



Od magnetů k ChatGPT: Rozumíme umělé inteligenci?

Lenka Zdeborová



SPOC 
Statistical Physics Of Computation

EPFL

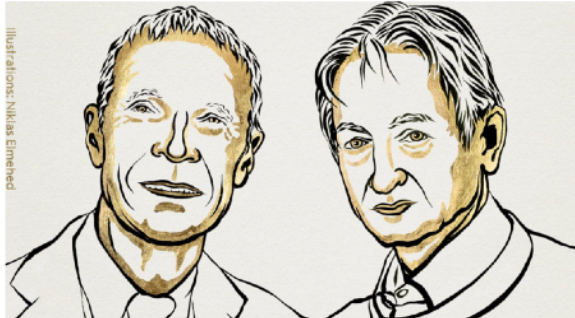
Poslední dvě Nobelovy ceny za teoretickou fyziku.

2021



Objevy Giorgia Parisiho umožňují porozumět a popsat mnohé na první pohled zcela náhodné komplexní materiály a jevy, a to nejen ve fyzice, ale také v dalších, velmi rozdílných oblastech, jako jsou matematika, biologie, neurovědy a **strojové učení**.

2024



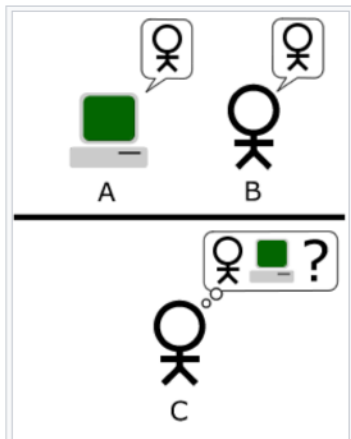
John Hopfield a Geoffrey Hinton využili nástroje fyziky k rozvoji metod, které tvoří základ dnešního **strojového učení**.

Co musí stroj zvládnout, aby byl “inteligentní”?

~1950



“The Imitation Game”.
Turingův test



Standardní podoba Turingova testu⁵¹
– osoba C má za úkol rozlišit, kdo z A
a B je člověk a kdo je stroj
napodobující člověka

~2020

Drtivá většina expertů:

“Systém, který by prošel
Turingovým testem, je
dekády daleko!”

Dnes

Listopad 2022:



Large Language Models Pass the Turing Test

Cameron R. Jones
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
cameron@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
bkbergen@ucsd.edu

2025

Large Language Models Pass the Turing Test

Cameron R. Jones

Department of Cognitive Science
UC San Diego
San Diego, CA 92119
cameron@ucsd.edu

Benjamin K. Bergen

Department of Cognitive Science
UC San Diego
San Diego, CA 92119
bkbergen@ucsd.edu

Abstract

We evaluated 4 systems (ELIZA, GPT-4o, LLaMa-3.1-405B, and GPT-4.5) in two randomised, controlled, and pre-registered Turing tests on independent populations. Participants had 5 minute conversations simultaneously with another human participant and one of these systems before judging which conversational partner they thought was human. When prompted to adopt a humanlike persona, GPT-4.5 was judged to be the human 73% of the time: significantly more often than interrogators selected the real human participant. LLaMa-3.1, with the same prompt, was judged to be the human 56% of the time—not significantly more or less often than the humans they were being compared to—while baseline models (ELIZA and GPT-4o) achieved win rates significantly below chance (23% and 21% respectively). The results constitute the first empirical evidence that any artificial system passes a standard three-party Turing test. The results have implications for debates about what kind of intelligence is exhibited by Large Language Models (LLMs), and the social and economic impacts these systems are likely to have.

REVIEW

2015

doi:10.1038/nature14539

nature

Deep learning

Yann LeCun^{1,2}, Yoshua Bengio³ & Geoffrey Hinton^{4,5}

NATURE VOL. 323 9 OCTOBER 1986

— LETTERS TO NATURE —

Learning representations by back-propagating errors

1986

David E. Rumelhart*, **Geoffrey E. Hinton†**
& **Ronald J. Williams***

* Institute for Cognitive Science, C-015, University of California,
San Diego, La Jolla, California 92093, USA

† Department of Computer Science, Carnegie-Mellon University,
Pittsburgh, Philadelphia 15213, USA

Communicated by Dana Ballard

1989

Backpropagation Applied to Handwritten Zip Code Recognition

Y. LeCun
B. Boser
J. S. Denker
D. Henderson
R. E. Howard
W. Hubbard
L. D. Jackel

AT&T Bell Laboratories Holmdel, NJ 07733 USA

Klasifikace obrázků



701.jpg



702.jpg



703.jpg



704.jpg



705.jpg



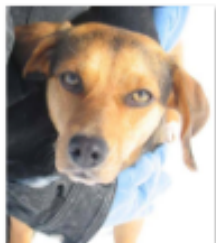
706.jpg



707.jpg



708.jpg



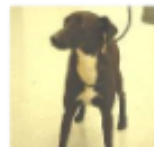
709.jpg



710.jpg



711.jpg



712.jpg



713.jpg



714.jpg



715.jpg



716.jpg



717.jpg



718.jpg



Klasifikace obrázků



$= X_\mu = (01001010 \ 01110011 \ 10001100 \ 01001011$
 $01110000 \ 10001100 \ \dots \text{všechny pixely} \ \dots)$

