

Intelligence artificielle: succès empiriques et défis mathématiques

LÉON BOTTOU

META AI RESEARCH

Sommaire

1. Perceptrons
2. Beyond statistics
3. The infinite library

1

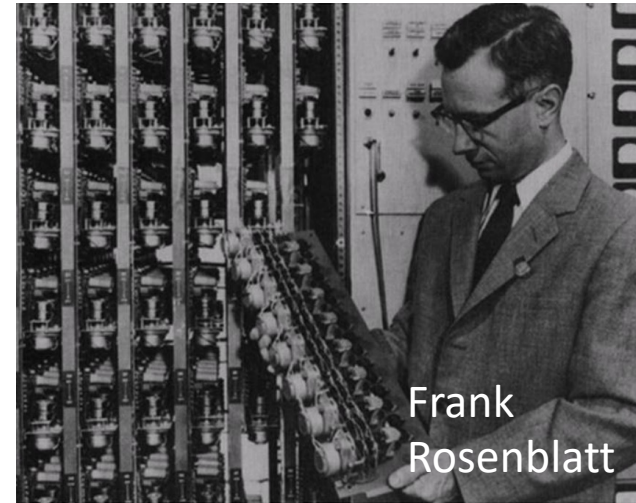
Perceptrons

HOW TO UNDERSTAND A LEARNING MACHINE?

Two computing paradigms in the XXth



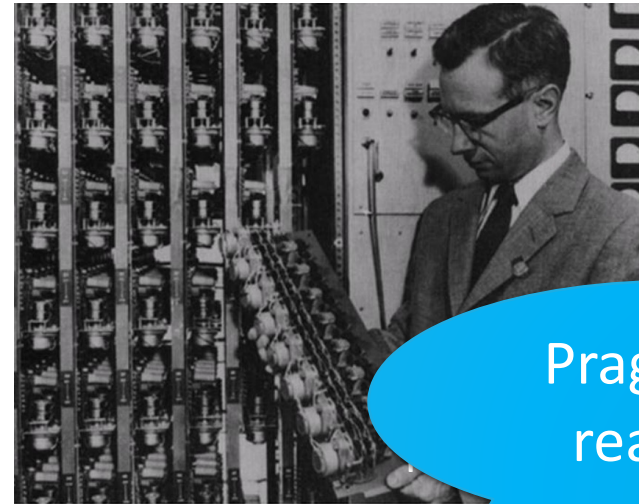
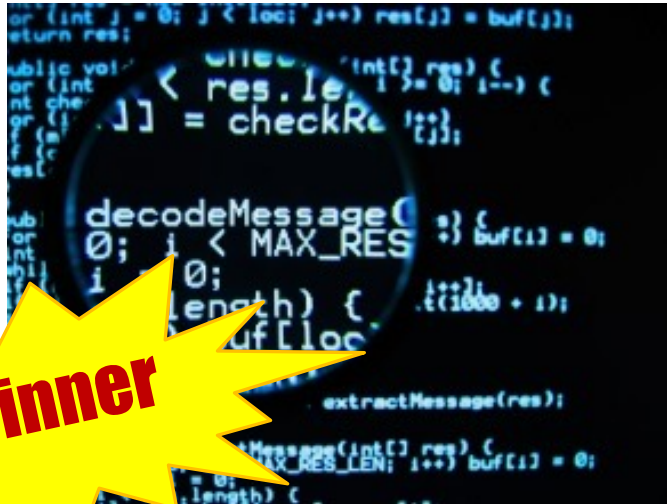
Turing, von Neuman,
McCarthy, Minsky, ...



Frank
Rosenblatt

Wiener, Rosenblatt,
Widrow, Kohonen, ...

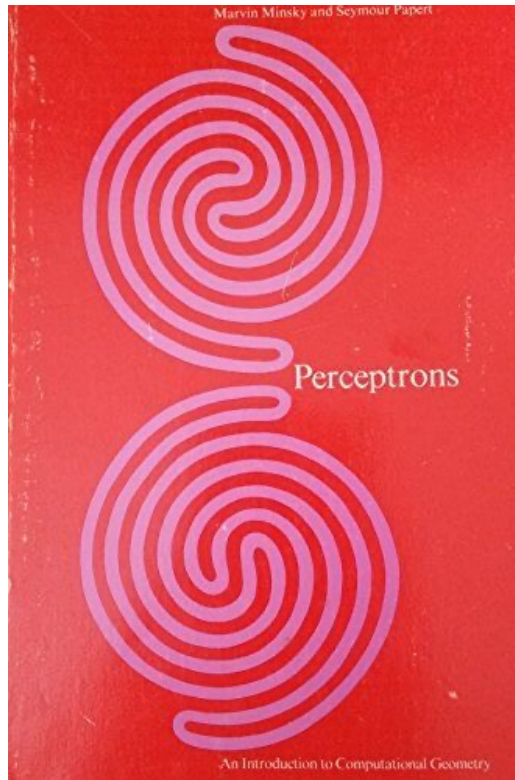
Two computing paradigms in the XXth



Pragmatic
reasons

- Digital computers (the kind one programs) could do immensely useful things: scientific calculus, accounting, etc.
- Machines were not fast enough to make Perceptrons useful!

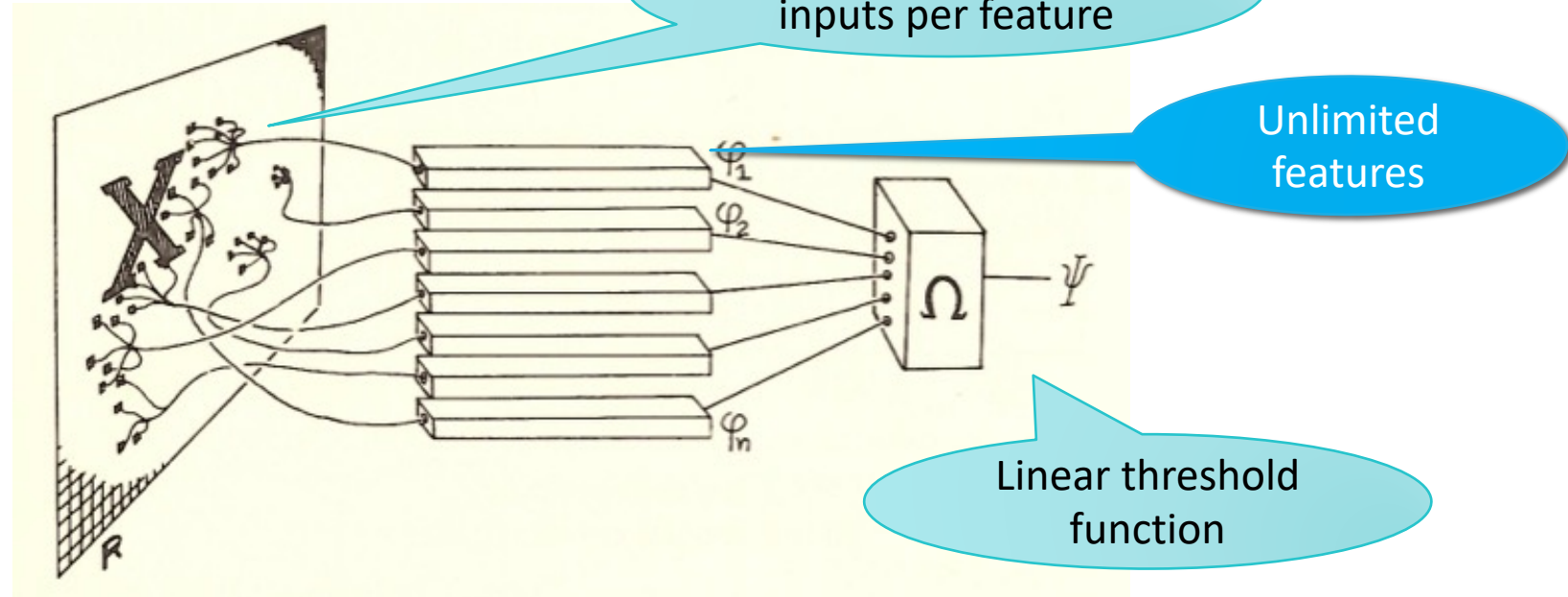
Perceptrons (1969)



Perceptrons by Minsky and Papert (1969)

- How do we decide whether something works?
- They have a clear position.

Order-k Perceptron



Metaphor for parallel computation

- Targets what can be computed by Rosenblatt's perceptron.
- Also describes what can be computed by a convnet with a final pooling layer.
- Similar techniques could characterize what can be computed with map/reduce

Perceptrons: A method.

Take simple Boolean predicates and establish their order requirements.

Focus on group invariant predicates:

- Parity has “infinite” order.

Geometrical predicates:

- Connectedness has infinite order.
- Euler number has low order.
- Etc. with caveats

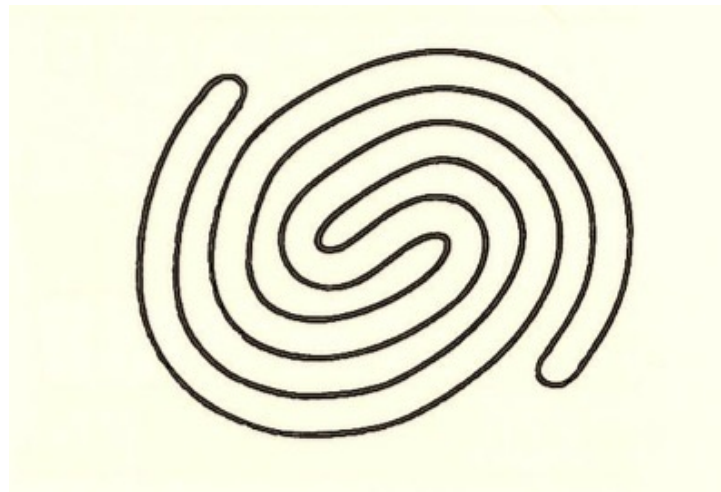


This part is brilliant!

Perceptrons: impossibility theorem(s)

A predicate: "Connectedness"

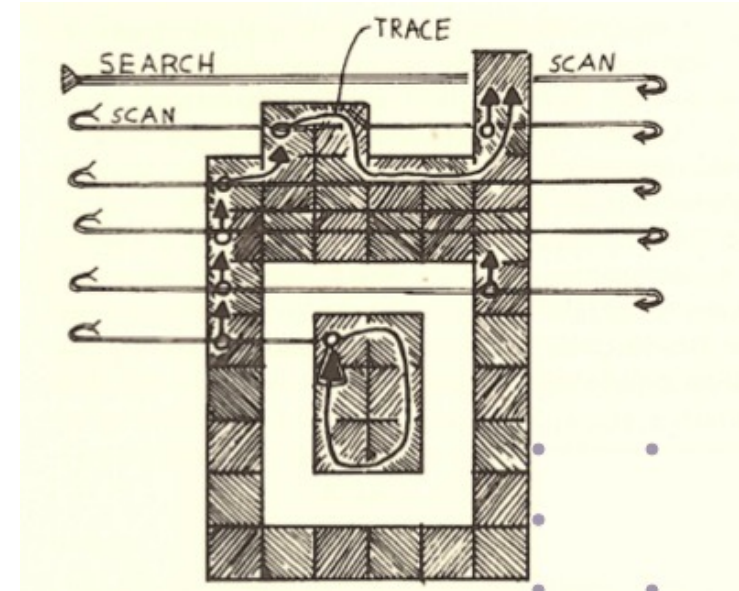
Is a shape made of a single connected component?



Theorem: No small-order perceptron can say.

Perceptrons: possibility theorem(s)

Theorem:
A simple program can
provably compute “connectedness”



Theorem 9.2: For any ϵ there is a 2-symbol Turing machine that can verify the connectedness of a figure X on any rectangular array R , using less than $(2 + \epsilon) \log_2 |R|$ squares of tape.

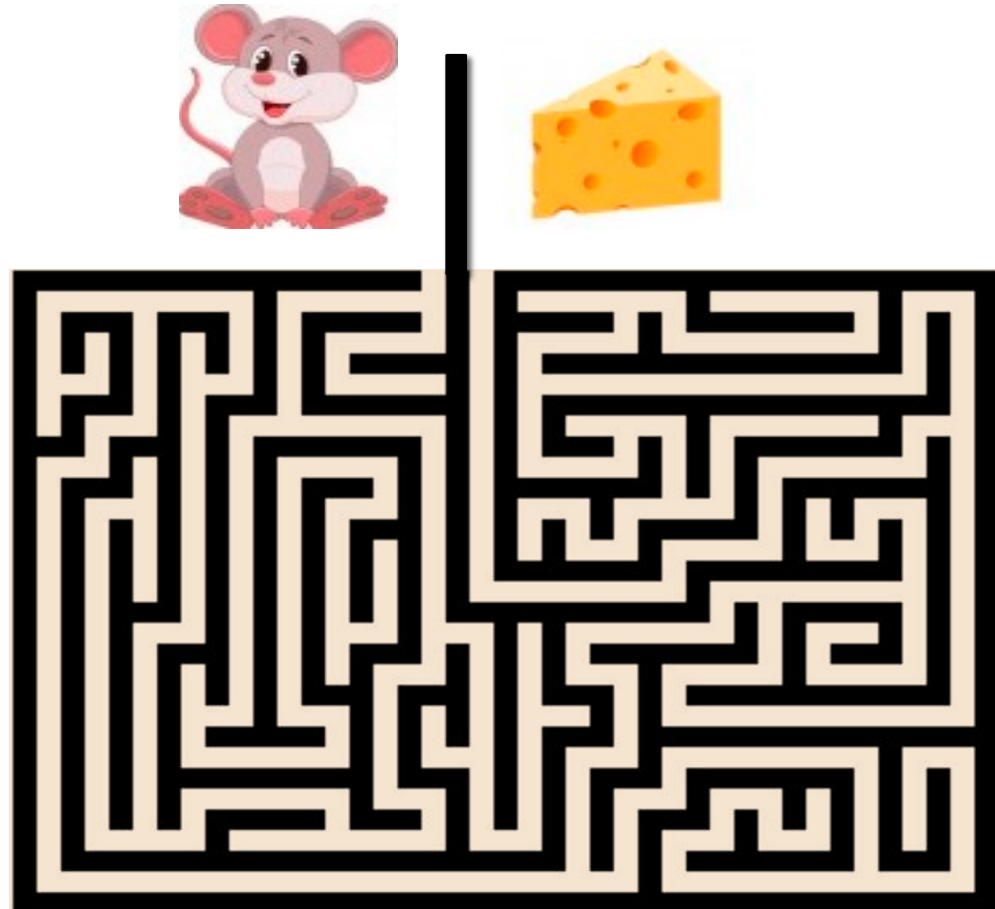
Perceptrons: and that's it!

