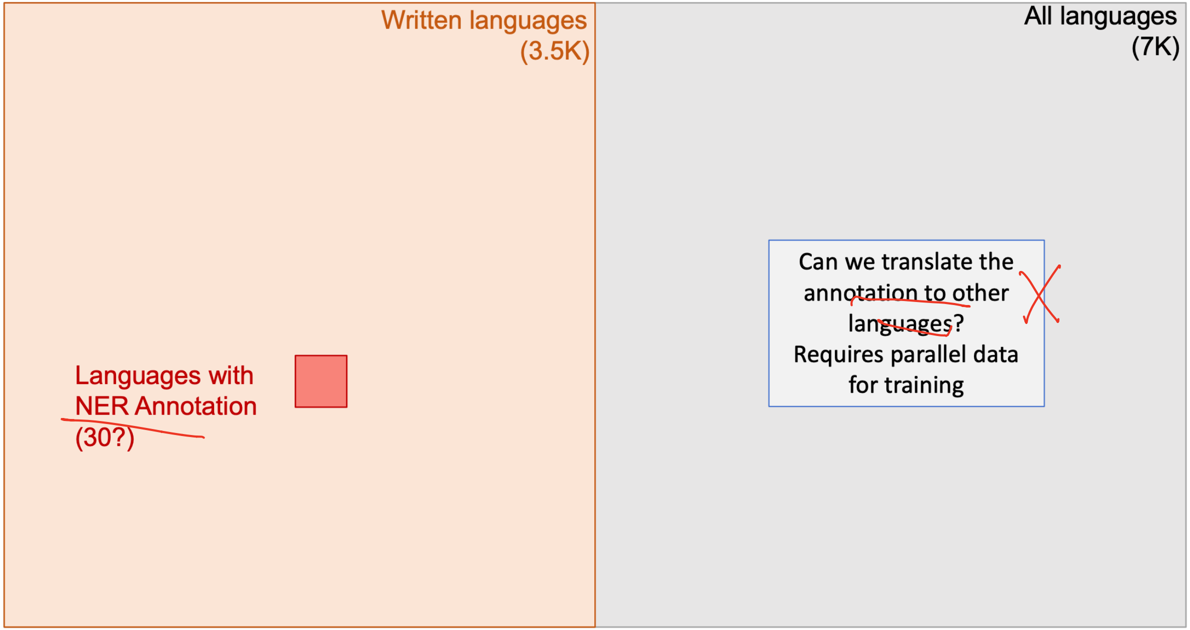
- Specific domain   Low-resource languages



[Figure courtesy: Dan Roth, CIS620]

Problems with few data (labels)

- Specific domain   Low-resource languages



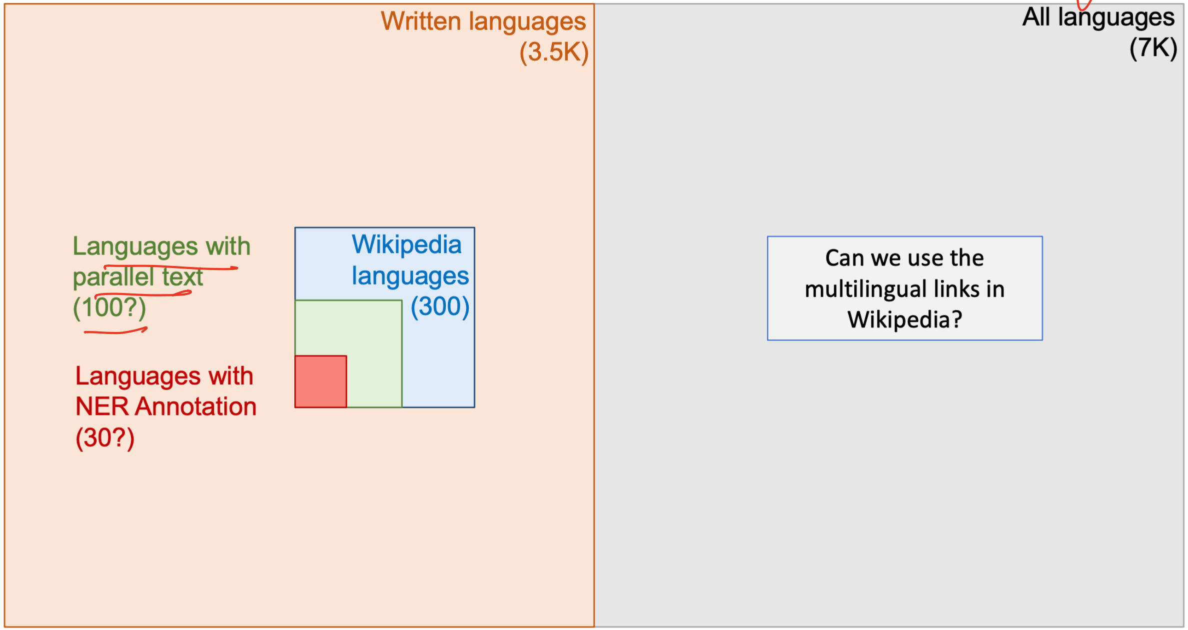
[Figure courtesy: Dan Roth, CIS620]