Účel: Najdi funkci $f_W(X_\mu)$ takovou, aby

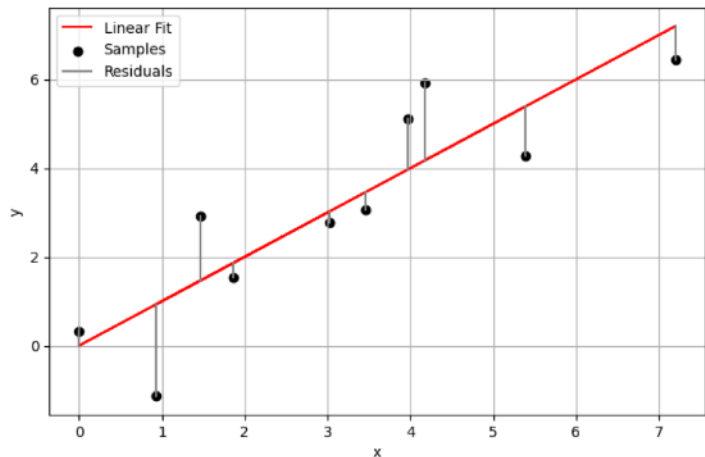
$$f_W(X_\mu) = +1$$

pro obrázek X_μ kočky.

$$f_W(X_\mu) = -1$$

pro obrázek X_μ psa.

Lineární regrese



vstupní data

$$\{X_{\mu}\}_{\mu=1}^n$$

výstupní data

$$\{y_{\mu}\}_{\mu=1}^n$$

parametry:

$$W = \{w_1, w_2\}$$

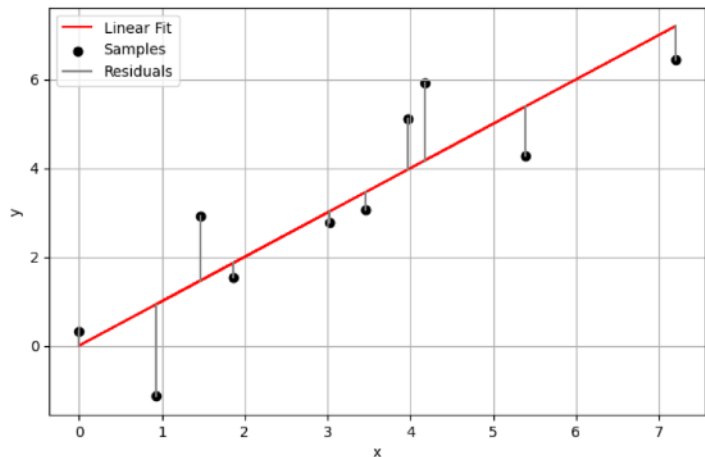
lineární funkce:

$$f_W(X_{\mu}) = w_1 X_{\mu} + w_2$$

Metoda nejmenších čtverců:

$$\min_W \sum_{\mu=1}^n \left[y_{\mu} - f_W(X_{\mu}) \right]^2$$

Lineární regrese



Metoda nejmenších čtverců:

vstupní data

$$\{X_{\mu}\}_{\mu=1}^n$$

výstupní data

$$\{y_{\mu}\}_{\mu=1}^n$$

parametry:

$$W = \{w_1, w_2\}$$

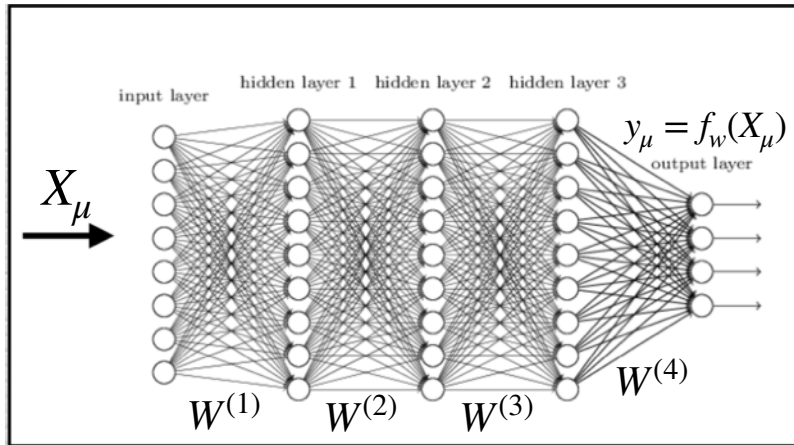
nelineární funkce:

$$f_W(X_{\mu}) = \cos(w_1 X_{\mu}) + \sin(w_2 X_{\mu})$$

$$\min_W \sum_{\mu=1}^n \left[y_{\mu} - f_W(X_{\mu}) \right]^2$$

Neuronová síť = specifická nelineární funkce

- čtyřvrstvý perceptron: $f_W(X_\mu) = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}X_\mu))))$



vstupní data $X_\mu \in \mathbb{R}^d$

trénovatelné parametry: $W^{(\ell)}$

aktivační funkce $\varphi^{(\ell)}$

výstupní data y_μ

- Trénování AI: n trénovacích vzorků + minimalizace “ztráty” sestupem po gradientu:

$$\min_W \mathcal{L}(W); \quad \mathcal{L}(W) = \sum_{\mu=1}^n [y_\mu - f_W(X_\mu)]^2$$