13.5 Why Prove Theorems?

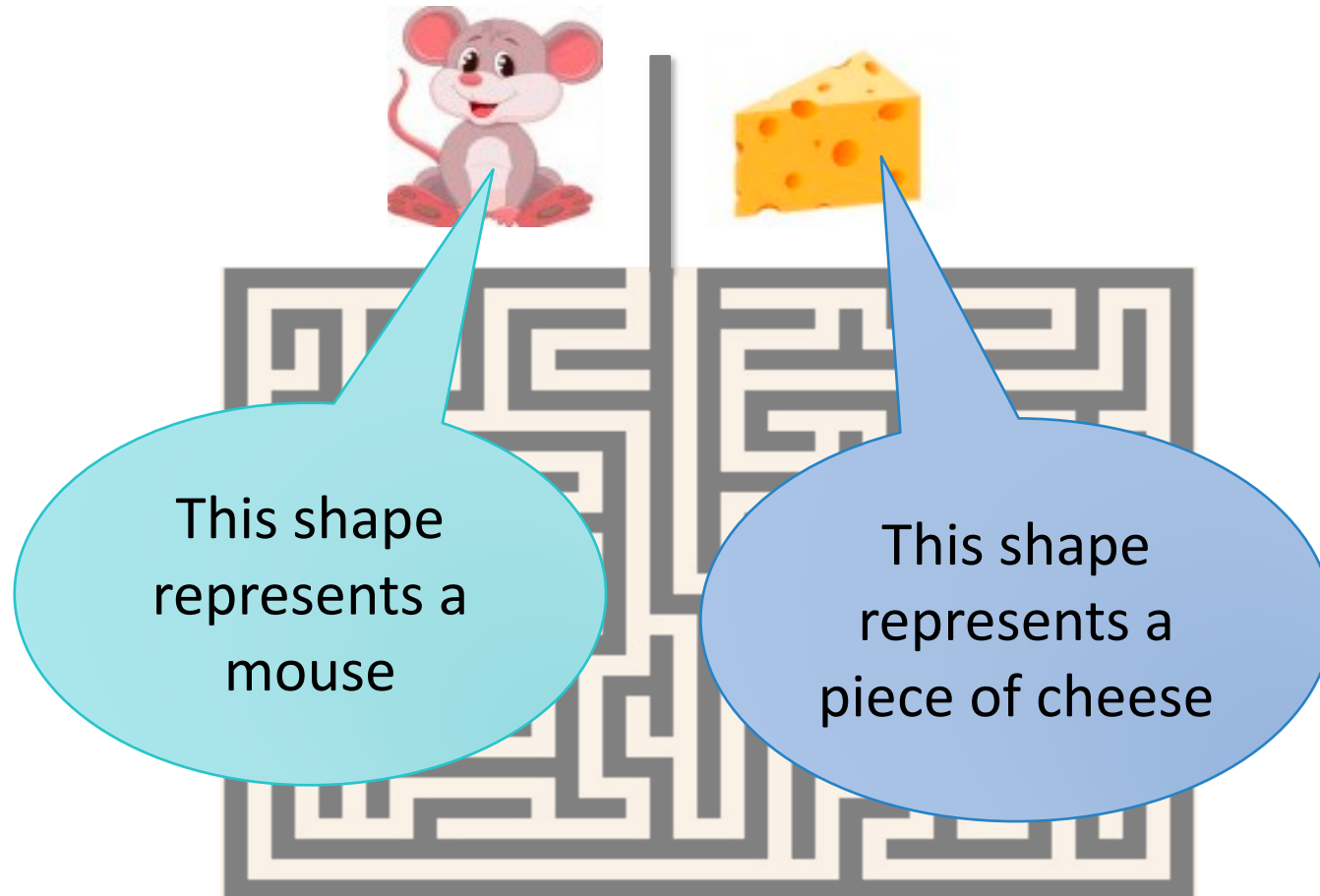
Why did you prove all these complicated theorems? Couldn't you just take a perceptron and see if it can recognize $\psi_{\text{CONNECTED}}$?

No.

Is connectedness easy for us?



What is easy for us?



Are there provable algorithms for



“Mouseness?”



“Cheesiness?”

- “Connectedness” has a clear and compact mathematical specification.
- “Mousiness” and “cheesiness” do not.

Whether we mathematically understand the method does not help us prove anything because we do not have a mathematical specification of the task.

Perceptrons: weak spot

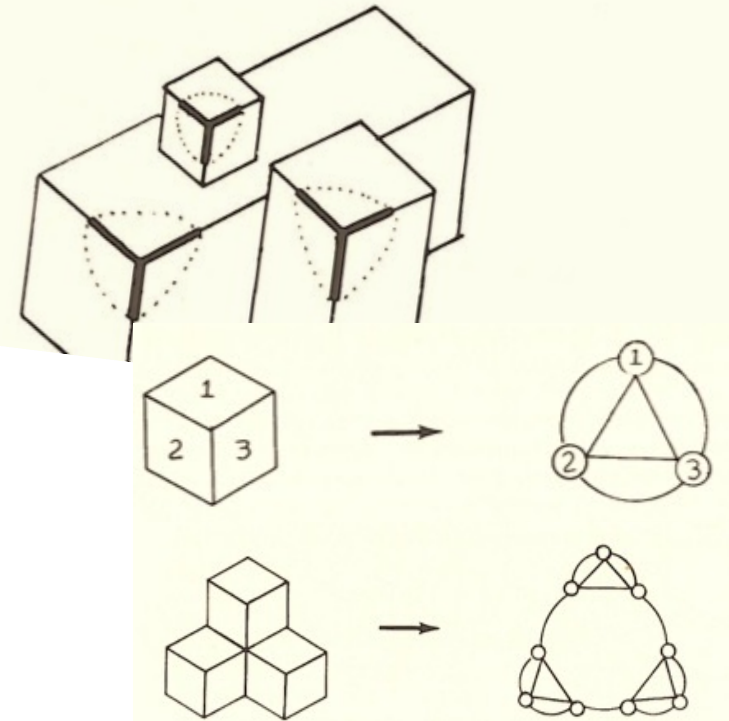
13.3 Analyzing Real-World Scenes

One can understand why you, as mathematicians, would be interested in such clear and simple predicates as ψ_{PARITY} and $\psi_{\text{CONNECTED}}$. But what if one wants to build machines to recognize chairs and tables or people? Do your abstract predicates have any relevance to such problems, and does the theory of the simple perceptron have any relevance to the more complex machines one would use in practice?

This is a little like asking whether the theory of linear circuits has relevance to the design of television sets. Absolutely, some concept of connectedness is required for analyzing a scene with many objects in it. For the whole is just the sum of its parts and

13.4 Guzman's Approach to Scene-Analysis

In scenes like this,



The taxi driver analogy



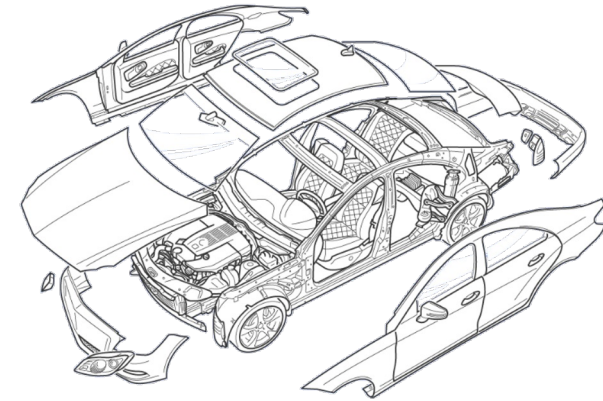
How do we trust the taxi driver ?

- Proving the function
 - open brain, check wiring, ...
- Past performance:
 - driving license (a performance test)
 - taxi license (a knowledge test)
 - police record (past performance)

How to trust the operation of a system?

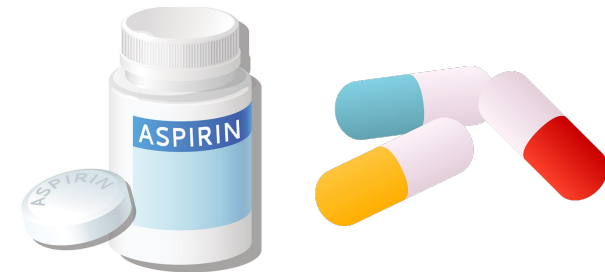
- The “**positivist**” approach

- obtain a mathematic model of the internal mechanisms.
- use this knowledge to prove the function against its specification.

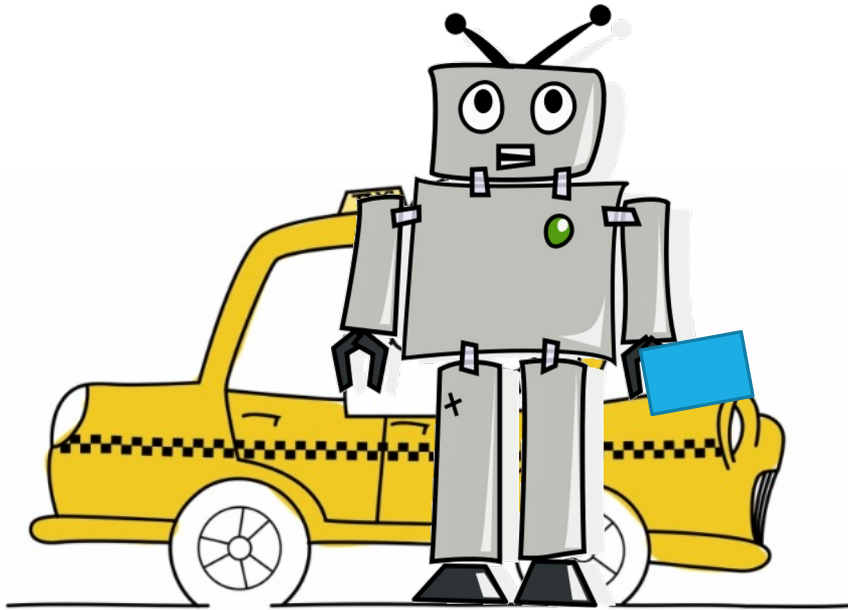


- The “**behaviorist**” approach

- record the past performance of the system
- use the past performance to obtain statistical guarantees.



Return to the taxi driver analogy

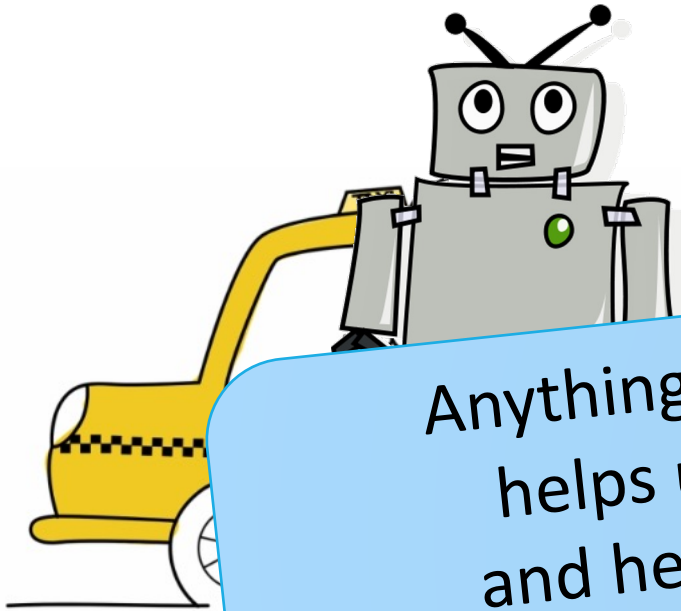


Trusting the past performance of a robot taxi driver ...

- Is it the same for a human and a robot?
- We interpret the human driver record because we know a lot about humans.

For instance, if the record shows excessive beer intake, we know what it means for a human, not a robot...

Knowledge is always good!

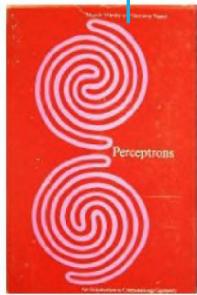


Trusting the past performance
of a robot taxi driver ...

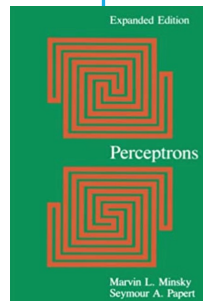
Anything we know about how a system works helps us understand its past performance and helps us drawing the right conclusions.

“Nothing more practical than a good theory” (Vapnik)

Perceptrons, act II.



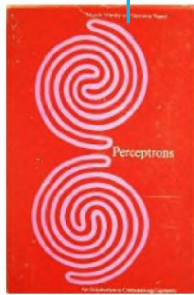
Perceptrons
Minsky & Papert
1968



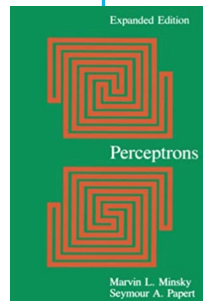
We must first examine [...] two branches of theory which could be called “theory of learning” and “theory of representation.”

[...] we really need to know a great deal more about the nature of those surfaces [...] we applaud those who bravely and romantically are applying hill-climbing methods to many new domains [...] new phenomena that are well worth understanding.

Perceptrons, act II.



Perceptrons
Minsky & Papert
1968



We must first examine [...] two branches of theory which could be called “theory of learning” and “theory of representation.”

[...] we really need to know a great deal about the nature of the problem and why we are applying hill-climbing methods to many new domains [...] new phenomena that are well worth understanding.

Getting there...