Problems with few data (labels)

Linguistics

⇒ language documentation

- Specific domain   Low-resource languages



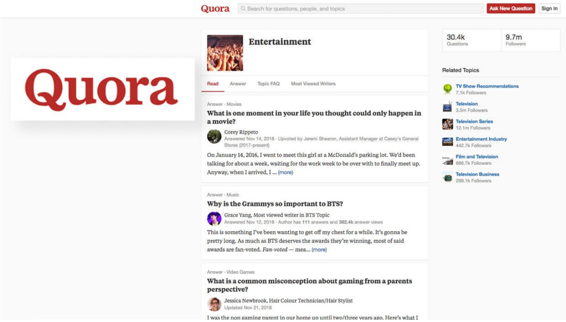
[Figure courtesy: Dan Roth, CIS620]

Problems with few data (labels)

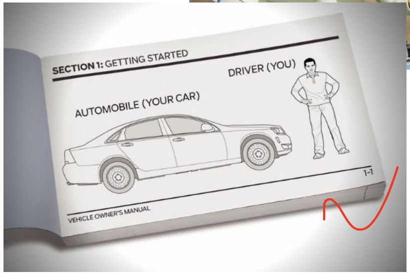
- Specific domain

Question answering

QA based on car manual?



reddit



Problems with few data (labels)

- Privacy, security issues
- Expensive to collect/annotate
- Difficult / expertise-demanding to annotate
- Specific domain

Machine learning solutions given few data (labels)

- How can we make more efficient use of data?
  - Clean but small-size
  - Noisy
  - Out-of-domain

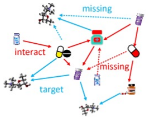
horse zebra

- Can we incorporate other types of experience in 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 thereof

Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

$$\min_\theta \mathcal{L}(\theta, \mathcal{E})$$

Optimization solver   Loss   Model architecture   Experience

GD → SGD, ADAM, ...

evolutionary algorithms

Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

This course discusses very little about model architecture

$$\min_{\theta} \mathcal{L}(\theta, \mathcal{E})$$

Optimization solver

Loss

Model architecture

Experience

Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

This course discusses very little about model architecture

Model of certain architecture whose parameters are the subject to be learned, $p_\theta(\boldsymbol{x}, \boldsymbol{y})$ or $p_\theta(\boldsymbol{y}|\boldsymbol{x})$

- Neural networks
- Graphical models
- Compositional architectures

θ

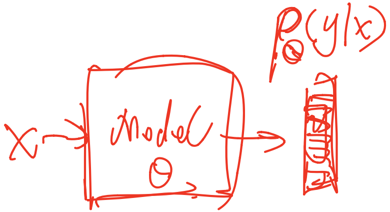
Components of a ML solution (roughly)

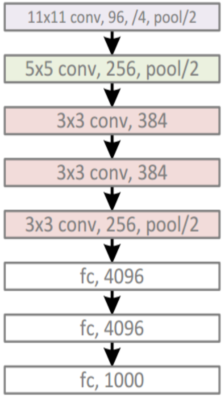
- Loss
- Experience
- Optimization solver
- Model architecture

This course discusses very little about model architecture

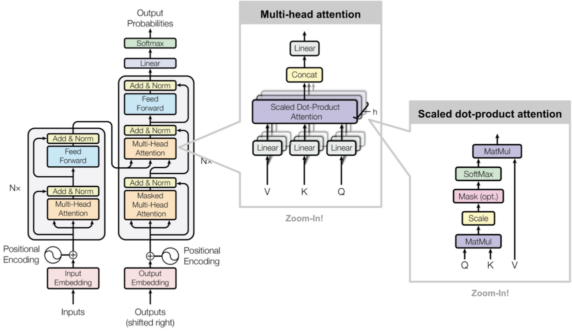
Model of certain architecture whose parameters are the subject to be learned, $p_\theta(\boldsymbol{x}, \boldsymbol{y})$ or $p_\theta(\boldsymbol{y}|\boldsymbol{x})$