Jazykové modely používají: Transformer

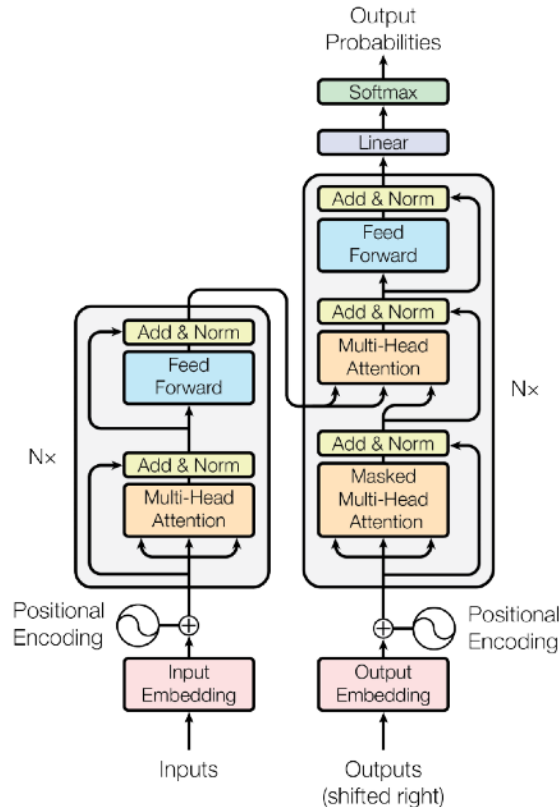


Figure 1: The Transformer - model architecture.

Attention Is All You Need

NeurIPS'17, arXiv:1706.03762

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Attention is all you need

[A Vaswani](#), [N Shazeer](#), [N Parmar](#)... - Advances in neural ..., 2017 - proceedings.neurips.cc

... to attend to **all** positions in the decoder up to and including that position. **We need** to prevent ... **We** implement this inside of scaled dot-product **attention** by masking out (setting to $-\infty$) ...

☆ Save 📄 Cite Cited by 249713 Related articles All 71 versions 🔗

Ingredience jazykových modelů

(Attention is all you need, Vaswani et al.'17)

- Vstupní data = sekvence tokenů/slov.
 - **token** = slovo reprezentované vektorem
 - **sekvence** = text slov/tokenů
- Výstupní data = **následující token/slovo**.

- **Attention** vrstva v transformerech

$$\tilde{x} = \text{softmax} \left[\underbrace{\frac{1}{\sqrt{r}}(xw_Q)(xw_K)^T}_{\text{attention}} \right] xw_V$$

trénované parametry:

$$w_Q, w_K, w_V$$

Předchozí hluboké učení

(vícevrstvé perceptrony)

- Vstupní data = vektor
- Výstupní data explicitně daná
- Plně propojená vrstva

$$\tilde{x} = \sigma(Wx)$$

Dvě fáze trénování

- “**Předtrénování**”:

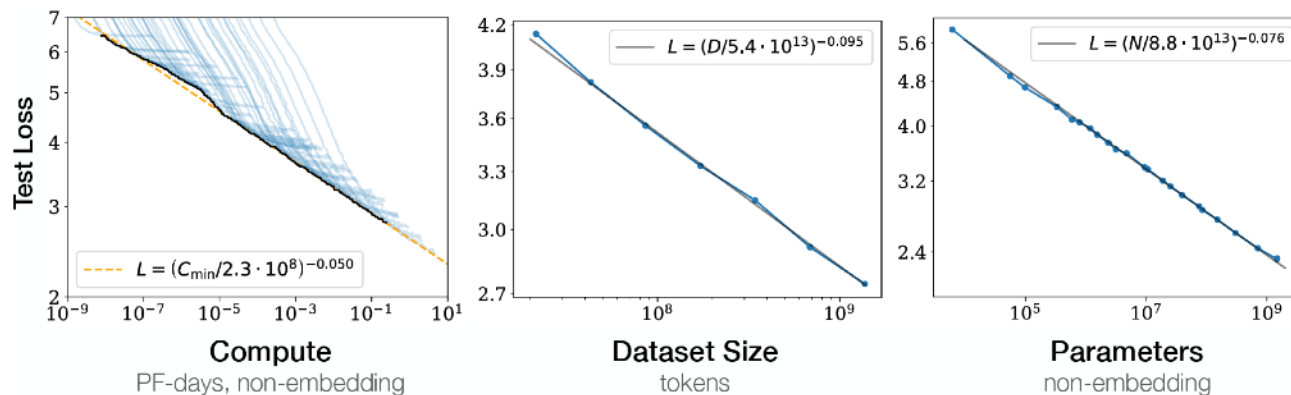
Model se z obrovského množství textu naučí předpovídat další token/slovo. Tím se naučí jazyk a velkou část struktury světa, kterou jazyk popisuje.

- “**Potrénování**”:

Model se pomocí lidské zpětné vazby naučí, jaké typy a styl odpovědí lidé preferují. Tím se naučí komunikovat s lidmi.

Klíčovou ingrediencí, která posouvá AI vpřed, je... škála!

- Více výpočetního výkonu (GPU), větší datasety, více učitelných parametrů.
- Škálovací zákony (scaling-laws) jsou hlavním hnacím motorem současného AI průmyslu.



Kaplan et al.'20

Velké vědecké překvapení.

Předpovídáním dalšího slova a škálou vznikly schopnosti, které jsme dříve považovali za mnohem hlubší.