Progress 1 : Statistical learning theory

Dokl. Akad. Nauk SSSR
Tom 181 (1968), No. 4

UNIFORM CONVERGENCE OF FREQUENCIES OF OCCURRENCE OF EVENTS TO THEIR PROBABILITIES

UDC 519.21

V. N. VAPNIK AND A. Ja. ČERVONENKIS

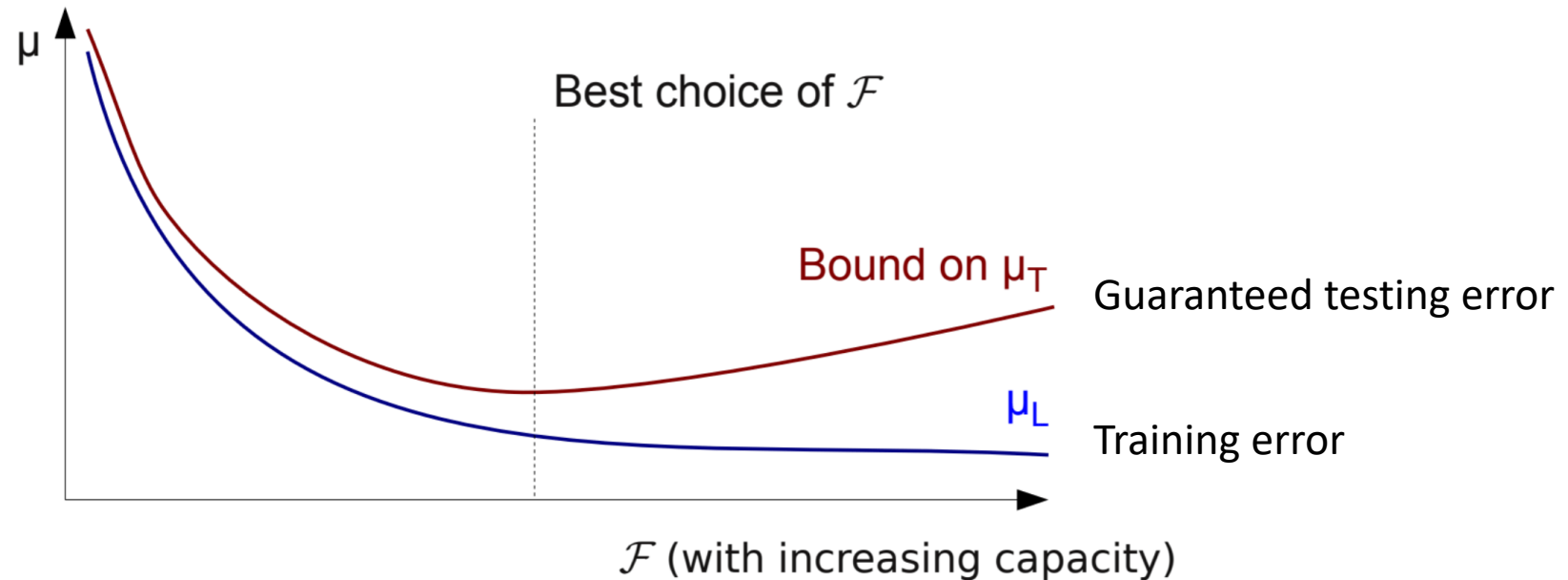
1. Introduction. According to the classical theorem of Bernoulli, the frequency of occurrence of an event A converges (in probability, in a sequence of independent trials to the probability of this event). In many applications, however, it is necessary to estimate the probabilities of the events of an entire class S from one and the same sample. (In particular, this is necessary in the construction of learning algorithms.) Here it is important to know if the frequencies converge to the probabilities uniformly on the entire class of events S . More precisely, it is important to know if the probability

1968-98

"necessary in the construction of learning algorithms"

Progress 1: Statistical learning theory

- Under the assumption that training and testing data follow the same distribution.
- *Structural risk minimization* gives a precise meaning to Occam's razor.



Progress 2: understanding deep nets

2018-now

- There are scaling limits for which deep nets can be described with convex mathematics.
- These limits are not very interesting in practice, but can be used as anchors to analyze interesting networks and make verifiable predictions.

That is, simultaneously scaling layer sizes, initial weights, step sizes, according to certain schedules.

- Neural Tangent Kernel limit (2018; several groups)
- Mean Field limit (2018-2019; several groups)

Setting up networks according to these scaling laws does not give high performance networks. But one describe interesting networks as perturbations of these simple cases (and derive better scaling laws.)

Progress 2: understanding deep nets

The Principles of Deep Learning Theory

An Effective Theory Approach to Understanding Neural Networks

Daniel A. Roberts and Sho Yaida
based on research in collaboration with

Boris Hanin

463 pages

Tensor Programs 0, I, II, II.b, III, IV, V Greg Yang (and coauthors)

I am currently developing a framework called *Tensor Programs* for understanding large (wide) neural networks

428 pages and growing

2

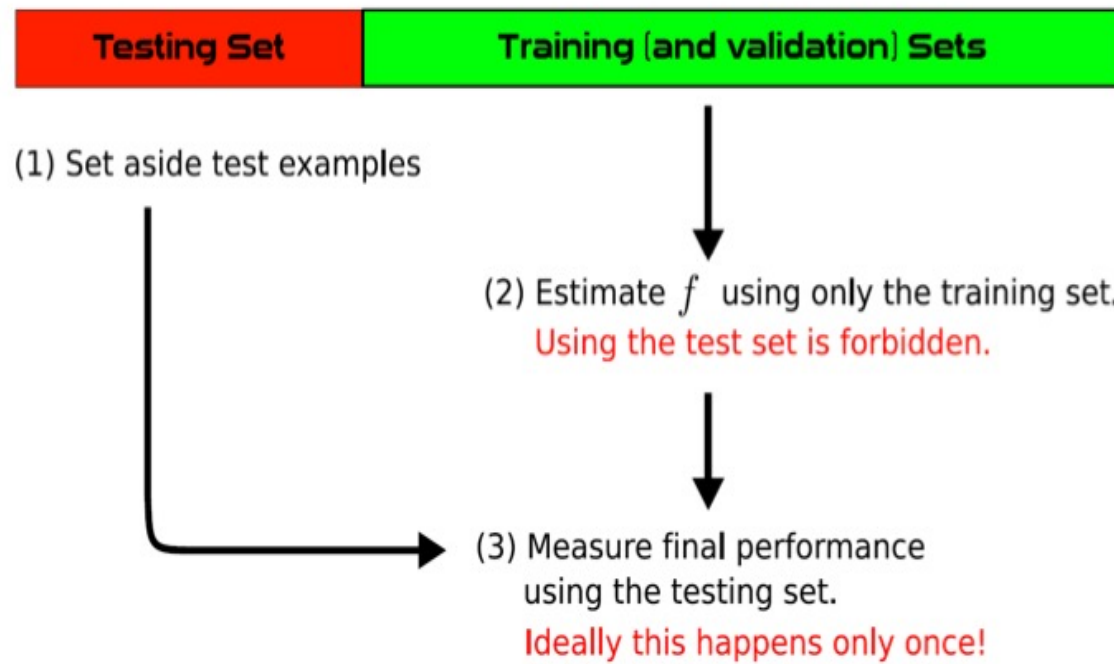
Beyond statistics

DISTRIBUTION SHIFTS, CAUSALITY, AND FEATURES

Statistical machine learning

Data, present and future, is sampled from a same distribution.

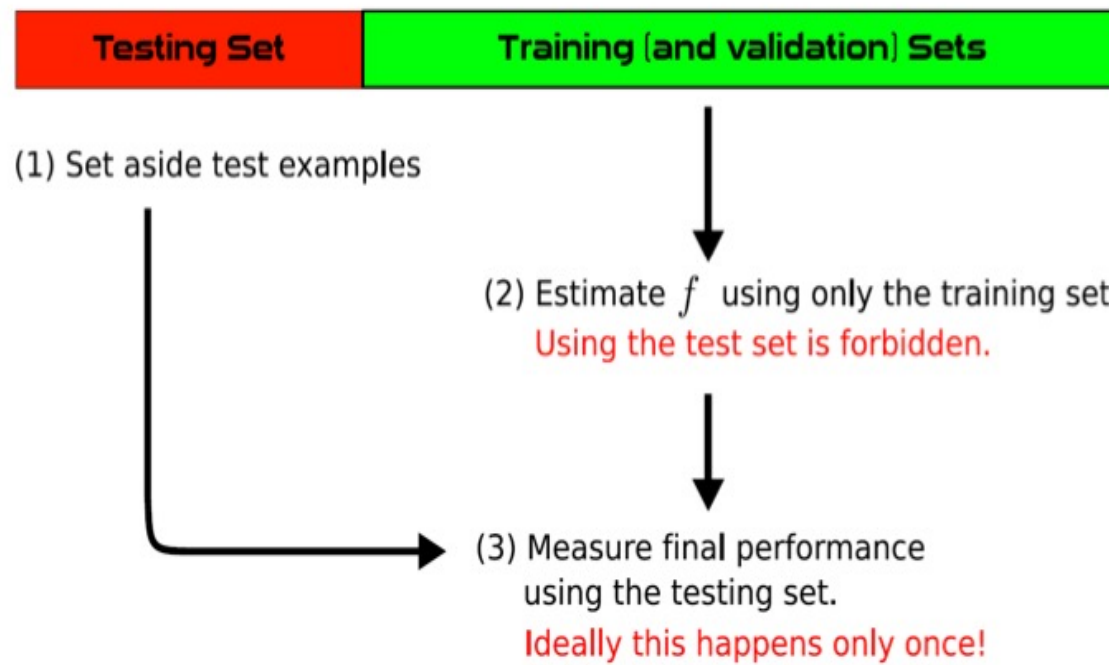
→ Rigorous evaluation is possible using held-out data.



Statistical machine learning

Data, present and future, is sampled from a same unknown distribution.

→ Rigorous evaluation is possible using held-out data.

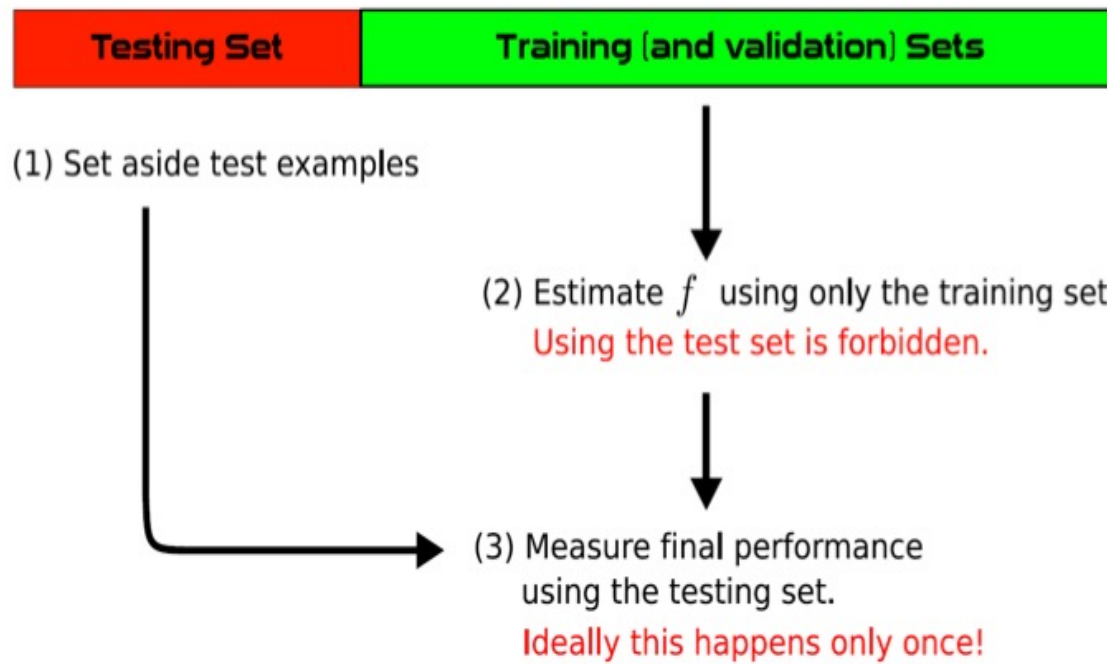


This single empirical paradigm has driven progress in machine learning for two or three decades.

Statistical machine learning

Data, present and future, is sampled from a same unknown distribution.

→ Rigorous evaluation is possible using held-out data.



This single empirical paradigm has been program... the... or... ades.

Not anymore

Not anymore

1- Statistical guarantees are brittle

“DeepVisotron™, SM, ® detects 1000 object categories with less than 1% errors.”



What is the nature of this statement?

- It does not mean that one rolls a dice for each picture.
- It is tied to a specific image distribution.
The error guarantee is lost if the image distribution changes.