- Neural networks
- Graphical models
- Compositional architectures





Convolutional networks



Transformers

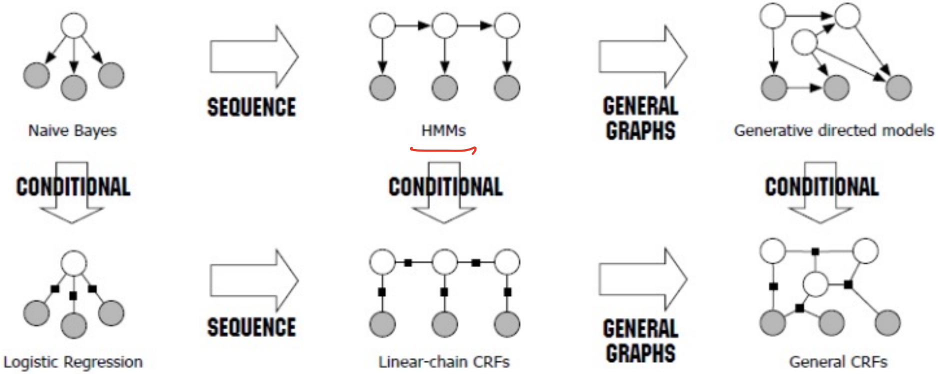
Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

This course discusses very little about model architecture

Model of certain architecture whose parameters are the subject to be learned, $p_\theta(\boldsymbol{x}, \boldsymbol{y})$ or $p_\theta(\boldsymbol{y}|\boldsymbol{x})$

- Neural networks
- Graphical models
- Compositional architectures



Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

This course discusses very little about model architecture

Assuming you know basic procedures:

- (Stochastic) gradient descent
- Backpropagation
- Lagrange multiplier
- …

constrained opt.

$$\min_\theta \mathcal{L}(\theta, \mathcal{E})$$

Optimization solver

Loss

Model architecture

Experience

Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

This course discusses very little about model architecture

Core of most learning algorithms

model

$$\min_\theta \mathcal{L}(\theta, \mathcal{E})$$

Optimization solver

Loss

Model architecture

Experience

GAN

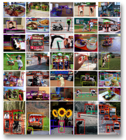
Gen. Adv. network

Algorith...

Machine learning solutions given few data (labels)

- (1) How can we make more efficient use of data?
  - Clean but small-size, Noisy, Out-of-domain

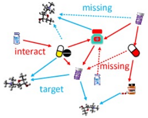
- (2) Can we incorporate other types of experience in learning?



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 thereof

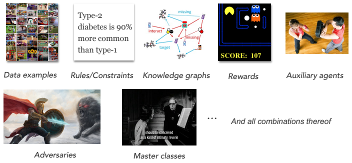
Machine learning solutions given few data (labels)

- (1) How can we make more efficient use of data?
  - Clean but small-size, Noisy, Out-of-domain, …
- Algorithms
  - **Supervised learning**: MLE, maximum entropy principle
  - **Unsupervised learning**: EM, variational inference, VAEs
  - **Self-supervised learning**: successful instances, e.g., BERT, GPT-3, contrastive learning, applications to downstream tasks
  - **Distant/weakly supervised learning**: successful instances
  - **Data manipulation:** augmentation, re-weighting, curriculum learning, …
  - **Meta-learning**

Mostly first half of the course

Machine learning solutions given few data (labels)

- (2) Can we incorporate other types of experience in learning?



  - Learning from auxiliary models, e.g., adversarial models:
    - Generative adversarial learning (GANs and variants), co-training, …
  - Learning from structured knowledge
    - Posterior regularization, constraint-driven learning, …
  - Learning from rewards
    - Reinforcement learning: model-free vs model-based, policy-based vs value-based, on-policy vs off-policy, extrinsic reward vs intrinsic reward, …

  - Learning in dynamic environment *(not covered)*
    - Online learning, lifelong/continual learning, …