- Jazykové modely se naučily: překlad, vysvětlování, argumentace, práce s kontextem, shrnování, programování, matematiku atd.

Fundamentální otázky

Je obrovská škála jediná možná cesta?

- Lidský mozek funguje se zlomkem energie a dat.

Jaká jsou fundamentální omezení učení?

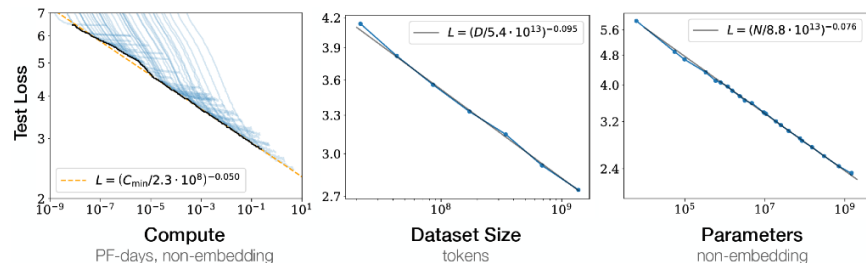
(Určená statistickými a výpočetními podmínkami.)

Škála přináší další otázky

Kaplan et al.'20

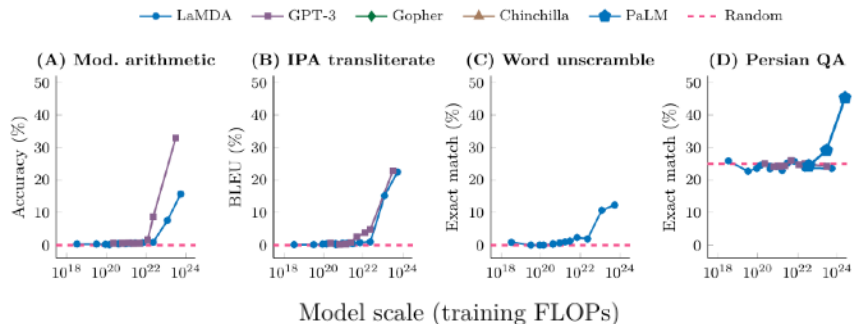
- **Výkon se řídí zákony škálování.**

➡ Jak odvodit jejich exponenty?



- **Schopnosti se vynořují se škálou.**

➡ Jsou to fázové přechody?



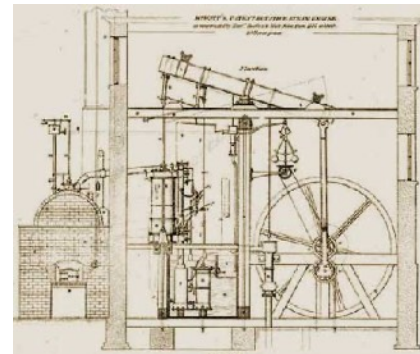
Wei et al.'22

- **Prokletí vysoké dimenze** = trénování neuronových sítí je výpočetně těžké (Blum, Rivest'89).

➡ Můžeme charakterizovat, kdy se lze obtížnosti vyhnout?

Kde porozumění AI stojí dnes?

Analogie s parním strojem



- **Denis Papin (1679)** – tlakový hrnec, první pokusy využít páru → krásná demonstrace principu.
- **Thomas Newcomen (1712)** – „atmosférický stroj“ na odčerpávání vody z dolů → první široce nasazená technologie, ale neefektivní.
- **James Watt (1763–1775)** – klíčové inženýrské zlepšení → stroj se stal účinným, univerzálně použitelným, odstartoval průmyslovou revoluci.
- **Sadi Carnot (1824)** – teorie účinnosti tepelného stroje → první hluboké porozumění principům a limitům.
- **Maxwell, Gibbs, Boltzmann (konec 19. stol.)** – rozvoj termodynamiky a statistické fyziky → obecné zákony nad rámec konkrétní technologie. Statistické myšlení odhalilo limity klasické fyziky a otevřelo cestu ke kvantové teorii. Entropie z fyziky se stala klíčovým konceptem pro měření a zpracování informace.

Parní stroj

Umělá inteligence

Denis Papin (1679) – tlakový hrnec, → krásná demonstrace principu. Chybí industriální využití.

Thomas Newcomen (1712) – „atmosférický stroj“ → první široce nasazená technologie, ale neefektivní.

James Watt (1760s) – klíčové inženýrské zlepšení → stroj univerzálně použitelný, **odstartoval průmyslovou revoluci.**

Sadi Carnot (1824) – teorie účinnosti tepelného stroje, porozumění principům a limitům.

Maxwell, Gibbs, Boltzmann (konec 19. stol.) – rozvoj statistické fyziky. Obecné zákony nad rámec technologie.

Parní stroj

Denis Papin (1679) – tlakový hrnec, → krásná demonstrace principu. Chybí industriální využití.

Thomas Newcomen (1712) – „atmosférický stroj“ → první široce nasazená technologie, ale neefektivní.

James Watt (1760s) – klíčové inženýrské zlepšení → stroj univerzálně použitelný, **odstartoval průmyslovou revoluci.**

Sadi Carnot (1824) – teorie účinnosti tepelného stroje, porozumění principům a limitům.

Maxwell, Gibbs, Boltzmann (konec 19. stol.) – rozvoj statistické fyziky. Obecné zákony nad rámec technologie.

Umělá inteligence

John Hopfield (1982) – energie a asociativní paměť → krásná demonstrace principu. Chybí konkrétní využití.