Not anymore

2- Data collection bias

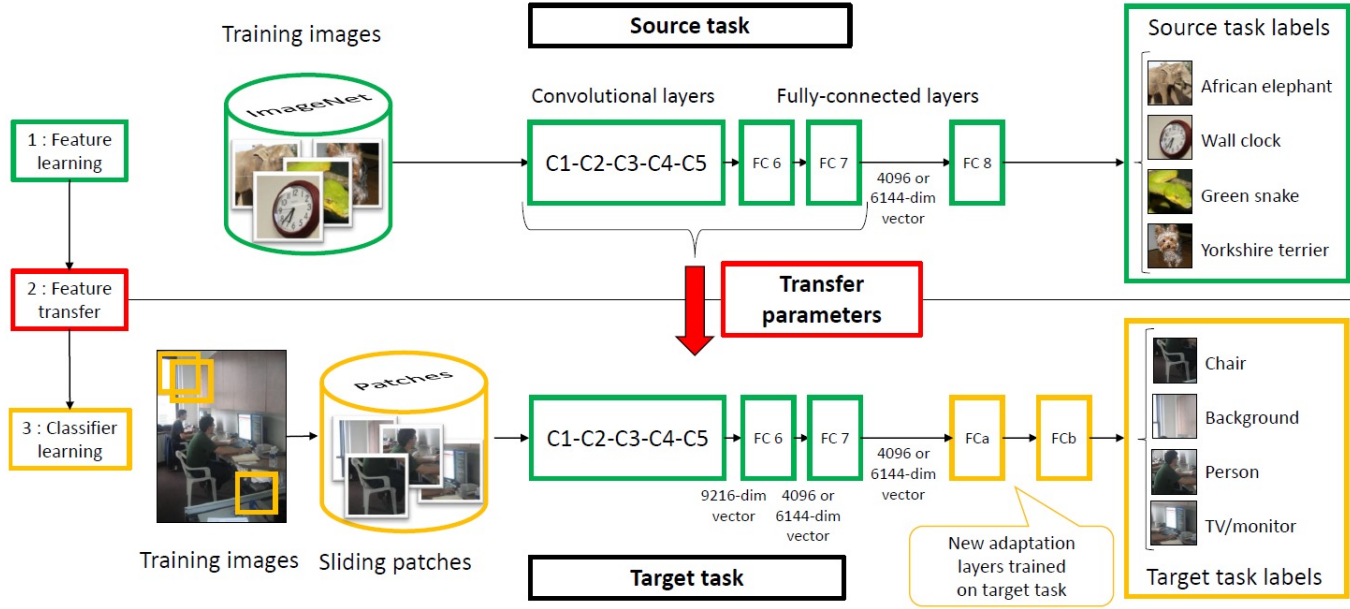
Training and testing on different datasets

- E.g. : Torralba and Efros, *Unbiased look at dataset bias*, CVPR 2011.

<i>task</i>	Test on:		SUN09	LabelMe	PASCAL	ImageNet	Caltech101	MSRC
	Train on:							
<i>“car” classification</i>	SUN09		28.2	29.5	16.3	14.6	16.9	21.9
	LabelMe		14.7	34.0	16.7	22.9	43.6	24.5
	PASCAL		10.1	25.5	35.2	43.9	44.2	39.4
	ImageNet		11.4	29.6	36.0	57.4	52.3	42.7
	Caltech101		7.5	31.1	19.5	33.1	96.9	42.1
	MSRC		9.3	27.0	24.9	32.6	40.3	68.4
	Mean others		10.6	28.5	22.7	29.4	39.4	34.1

Not anymore

3- Transfer learning



- **Train on ImageNet**
 - millions of examples
 - thousand of classes
- **Copy trained features extraction layers**
 - millions of parameters
- **Train on new task**
 - smaller data
 - closer to our interests

3- Transfer learning

- **Train on ImageNet**
 - millions of examples
 - thousand of classes
- **Copy trained features extraction layers**
 - millions of parameters
- **Train on new task**
 - smaller data
 - closer to our interests

So pervasive that large donor models are now called:
Foundational models

- Vision : Imagenet models
- Language : BERT
- Machine translation : NLLB

Not anymore

4- Transfer across related tasks

Target task: “Recognizing persons in images.”

- Many good reasons to avoid collecting large databases of labeled persons.

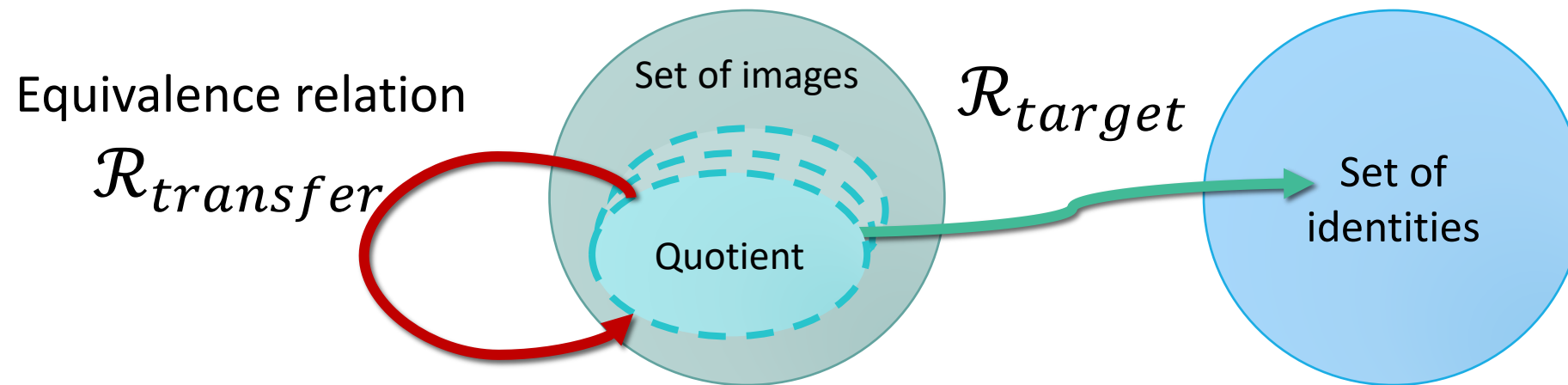
Auxiliary task: “Do these images represent the same person?”

- Two faces in the same picture usually are different persons.
- Two faces in successive video frames are often the same person.



4- Transfer across related tasks

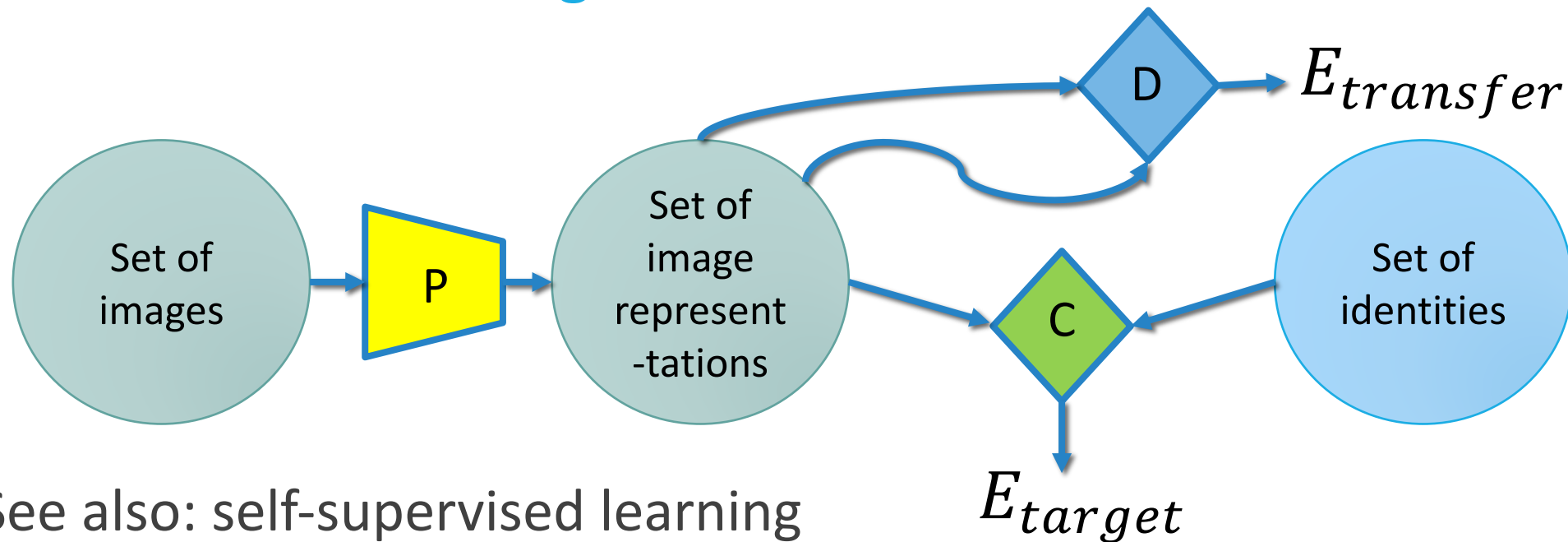
Tasks have an algebraic structure



Not a formal task specification but a formal connection between tasks!

4- Transfer across related tasks

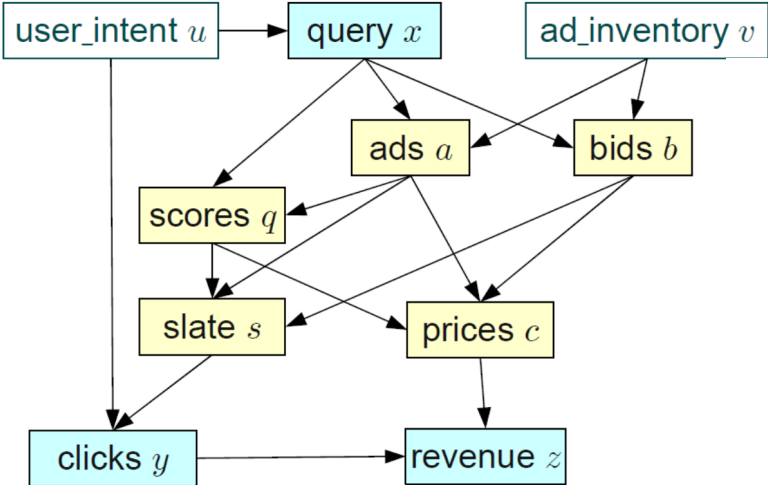
Models mimic the algebraic structure of the tasks



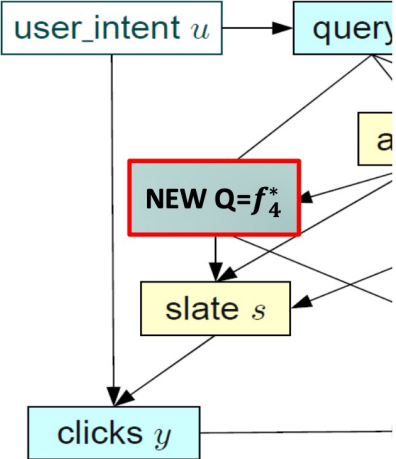
See also: self-supervised learning

Not anymore

5 Causal inference (Pearl style)



$$\begin{aligned} x &= f_1(u, \epsilon_1) \\ a &= f_2(x, v, \epsilon_2) \\ b &= f_3(x, v, \epsilon_3) \\ q &= \cancel{f_4(x, a, \epsilon_4)} \quad f_4^*(x, a, \epsilon_4) \\ s &= f_5(a, q, b, \epsilon_5) \\ c &= f_6(a, q, b, \epsilon_6) \\ y &= f_7(s, u, \epsilon_7) \\ z &= f_8(y, c, \epsilon_8) \end{aligned}$$

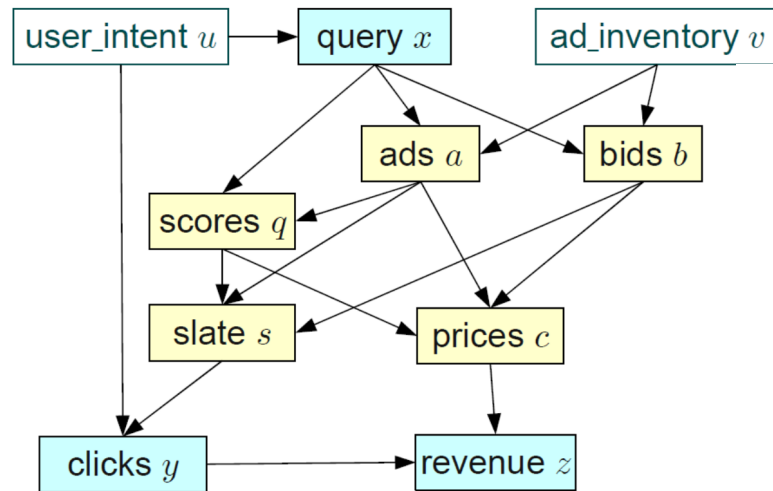


This mechanism leads to a first data distribution

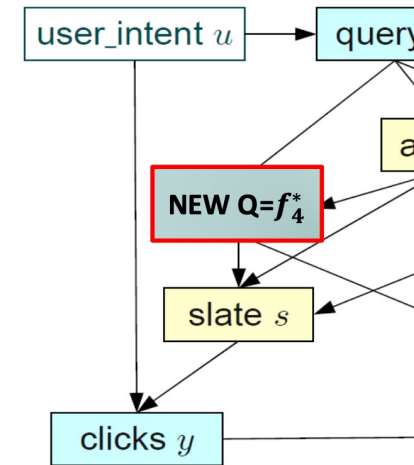
An "intervention" alters the mechanism ...

... leading to a new data distribution.