Second half of the course

Algorithm marketplace

Designs driven by: experience, task, loss function, training procedure …

maximum likelihood estimation

reinforcement learning as inference

data re-weighting

inverse RL

active learning

policy optimization

data augmentation

reward-augmented maximum likelihood

label smoothing

imitation learning

softmax policy gradient

actor-critic

adversarial domain adaptation

GANs

posterior regularization

knowledge distillation

intrinsic reward

constraint-driven learning

prediction minimization

generalized expectation

regularized Bayes

learning from measurements

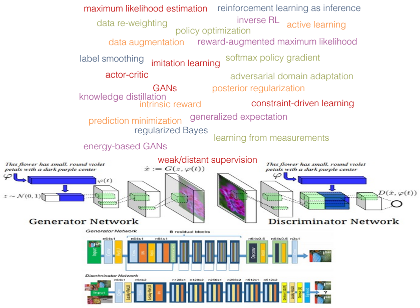
energy-based GANs

weak/distant supervision

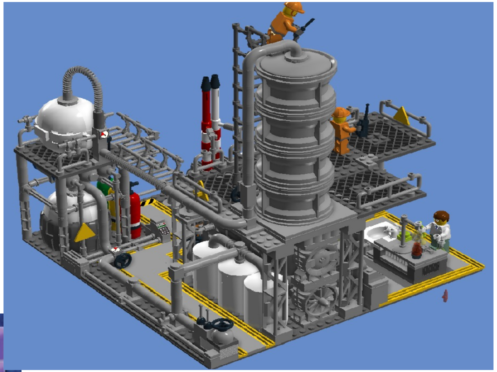
Where we are now? Where we want to be?

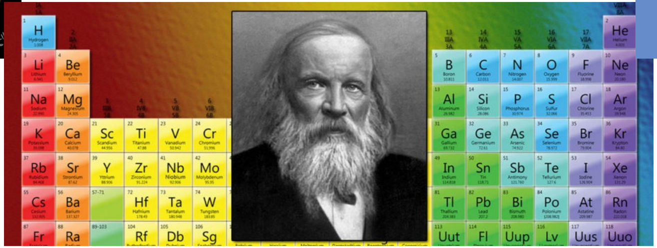
- Alchemy vs chemistry

$\approx$



$\neq\neq$





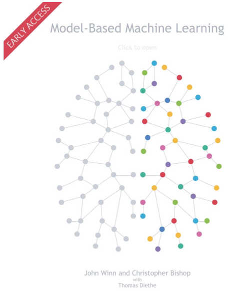
Quest for more standardized, unified ML principles

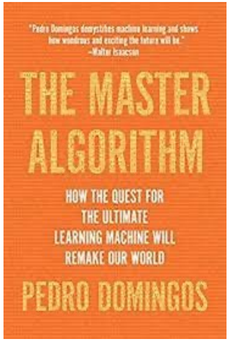
Machine Learning 3: 253–259, 1989
© 1989 Kluwer Academic Publishers – Manufactured in The Netherlands

EDITORIAL

Toward a Unified Science of Machine Learning

[P. Langley, 1989]





REVIEW Communicated by Steven Nowlan

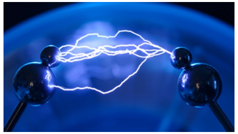
A Unifying Review of Linear Gaussian Models

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

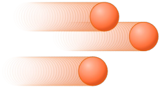
Zoubin Ghahramani*
Department of Computer Science, University of Toronto, Toronto, Canada

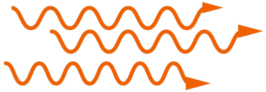
Physics in the 1800's

- 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 equations” 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
$P=-\xi p \quad Q=-\xi q \quad R=-\xi r$		Ohm’s Law
$P=kf \quad Q=kg \quad R=kh$		The electric elasticity equation (**E** = **D**/ε)
$\frac{de}{dt}+\frac{dp}{dx}+\frac{dq}{dy}+\frac{dr}{dz}=0$		Continuity of charge

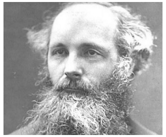
Maxwell’s Eqns simplified w/ rotational symmetry

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

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

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

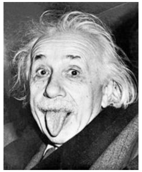
$$\nabla\times\mathbf{H}=\frac{\partial\mathbf{D}}{\partial t}+\mathbf{J}$$



Maxwell’s Eqns further simplified w/ symmetry of special relativity

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

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

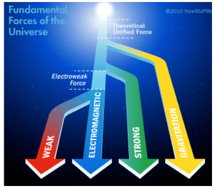


Standard Model w/ Yang-Mills theory and US(3) symmetry

$$\mathcal{L}_{\mathrm{gf}}=-\frac{1}{2}\mathrm{Tr}(F^2)$$

$$=-\frac{1}{4}F^{a\mu\nu}F^a_{\mu\nu}$$

Unification of fundamental forces?



1861 1910s 1970s ?

A “standardized formalism” of ML

Data examples

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

Constraints

Rewards

Auxiliary agents

Adversaries

Imitation

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

Uncertainty Divergence Experience

- Panoramically learn from all types of experience
- Subsumes many existing algorithms as special cases

Will discuss in later in the class

Questions?