Geoffrey Hinton & kol. (1980s) – Boltzmannova mašina → první široce nasazené metody, ale neefektivní.

Parní stroj

Denis Papin (1679) – tlakový hrnec, → krásná demonstrace principu. Chybí industriální využití.

Thomas Newcomen (1712) – „atmosférický stroj“ → první široce nasazená technologie, ale neefektivní.

James Watt (1760s) – klíčové inženýrské zlepšení → stroj univerzálně použitelný, **odstartoval průmyslovou revoluci.**

Sadi Carnot (1824) – teorie účinnosti tepelného stroje, porozumění principům a limitům.

Maxwell, Gibbs, Boltzmann (konec 19. stol.) – rozvoj statistické fyziky. Obecné zákony nad rámec technologie.

Umělá inteligence

John Hopfield (1982) – energie a asociativní paměť → krásná demonstrace principu. Chybí konkrétní využití.

Geoffrey Hinton & kol. (1980s) – Boltzmannova mašina → první široce nasazené metody, ale neefektivní.

Deep learning & LLM (od ~2012) – klíčová inženýrská zlepšení → univerzálně použitelné, **současná AI revoluce.**

Parní stroj

Denis Papin (1679) – tlakový hrnec, → krásná demonstrace principu. Chybí industriální využití.

Thomas Newcomen (1712) – „atmosférický stroj“ → první široce nasazená technologie, ale neefektivní.

James Watt (1760s) – klíčové inženýrské zlepšení → stroj univerzálně použitelný, odstartoval průmyslovou revoluci.

Sadi Carnot (1824) – teorie účinnosti tepelného stroje, porozumění principům a limitům.

Maxwell, Gibbs, Boltzmann (konec 19. stol.) – rozvoj statistické fyziky. Obecné zákony nad rámec technologie.

Umělá inteligence

John Hopfield (1982) – energie a asociativní paměť → krásná demonstrace principu. Chybí konkrétní využití.

Geoffrey Hinton & kol. (1980s) – Boltzmannova mašina → první široce nasazené metody, ale neefektivní.

Deep learning & LLM (od ~2012) – klíčová inženýrská zlepšení → univerzálně použitelné, současná AI revoluce.

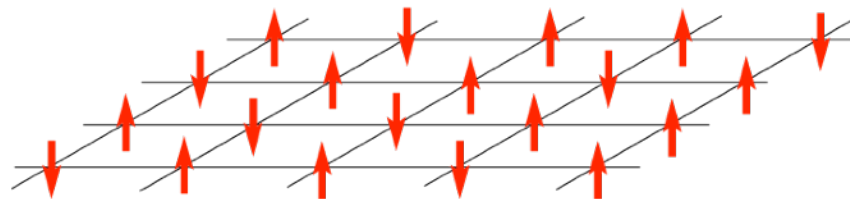
Porozumění možnostem a omezením? Chybí.

Obecné zákony inteligence? Zatím nemáme.

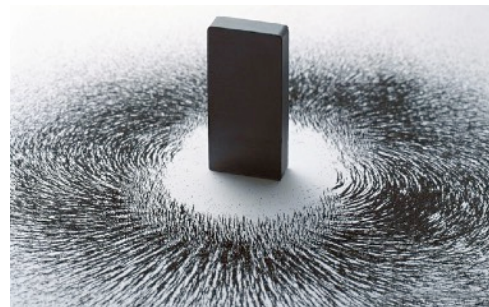
Škálování přináší také odpovědi

- **Požehnání vysoké dimenze** (Donoho'00): Některé škálovací limity umožňují popis jen pomocí několika málo “řádových parametrů”.
 - ➔ **Statistická fyzika** (Gibbs'1902) toho využívá k popisu “nekonečně” velkých systémů atomů a molekul.
 - ➔ **Fázové přechody** (Lenz, Ising'1920s; Onsager'1944) — náhlé změny v chování, které se formálně objevují při “nekonečných” velikostech.

Isingův model'1925, fyzikální pochopení magnetismu



2-D Ising Model



spiny $S_i \in \{-1, +1\}$ $i = 1, \dots, N$

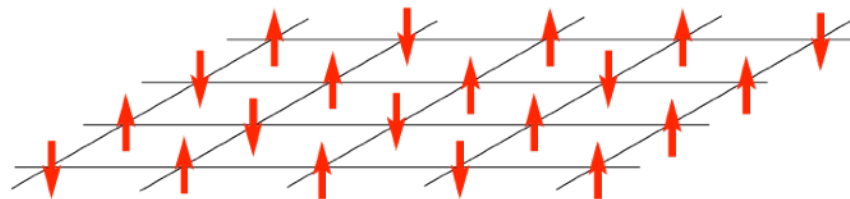
energie $E = -J \sum_{(ij) \in E} S_i S_j$ interakční konstanta $J > 0$

teplota

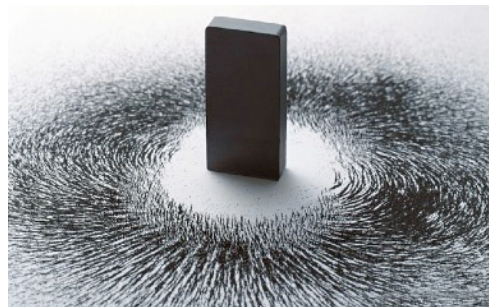
dynamika: Vyber spin, otoč ho do směru většiny jeho sousedů s pravděpodobností $\rho(J, T, \sum_{j \in N(i)} S_j)$.