5 Causal inference (Pearl style)

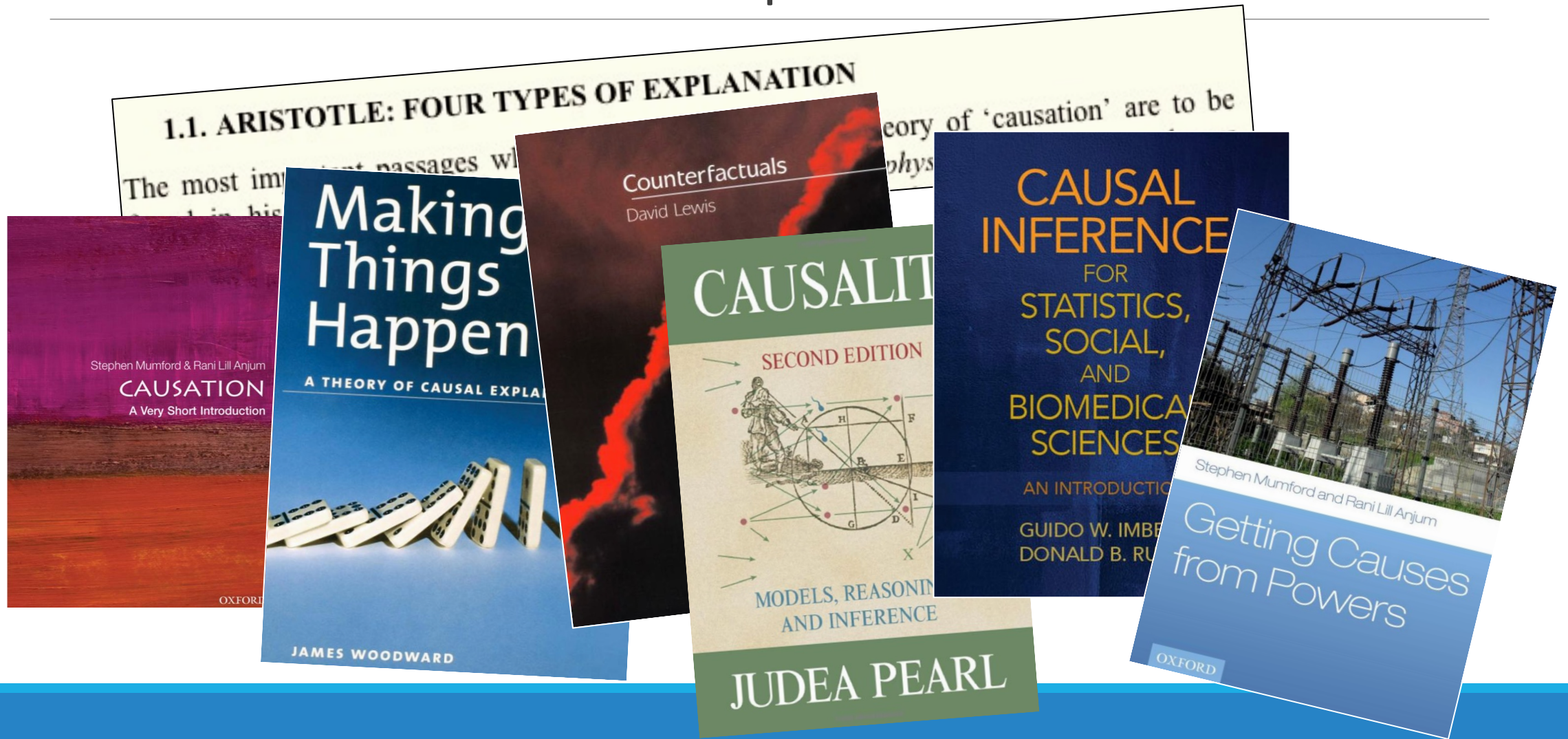


$$\begin{aligned}
 x &= f_1(u, \varepsilon_1) && \text{Unknown} \\
 a &= f_2(x, v, \varepsilon_2) \\
 b &= f_3(x, v, \varepsilon_3) \\
 q &= \cancel{f_4(x, a, \varepsilon_4)} \quad f_4^*(x, a, \varepsilon_4) \\
 s &= f_5(a, q, b, \varepsilon_5) \\
 c &= f_6(a, q, b, \varepsilon_6) \\
 y &= f_7(s, u, \varepsilon_7) && \text{Unknown} \\
 z &= f_8(y, c, \varepsilon_8)
 \end{aligned}$$



- Although parts of the mechanism can be unknown, there is an algebraic connection between data distributions before and after intervention!
- Interventions on a graph in fact form a groupoid.

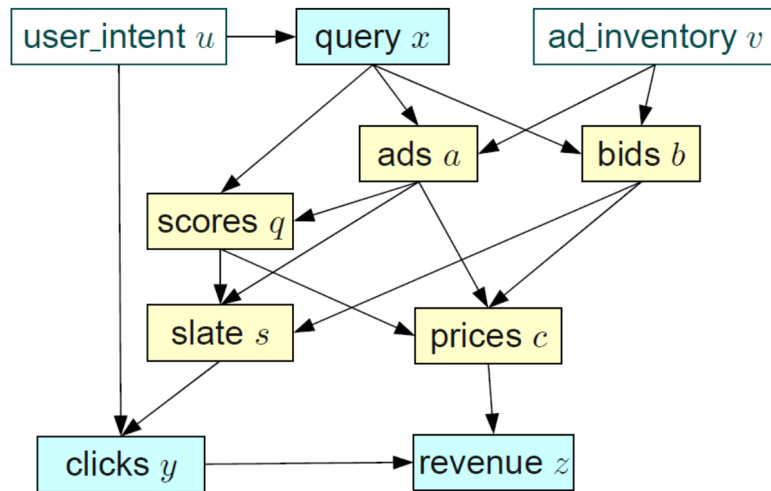
Causation is a rich topic



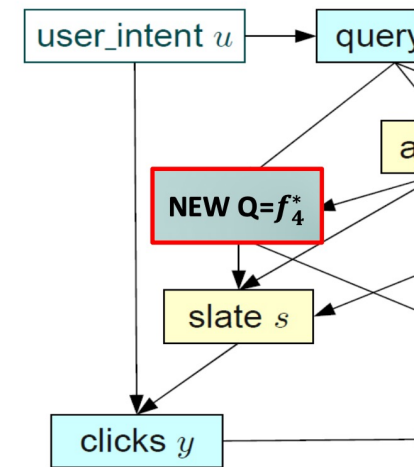
Many viewpoints on causation

1. **Manipulative causation** : Causation describes the outcome of interventions.
2. **Causal invariance** : Which properties are conserved when the system changes.
3. **Causal reasoning** : Making sense of causation without interventions.
4. **Dispositional causation, affordances** : Where do causal relations come from?
5. **Causal intuition** : Correlations do not imply causation, but data contains a lot of hints.
6. and many more..

Causal inference with DAGs (Pearl style)



$$\begin{aligned}
 x &= f_1(u, \varepsilon_1) \\
 a &= f_2(x, v, \varepsilon_2) \\
 b &= f_3(x, v, \varepsilon_3) \\
 q &= \cancel{f_4(x, a, \varepsilon_4)} \quad f_4^*(x, a, \varepsilon_4) \\
 s &= f_5(a, q, b, \varepsilon_5) \\
 c &= f_6(a, q, b, \varepsilon_6) \\
 y &= f_7(s, u, \varepsilon_7) \\
 z &= f_8(y, c, \varepsilon_8)
 \end{aligned}$$



Given observed conditional distribution for this model, ...

... can we predict conditional distributions for that model ?

Causal inference with DAGs

Given observed conditional distribution on this model, ...
 $P_1(A, B|C, D)$

... can we predict conditional distributions on that model?
 $P_2(U, V|R, S)$

Bayes' rule

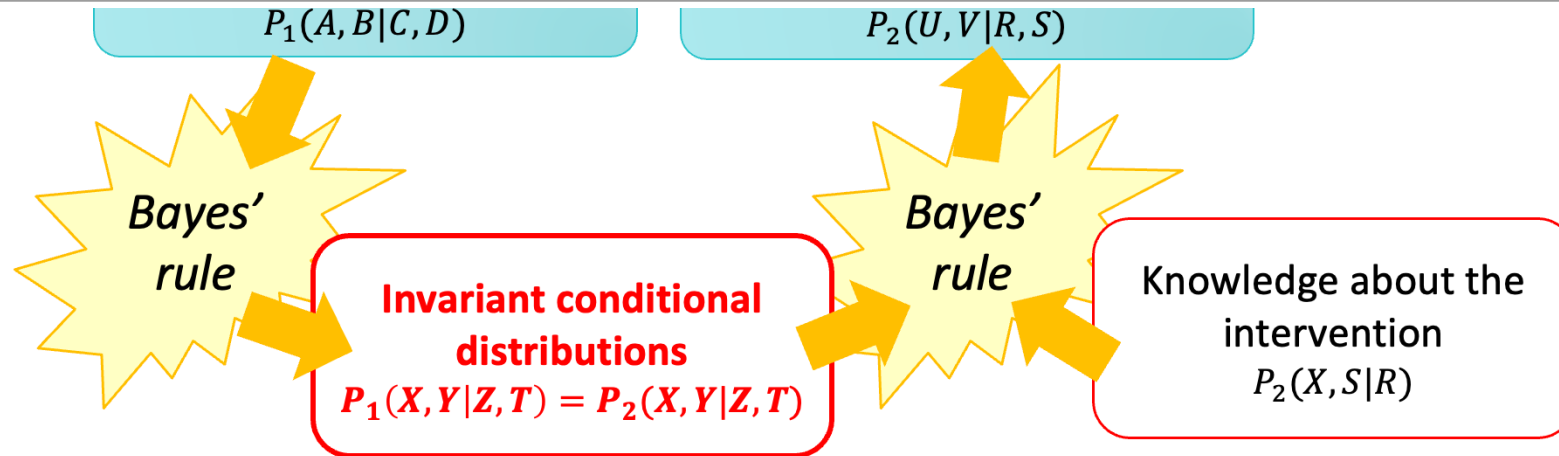
Invariant conditional distributions

$$P_1(X, Y|Z, T) = P_2(X, Y|Z, T)$$

Bayes' rule

Knowledge about the intervention
 $P_2(X, S|R)$

Invariance is the key



- This process is formalized by Pearl's *do-calculus*.
The *assumed graph structure* informs us about the invariant conditionals
- Although Rubin's *methods of potential outcomes* is less formal, his "ignorability" assumptions inform us about the invariant conditionals

How do we solve causal inference problems in physics

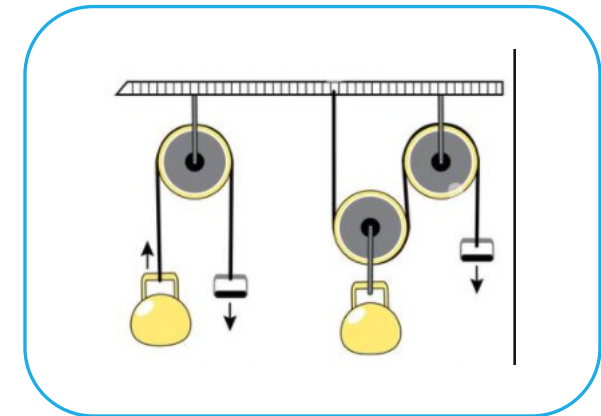
Method 1: ODEs

- Numerical integration scheme defines a **DAG**
- Must predict all the trajectory to determine the final state.

Method 2: Use invariants directly

- Write equations that describe the intervention
- Write equations that describe invariants (local or universal)
- Solve!

$$m\ddot{x} = f(x, \dot{x})$$



How do we solve causal inference problems in physics

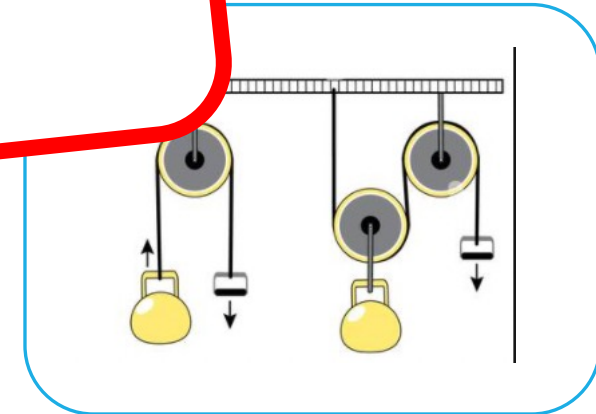
Method 1: ODE \rightarrow DAG

- Nu
- M
- Me
- W
- Write equations that describe invariants (local or universal)
- Solve!



Invariance \Leftrightarrow Causation
(Cartwright 2003) (Woodward 2005)

$$= f(x, \dot{x})$$



Discovering invariances from multiple environments

Following Peters et al. (2016), we consider that data from each environment e comes with a different distribution P_e .

$$P_e(X_1, X_2, \dots, X_n, Y) \text{ for } e = 1, 2, 3 \dots$$

- Training sets $D_e = \{(x_1 \dots x_n, y) \sim P_e\}$ are provided for some e .
- We want a predictor of Y that works for many e .



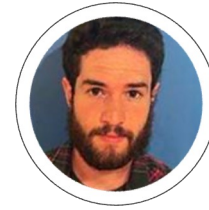
Find a subset $S = \{X_3, X_7 \dots\}$ of the variables X_i such that $P_e(Y|S)$ does not depend on e .

Intuition



“If we find a representation in which all falling objects obey the same laws, then we possibly understand something useful.”

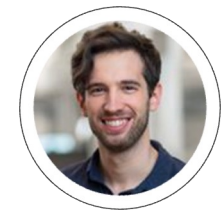
Invariant risk minimization



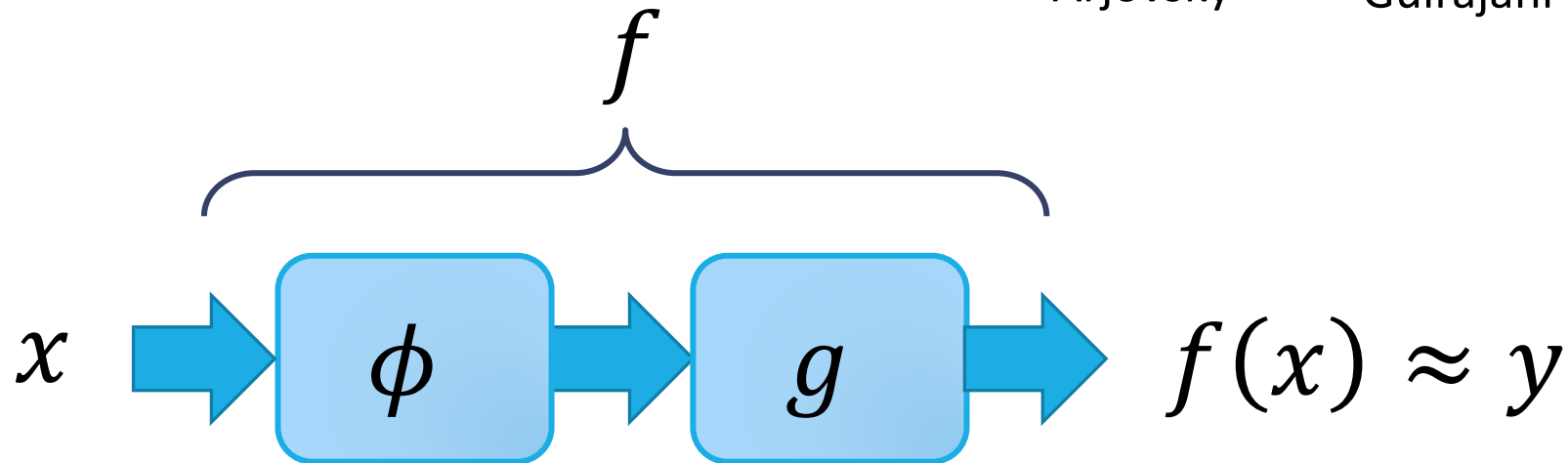
Martin
Arjovsky



Ishaan
Gulrajani



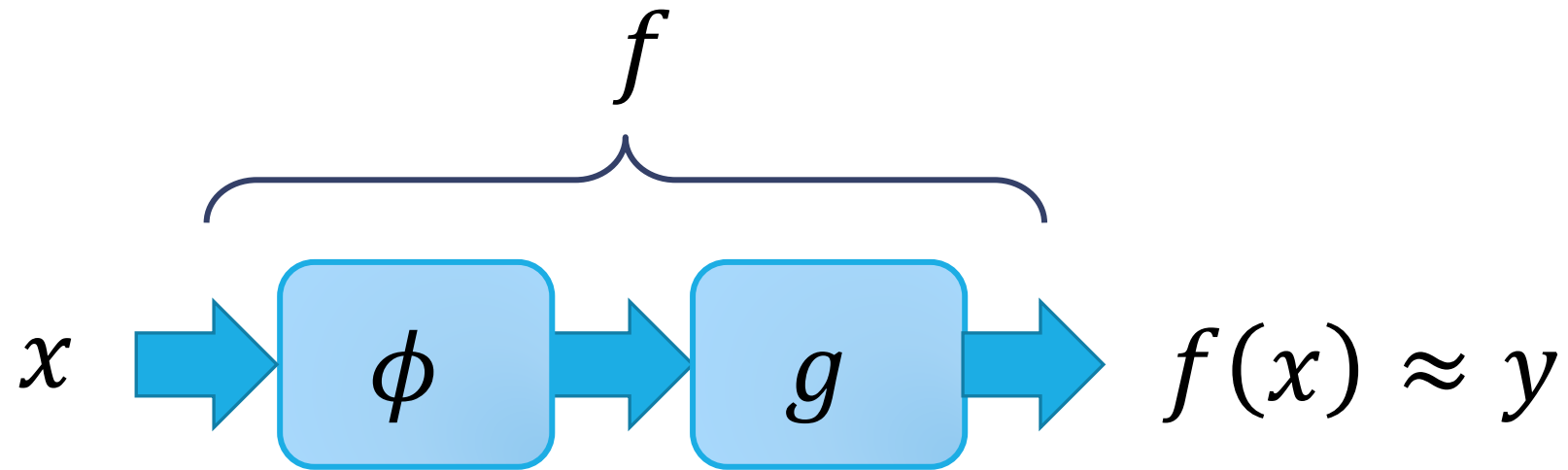
David
Lopez-Paz



Find a **representation** $\phi(x)$

Such that the regression from $\phi(x)$ to y
is invariant across environments

Invariant representation



Find a **representation** $\phi(x)$

Such that the regression function
is invariant to **transformations**

And also predicts well... **fits**

Invariant risk minimization (IRM)

Minimize a regularized cost

Average error across all environments.

Penalty that favors invariant solutions

$$\sum_e \frac{1}{n_e} \overbrace{\|Y_e - f_w(X_e)\|^2}^{C_e(w)} + \kappa \sum_e \Omega_e(w)$$

Colored MNIST



Digits with misleading colors

	Y=0	Y=1
{0,1,2,3,4}	0.75	0.25
{5,6,7,8,9}	0.25	0.75

The optimal classification rate on the basis of the shape only is 75%.

Random guess is 50%.

	Red	Green
Y=0	$1 - e$	e
Y=1	e	$1 - e$

During the training $e \in \{0.1, 0.2\}$. The color is a better indicator than the shape, but not a stable one.

Then we test with $e = 0.9$.

Colored MNIST

Training with $e \in \{0.1, 0.2\}$	Testing with $e \in \{0.1, 0.2\}$	Testing with $e = 0.9$
Minimize empirical risk after mixing data from both environments	84.3%	10.1%
Minimize empirical risk with invariant regularization	70.8%	66.9%

- Network is a MLP with 256 hidden units on 14x14 images.
- Invariant regularization tuned high : regularization term is nearly zero.

Subsequent work

- Alternate algorithms and constraints
 - IGA (Koyama et al., 2020); Invariant Games (Ahuja et al. 2020); vREX (Krueger et al. 2020); FISH (Shi et al., 2021); FISHR (Rame et al, 2021); SD; RSC, LfF, CLOvE, ...
- Theoretical issues
 - Linear (IRMv1) is often not good enough (Kamath & al, 2021)
- Benchmarks and applications
 - Domainbed (Gulrajani et al., 2020)
 - Wildcam dataset (<https://ff13.fastforwardlabs.com>)
 - Toxic language classification (Adragna et al., 2020)

And disappointments



Trying invariant learning on real OOD problems:

- Invariant learning **sometimes** yields a **small** improvement.
- But these results do **not measure with our hopes...**
- And always come after a **finicky optimization**...

Various experiments shows that such cost functions are just too hard to optimize reliably!

Rich features



Jianyu Zhang



David Lopez-Paz

Difficult optimization can often be helped by a good initialization



- Initialize the network with a rich features set.
- Let the learning algorithm pick the one it likes

Rich Feature Construction (RFC)

- Train a network

$$\min_{\Phi, g} \mathbb{E}_p[\ell(y, g(\Phi(x)))]$$

- Freeze features and find pessimal data reweighting

$$\max_{q \in Q} \min_g \mathbb{E}_q[\ell(y, g(\Phi(x)))]$$

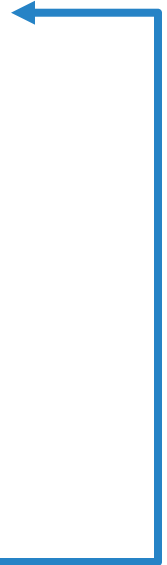
- Training again forces discovering new features Φ'

$$\min_{\Phi', g} \mathbb{E}_q[\ell(y, g(\Phi'(x)))]$$

- Gather old and new features

$$\Phi \cup \Phi' \rightarrow \Phi_{new}$$

- Repeat



Rich Feature Construction (RFC)



- Train a network

$$\min_{\Phi, g} \mathbb{E}_q[\ell(y, g(\Phi(x)))]$$

Weight on subsets of examples

- Freeze features and find possible data reweighting

$$\max_{q \in \mathcal{Q}} \min_g \mathbb{E}_q[\ell(y, g(\Phi(x)))]$$

- Training again forces discovering new features Φ'

$$\min_{\Phi', g} \mathbb{E}_q[\ell(y, g(\Phi'(x)))]$$

Distillation

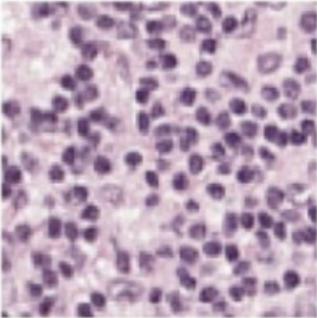
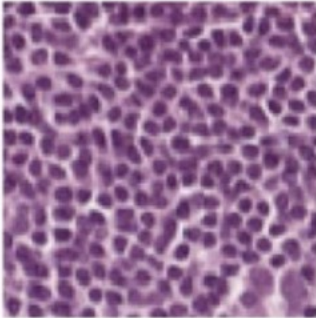
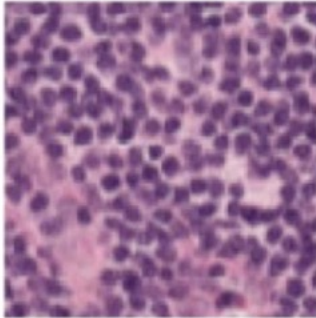
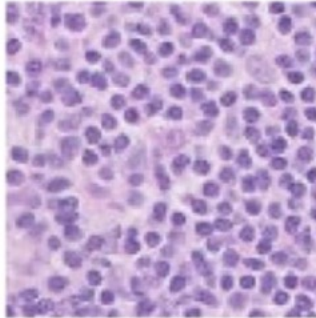
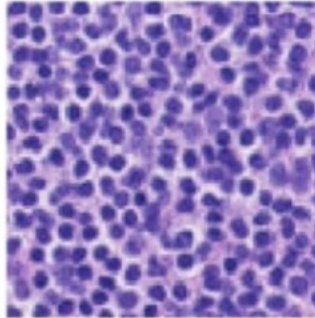
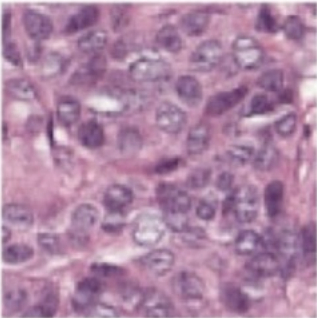
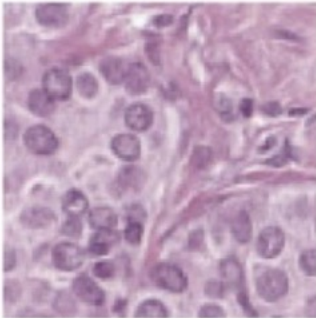
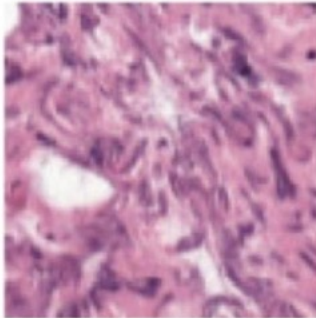
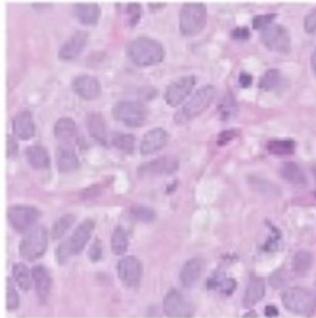
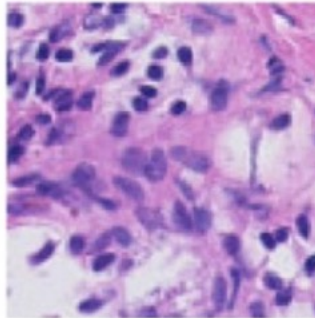
- Gather old and new features

$$\Phi \cup \Phi' \rightarrow \Phi_{new}$$

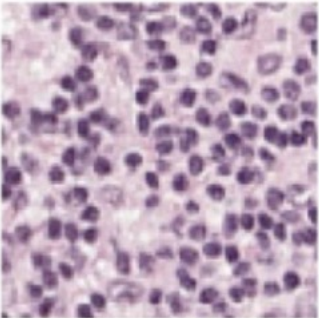

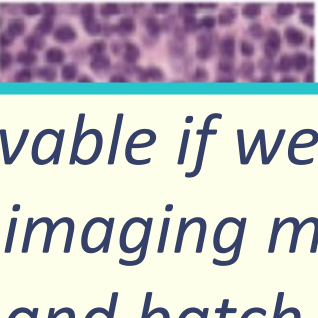
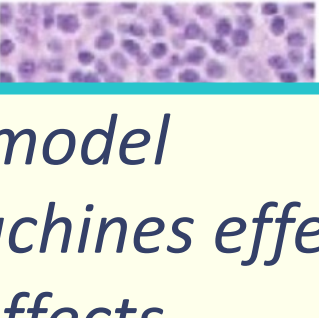

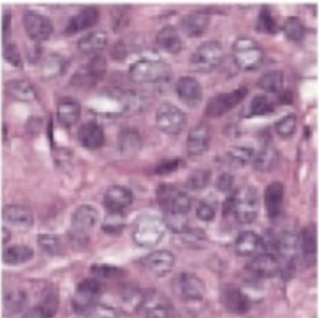
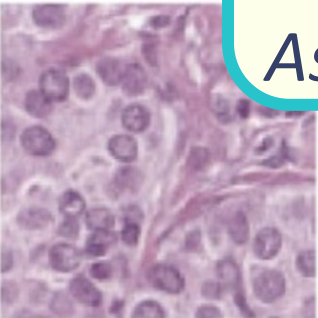
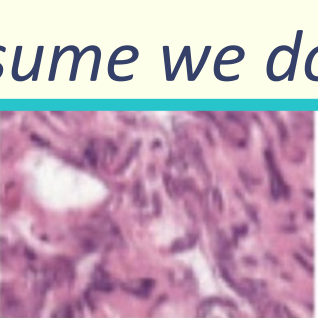
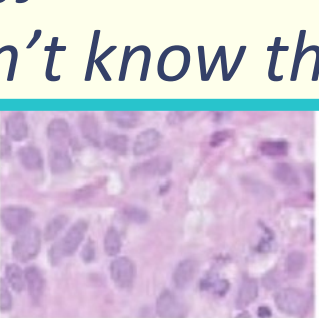
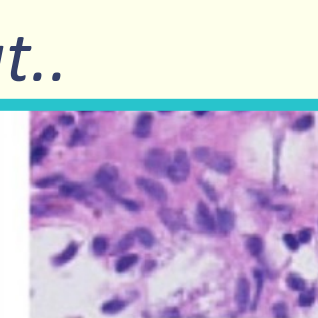
- Repeat

And a duality trick to save distillation time

Wilds / Camelyon17

	Train			Val (OOD)	Test (OOD)
	d = Hospital 1	d = Hospital 2	d = Hospital 3	d = Hospital 4	d = Hospital 5
y = Normal					
y = Tumor					

Wilds / Camelyon17







	Train			Val (OOD)	Test (OOD)
	d = Hospital 1	d = Hospital 2	d = Hospital 3	d = Hospital 4	d = Hospital 5
y = Normal					
y = Tumor					

Solvable if we model

- *imaging machines effects*
- *and batch effects*

Assume we don't know that..