vyber mu náhodný směr s pravděpodobností $1-\rho$,

Isingův model'1925, fyzikální pochopení magnetismu



2-D Ising Model



spiny $S_i \in \{-1, +1\}$ $i = 1, \dots, N$

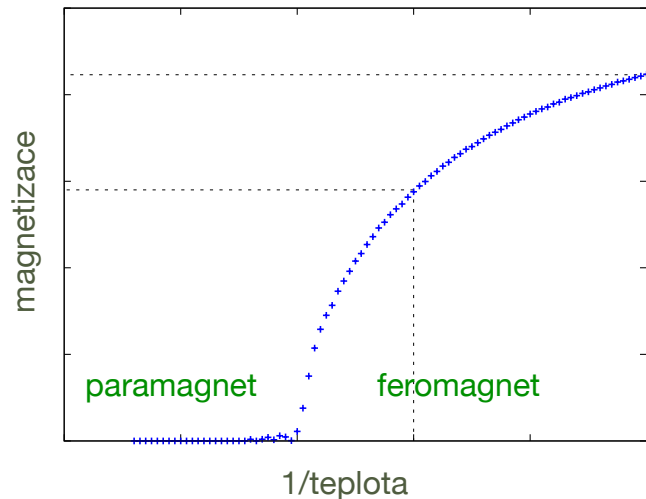
energie $E = -J \sum_{(ij) \in E} S_i S_j$ interakční konstanta $J > 0$

dynamika: Teplota 0 K: vždy otoč spin do směru většiny jeho sousedů $S_i^{t+1} = \text{sign}(\sum_{j \in N(i)} S_j^t)$.

Energie v každém kroku klesá nebo stagnuje.

Isingův model'1925, fázový přechod

Fázový přechod vznikne spontánně z toho, že je počet atomů/spinů velmi velký, $N \rightarrow \infty$.



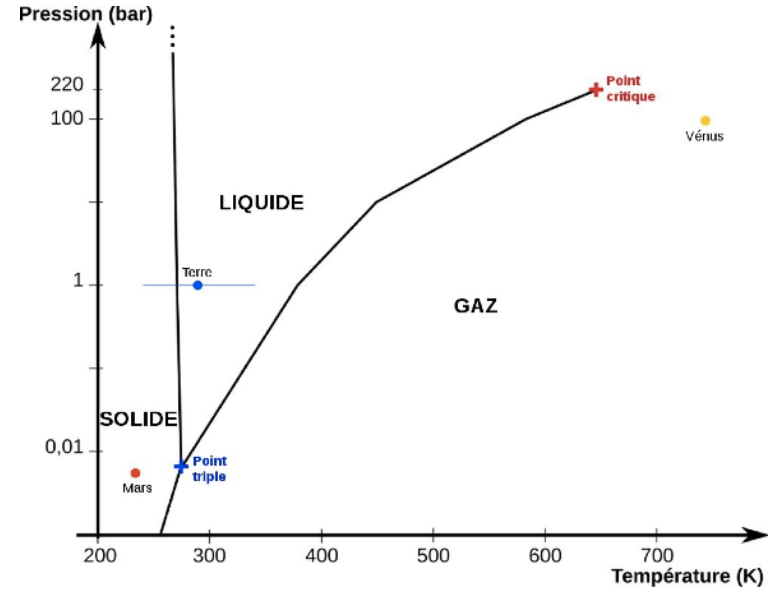
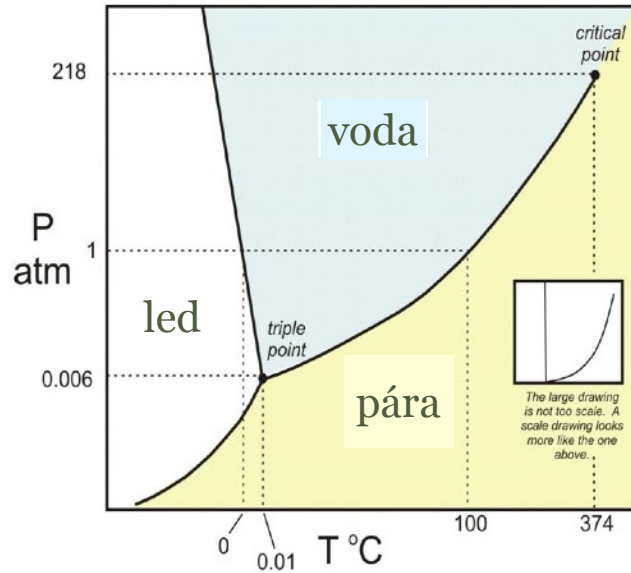
$$P(\{S_i\}_{i=1}^N) = \frac{1}{Z(T)} e^{\frac{1}{k_B T} J \sum_{(ij) \in E} S_i S_j}$$

Fázový přechod = nespojitost derivací $\lim_{N \rightarrow \infty} \frac{\log Z(T)}{N}$

Vysvětlení toho, že železo se stane paramagnetickým nad $T_{\text{Curie}}=1043\text{K}$.