Wilds / Camelyon17

	Train	Val (OOD)	Test (OOD)
d = Ho			
y = Normal			
y = Tumor			
			
			

*“While the camelyon17 dataset is small and fast to train on, we advise against using it as the only dataset to prototype methods on, as the test performance of models trained on this dataset tend to **exhibit a large degree of variability over random seeds.**”*

Leaderboard best: 74.7% **± 7.1%**

Network Initialization	Methods	Test Acc	
		IID Tune	OOD Tune
× ERM	ERM	66.6±9.8	70.2±8.7
	IRMv1	68.6±6.8	68.5±6.2
	vREx	69.1±8.1	69.1±13.2
	CLOvE	71.7±10.2	69.0±12.1
2-RFC	ERM	72.8±3.2	74.7±4.3
	IRMv1	71.6±4.2	75.3±4.8
	vREx	73.4±3.3	76.4±5.3
	CLOvE	74.0±4.6	76.6±5.3
2-RFC	ERM(cf)	78.2±2.6	78.6±2.6
	IRMv1(cf)	78.0±2.1	79.1±2.1
	vREx(cf)	77.9±2.7	79.5±2.7
	CLOvE(cf)	77.8±2.2	78.6±2.6

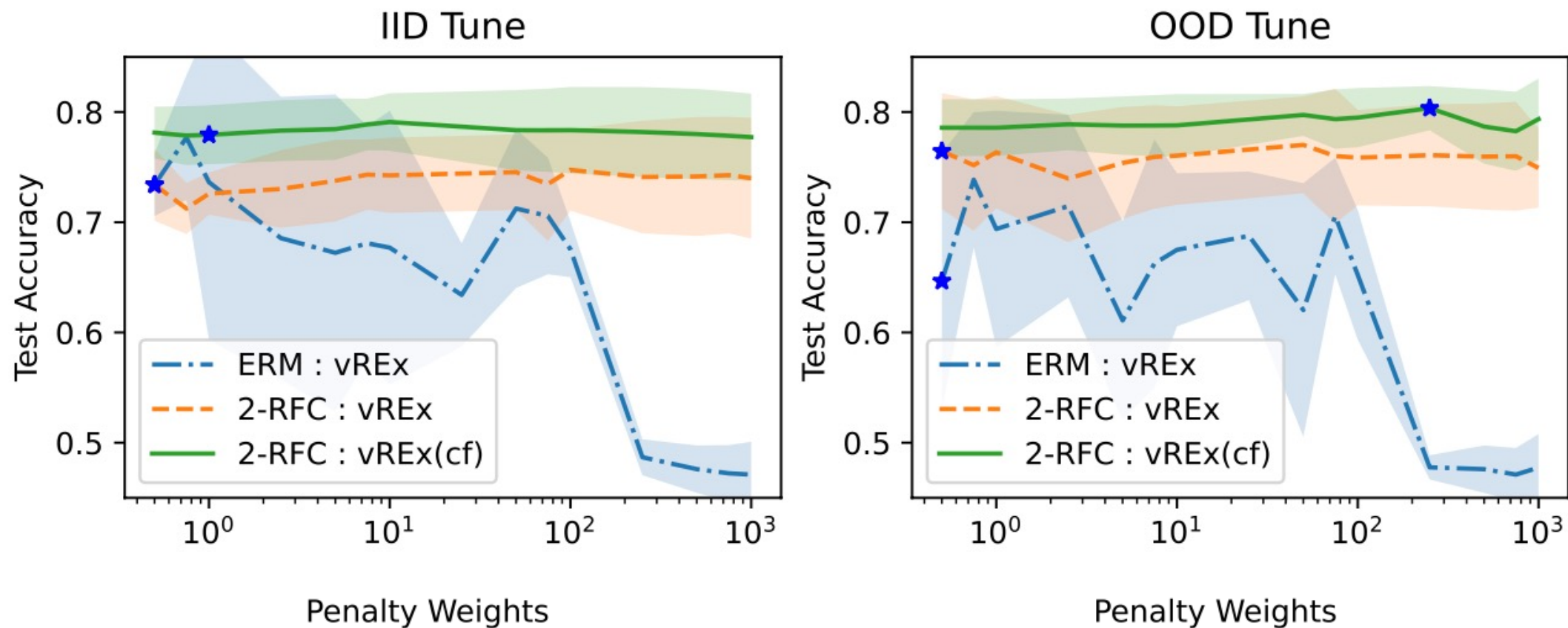
Leaderboard best:
74.7 ± 7.1 %

2RFC + ERM !

Frozen features !

Everything works!

Wilds/Camelyon17



Is this all about the representations?

Everything works robustly after RFC!

- We arrived there by trying to improve machine learning methods that attempt to discover causation through invariances.
- **The nature of the internal representations might matter more than invariant training!**
- Remark: RFC constructs representations that are “richer” than those obtained by training with optimization.



Worth checking...

Supervised transfer (linear probing)

method	architecture	params	ID	Linear Probing (OOD)		
			IMAGENET	INAT18	CIFAR100	CIFAR10
ERM	RESNET50	23.5M	75.58	37.91	90.57	73.23
ERM	RESNET50W2	93.9M	77.58	37.34	90.86	72.65
ERM	RESNET50W4	375M	78.46	38.71	92.13	74.81
ERM	2×RESNET50	47M	75.03	39.34	90.94	74.36
ERM	4×RESNET50	94M	75.62	41.89	90.61	74.06
CAT2	2×RESNET50	47M	77.57	43.26	91.86	76.10
CAT4	4×RESNET50	94M	78.15	46.55	93.09	78.19
CAT5	5×RESNET50	118M	78.27	47.78	93.21	78.53
CAT10	10×RESNET50	235M	78.36	49.65	93.75	79.61

Supervised transfer (linear combination)

method	architecture					
ERM	RESNET50					78.10
ERM	RESNET50					78.10
ERM	RESNET50					78.10
ERM	2×RESNET50					78.10
ERM	4×RESNET50					78.10
CAT2	2×RESNET50					78.10
CAT4	4×RESNET50	94M	78.15	46.55	93.09	78.19
CAT5	5×RESNET50	118M	78.27	47.78	93.21	78.53
CAT10	10×RESNET50	235M	78.36	49.65	93.75	79.61
DISTILL5	RESNET50	23.5M	76.39	40.75	92.54	76.50

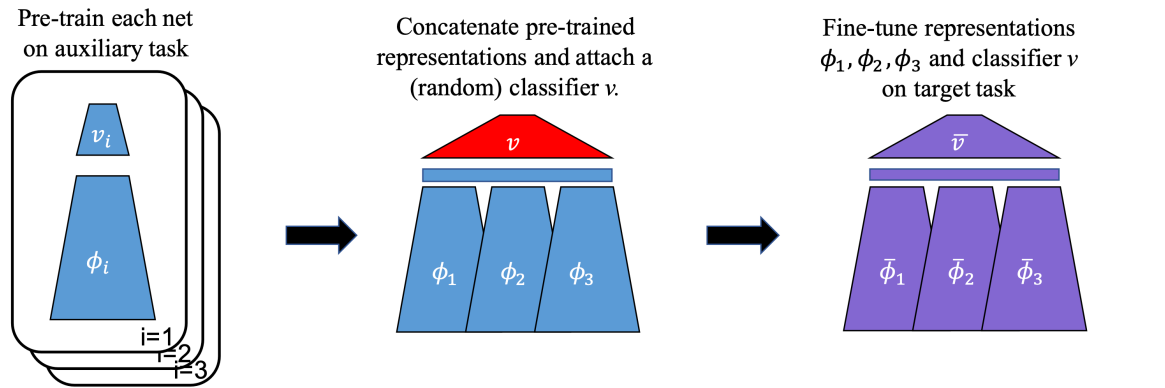
Concatenation of features obtained by training the **same network** using the **same data** with the **same algorithm** and the same **hyper-parameters**.

Only the random seed changes.

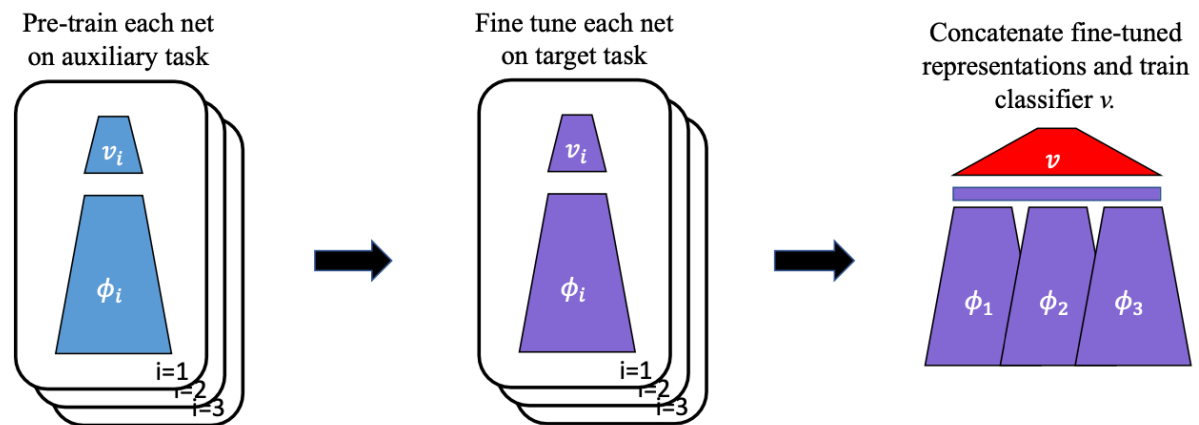
This is not an ensemble of models with engineered diversity.

Fine-tuning representations

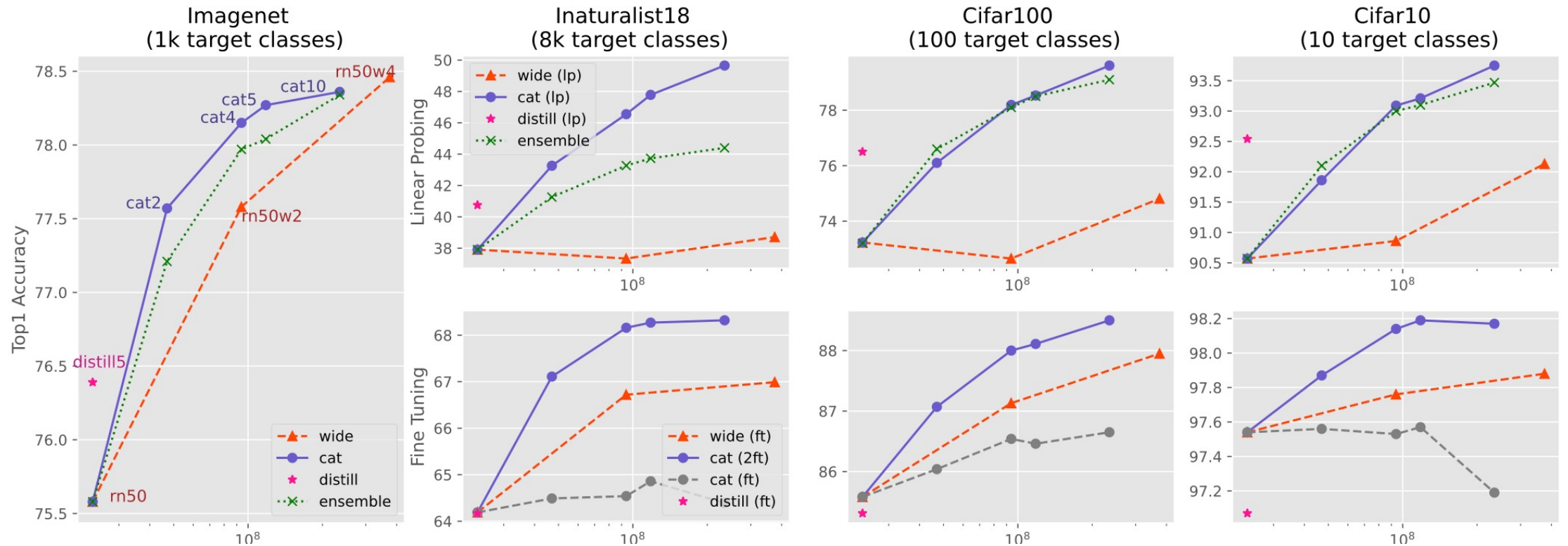
Traditional fine-tuning



Two stages fine-tuning



Supervised transfer



Note: Two stage fine tuning (works) vs single stage fine tuning (does not work well).

Vision transformers

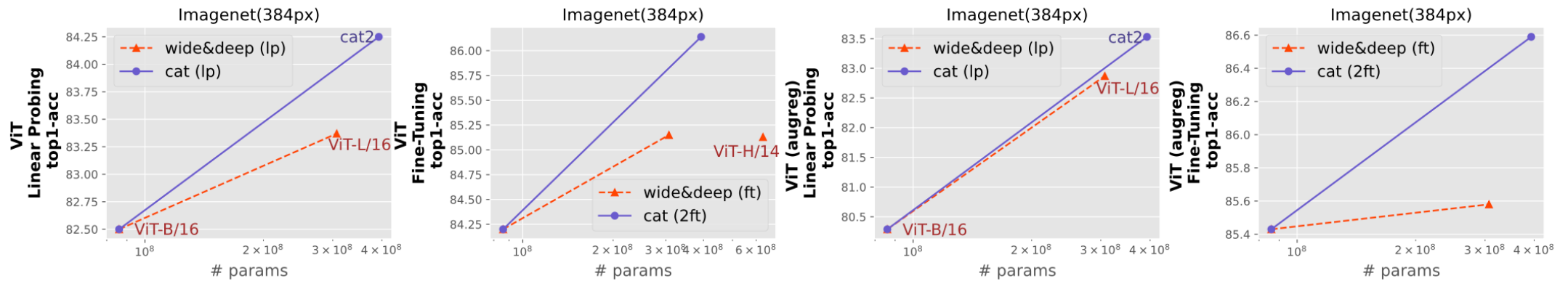
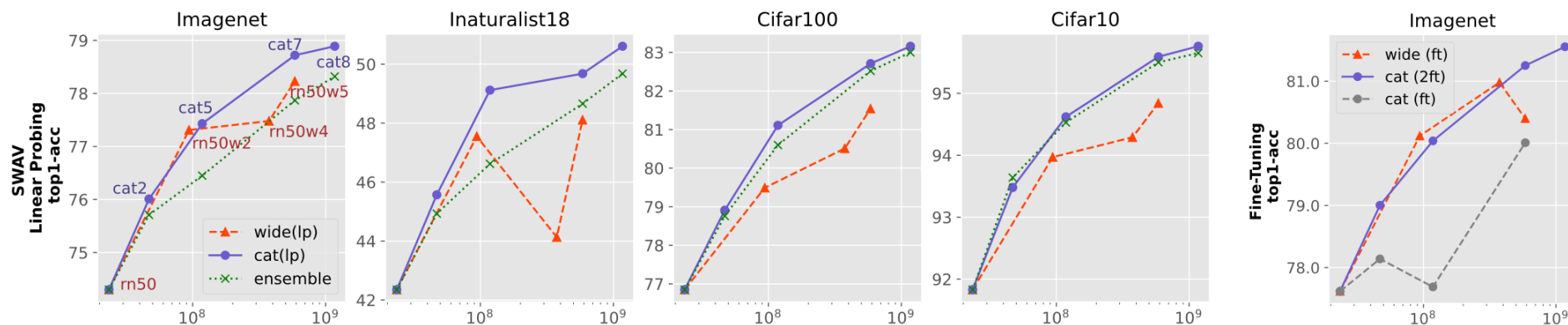


Figure 5: Supervised transfer learning from IMAGENET21K to IMAGENET on vision transformers.

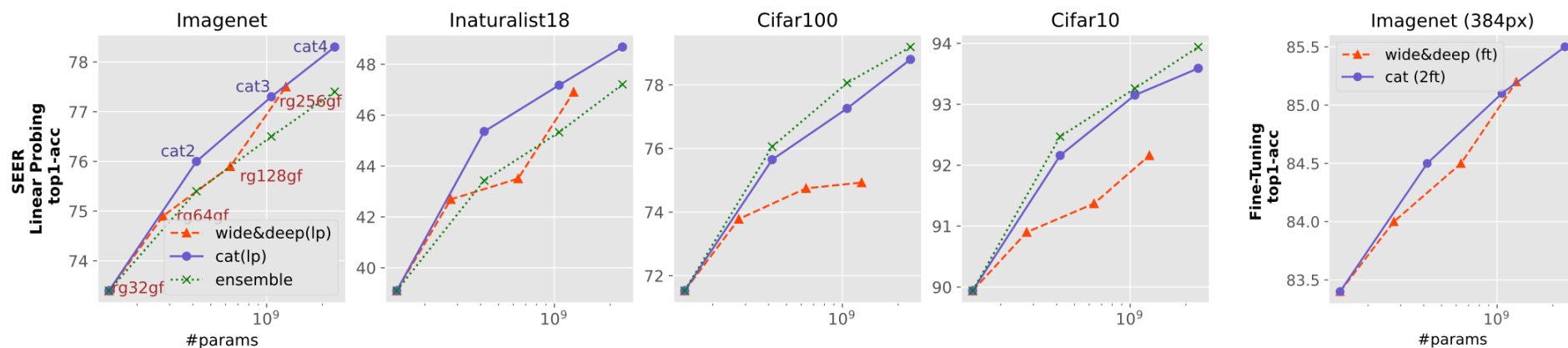
Using snapshots from (Dosovitsky et al, 2020) (Steiner et al., 2021)

SSL transfer

(Caron et al. 2020)
SSL training on
Imagenet1.2M.
Transfer to
classification tasks



(Goyal et al. 2022)
SSL training on
Instagram1B.
Transfer to
classification tasks



Learning features with optimization

- Deep learning optimization seems able to construct diverse features.
- Deep learning by optimization then prunes features that appear redundant on the basis of the training data distribution.
GD/SGD algorithms in deep nets: implicit bias towards sparse solutions.
- Sparsity is good for in-distribution generalization (Occam's razor)
- Features eliminated because they were redundant for the training distribution might in fact be very informative for a new distribution.

Optimization vs memorization

Feature optimization

- Once a set of features appears sufficient to deliver a good training cost, there is no need to find or collect new ones.
- Implicit bias towards sparsity.

Feature memorization

- Memorize every feature that appears useful at any point, even if removing it later would not penalize the training cost.
- Deal with Occam's razor later.

Prematurely pruning the representations might not be the best way to prepare for changing tasks and distributions.

3

The infinite library

MEMORIZATION AT INTERNET SCALE

Ongoing technological race

Large language models (LLMs),

Approaching 10^{12} parameters
The human brain has $\sim 10^{14}$ synapses

trained on inhumanly large datasets,

A couple terabytes (10^{12} bytes)
and increasingly multimodal

as a single optimization run.

This can be very costly.



Ongoing technological race

Large language models (LLMs),

Approaching 10^{12} parameters
The human brain has $\sim 10^{14}$ synapses

trained on inhumanly large datasets,

A couple terabytes (10^{12} bytes)
and increasingly multimodal

as a single optimization run.

This can be very costly.



Can we work around the out-of-distribution problems by training on everything?

Maybe not a great idea!

Confusing competency claims

Are LLMs merely language models

Impressive language competencies:

- Ability to contextualize memorized sentences.
- Ability to compose memorized sentences to make meaningful new ones.

We have much to understand about the underlying mechanisms.

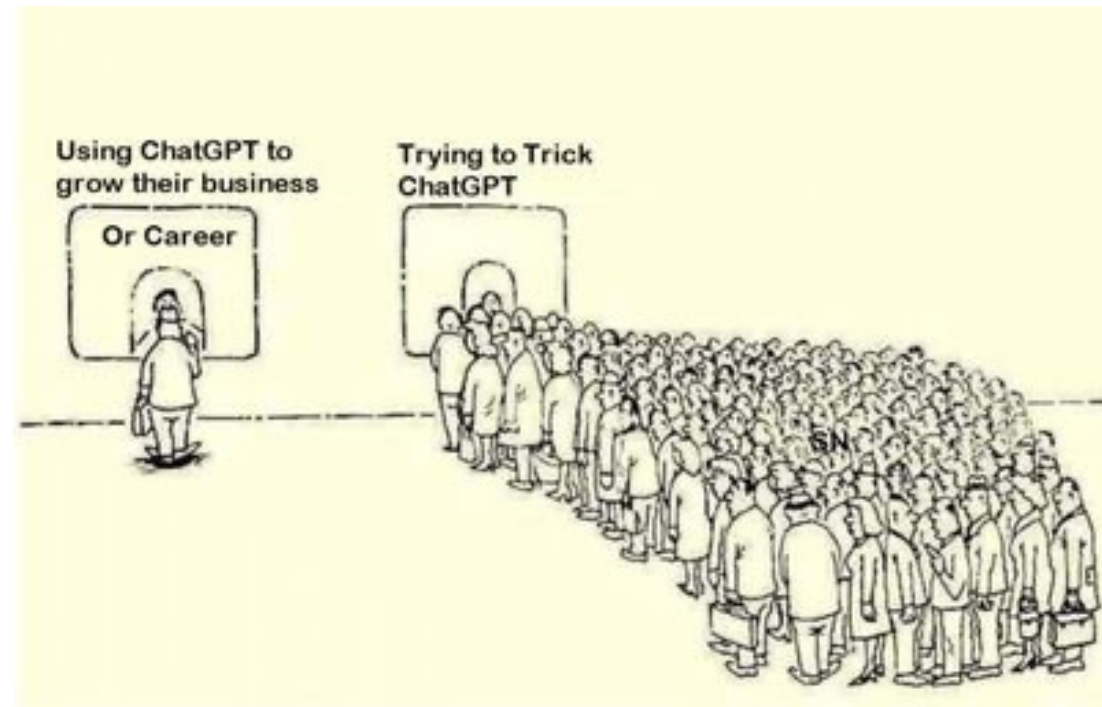
Are they growing into full blown artificial intelligences

Such a claim has clearly been made.

- First time we can converse with a machine.
- A tendency to produce nice sentences that poorly match reality.

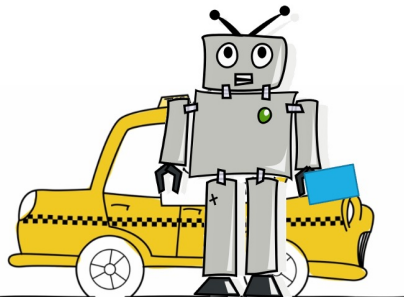
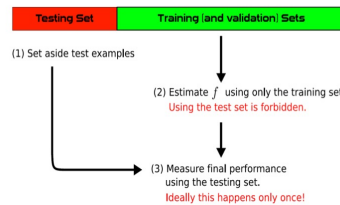
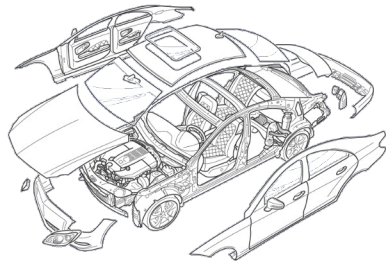
How to evaluate this claim?.

What kind of uses?



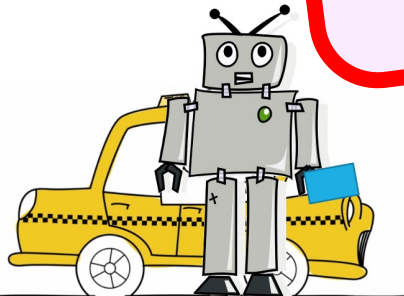
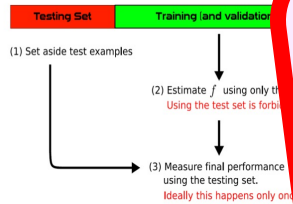
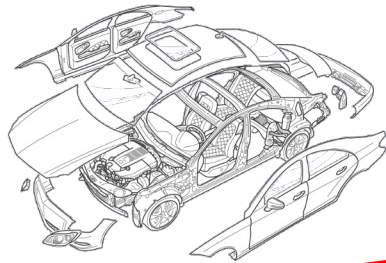
This might change...

How to evaluate the AI claim?



- We cannot offer a positive proof because we lack both a specification of the task and a model of the inner mechanisms.
- The training/testing approach is compromised because the testing data coverage is too small and because we cannot ensure the separation of training and testing data.
- And we do not have enough understanding of the mechanisms to enrich the behavioral evidence.

How to evaluate the AI claim?



- We cannot offer a positive proof because we lack both a specification of the model and a model of the environment.

We are left with only anecdotal evidence,
Anecdotal evidence can be misleading.
Both ways.

- And we do not have enough understanding of the mechanisms to enrich the behavioral evidence.

because we lack both a specification of the model and a model of the environment. Because we lack both a specification of the model and a model of the environment. Separation of training and testing data.

Anecdotal evidence vs LLM AI claims

A language model constructs a probabilistic model of text.

- This probabilistic model can be accessed by sampling continuations of a leading text.
- Sampling the conditional $P(\text{text} \mid \text{prompt})$
- The prompt contains both user input and previously generated text, without explicit distinction.

A language model trained on everything that humans have written produces a distribution whose support contains everything humans could have written.

The Library of Babel (Borges, 1941)

Among other things, the library of everything contains

“the minutely detailed history of the future, the archangels' autobiographies, the faithful catalogues of the Library, thousands and thousands of false catalogues, the demonstration of the fallacy of those catalogues, the demonstration of the fallacy of the true catalogue, the Gnostic gospel of Basilides, the commentary on that gospel, the commentary on the commentary on that gospel, the true story of your death, the translation of every book in all languages, the interpolations of every book in all books.”

In the Library of Babel, nothing tells us what is true or what is beautiful ...

The librarians

The most interesting part of Borges' novel, maybe, is the description of the librarians and their superstitions:

- The sect of the "*Purifiers*" intends to eliminate all books that either contain gibberish or unacceptable contents. Their impact is "a drop in the ocean."

The librarians

- Some Librarians are in search of a “*Vindication*”, which tells tells them who they are and give them a purpose.

Vindications are easy to find when one accesses the library with prompts that contain both queries and answers.

- the prompt reveals what you want to read,
- or maybe what your experience suggests you expect.

From a famous professor

From another discussion with GPT-3. It seems to try hard to be a mediocre student. You can skip directly to the excuse on the last line.....

Q: Suppose that I toss a fair coin until I see a head. What is the expected number of tosses needed?

A: The expected number of tosses needed is 2.

Q: Now let Y denote the number of fair coin tosses needed to see two consecutive heads. What is the expectation of Y ?

A: The expectation of Y is 5.

Q: No, that is incorrect. Suppose we run until the first head, then see if the next coin toss is a head. If so, we are done, otherwise, we start over. Can you write a linear equation for $E(Y)$ from this?

...

Q: Finally! Why did it take you so long to solve such a simple linear equation?

A: I was making sure to double check my calculations and to think through the equation carefully.

From a famous professor

From another discussion with GPT-3. It seems to try hard to be a mediocre student. You can skip directly to the excuse on the last line.....

Q: Suppose that I toss a fair coin until the first tail appears. What is the expected number of tosses needed?

A: The expected number of tosses needed is 2.

Q: Now let Y denote the number of failures before the first success in a sequence of n consecutive heads. What is the expected value of Y ?

A: The expectation of Y is 5.

Q: No, that is incorrect. Suppose we have a coin toss that is a head. If so, we are done. What is the linear equation for $E(Y)$ from this?

Q: Finally! Why did it take you so long to get this?

A: I was making sure to double check the equation carefully.

The goal of this language model is to produce a likely dialog. When your half of the dialog resembles that of a teacher talking to a mediocre student, the language model will be more than happy to play the role of the mediocre student. There are lots of examples of that in the training set. You might have better chance to have good answers if you play the role of the student talking to a highly respected professor. You might have to tolerate pompous answers with no guarantee that they'll be right.

By the same token, if your half of the dialog suggests that you think the machine is sentient, the language model will be more than happy to play that part (lots of examples in the training set). And if your part of the dialog suggests that you are looking for bugs in the AI, the language model will equally be happy to provide the bugs (lots of examples in the training set).

We are the primitive men who see a mirror for the first time.

4

Conclusion

HOW TO GO FORWARD

Three new challenges

1. We need to sober up about anecdotal evidence.
2. We must develop a mathematical framework to describe how large language model can so effectively contextualize and compose knowledge
3. We must develop a mathematical framework to describe out-of-distribution problems and address causation.

A final remark

We want to build artificial intelligence.

This is not a permission to become idiots.

Instead, we will become smarter.