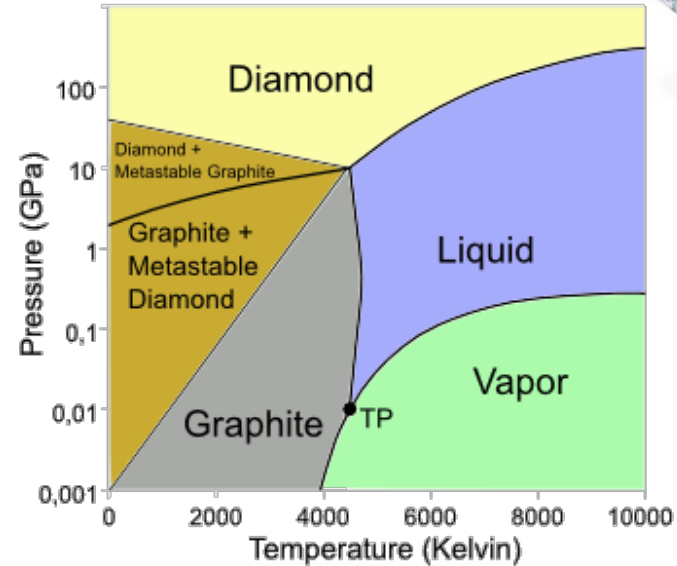
Fázový diagram vody (přechod prvního druhu)

Fázový diagram



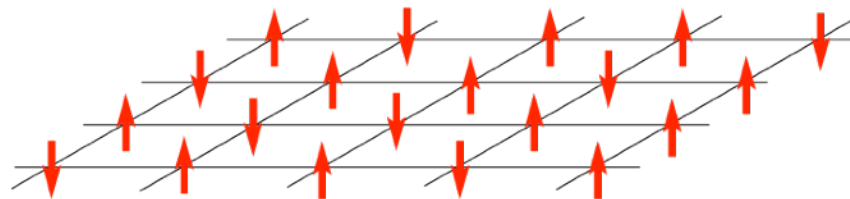
Metastabilita (u fázových přechodů prvního druhu)

Podchlazený (metastabilní) déšť mrzne v okamžiku při doteku s nečistotou.



Diamant za normálního tlaku a teploty je metastabilní. Grafit (tuha v tužce) je stabilní.

Spinové sklo '1975, komplexní a neuspořádané systémy



2-D Ising Model



spiny $S_i \in \{-1, +1\}$ $i = 1, \dots, N$

energie $E = - \sum_{(ij) \in E} J_{ij} S_i S_j$ interakční konstanta $J_{ij} \in \left\{ \frac{-J}{\sqrt{N}}, \frac{+J}{\sqrt{N}} \right\}$

Užitečný model pro celou řadu tzv. neuspořádaných systémů.

Budování teorie AI pomocí zjednodušených modelů

AI, až ji pochopíme



AI dnes (& mozek)

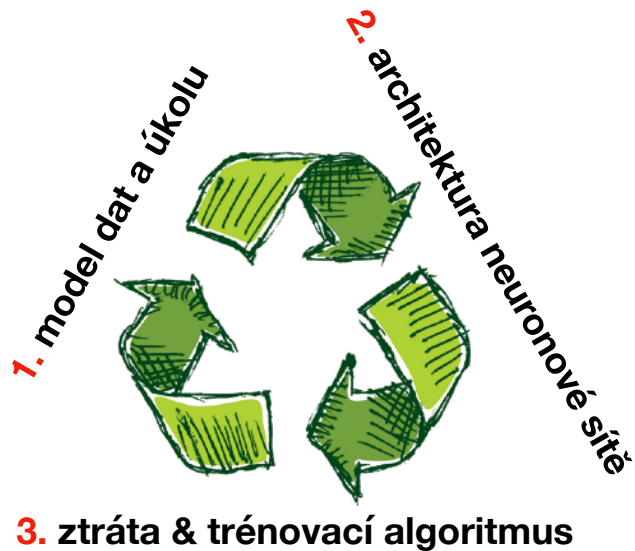


AI teorie



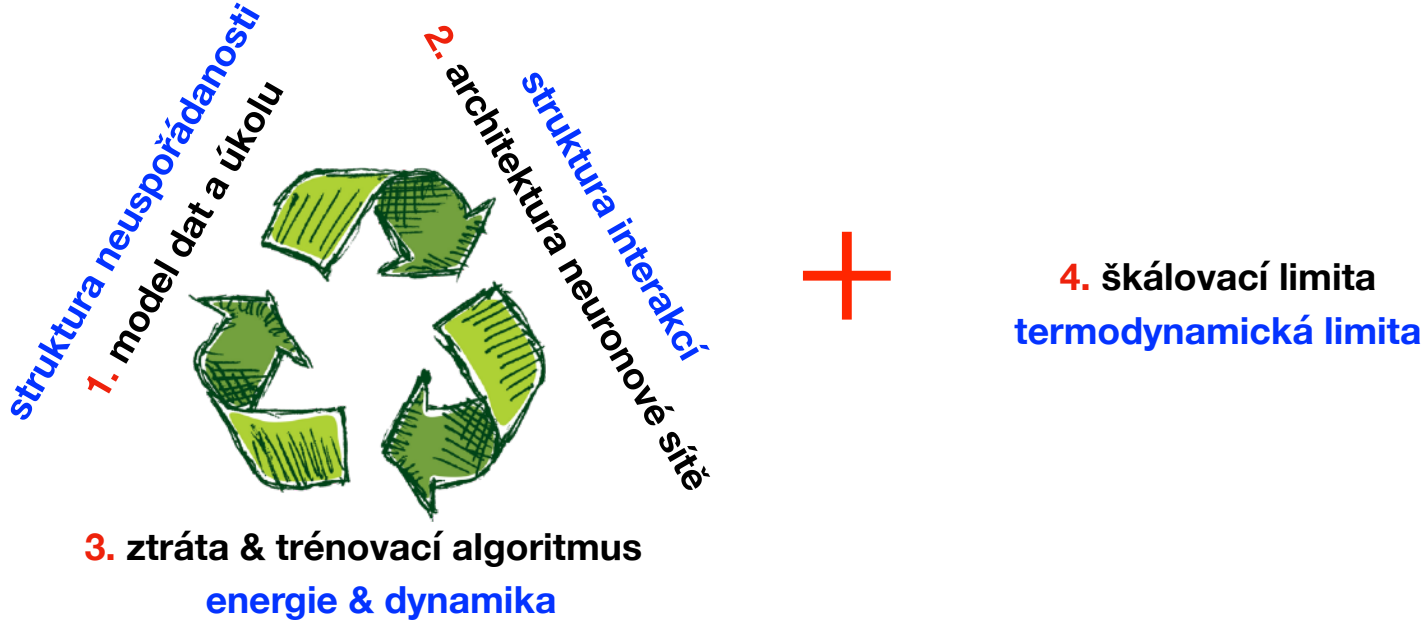
Porozumět optimální AI může být snazší než pochopit AI systémy současné.

Zjednodušené modely AI: čtyři **neoddělitelné** ingredience



4. škálovací limita

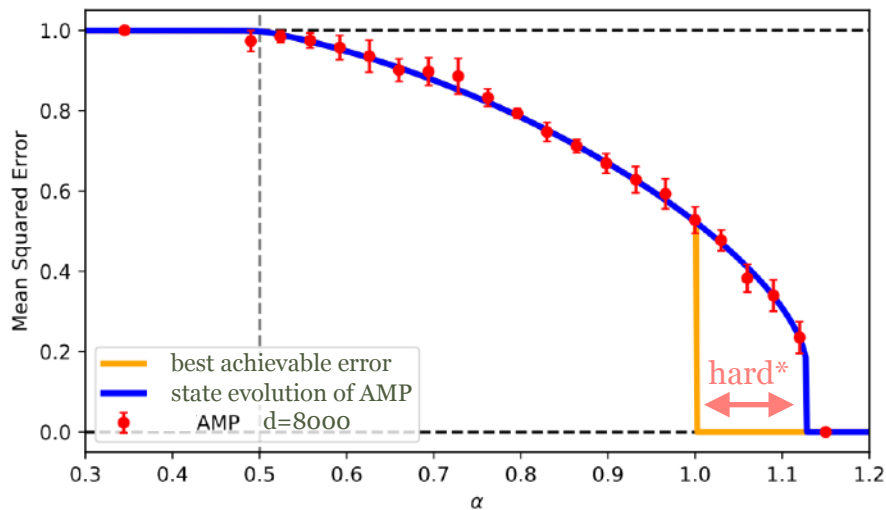
Zjednodušené modely AI = “spinová skla”



učitelné váhy = “spiny”

Učení v jednovrstvém perceptronu

Maillard, Loureiro, Krzakala, LZ, arXiv:2006.05228, NeurIPS'20.



$$y_{\mu} = \left| \sum_{i=1}^d X_{\mu i} w_i^* \right|^2$$

$$w_i^* \sim \mathcal{N}(0,1)$$

$$X_{\mu i} \sim \mathcal{N}(0,1/d)$$

$$\alpha = \frac{n}{d}, n \rightarrow \infty$$

$$\alpha_{\text{WR}} = 1/2$$

$$\alpha_{\text{IT}} = 1$$

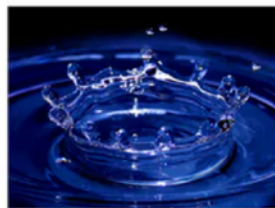
$$\alpha_{\text{AMP}} = 1.13$$

počet vzorů potřebných, aby učení začalo, **fázový přechod druhého druhu**.

počet vzorů potřebných, aby učení mohlo být přesné, **f.p. prvního druhu**.

počet vzorů potřebných, aby učení mohlo být přesné a algoritmicky dosažitelné.

Fázové přechody v učení



voda

nemožné



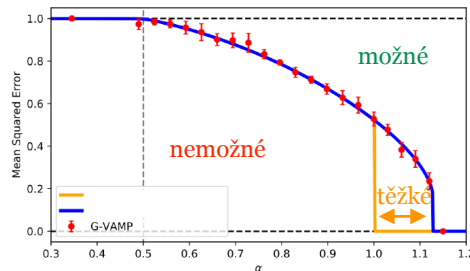
podchlazená voda

algoritmicky těžké



led

možné



Fundamentální statistické a výpočetní omezení učení jsou určeny fázovými přechody.

Parní stroj

Denis Papin (1679) – tlakový hrnec, → krásná demonstrace principu. Chybí industriální využití.

Thomson (1857) – první prakticky použitelný, odstartoval průmyslovou revoluci.
→ **Můj odhad: v 21. století nás čeká ještě celá řada Nobelových cen za porozumění AI.**

James Watt (1769) – univerzálně použitelný, odstartoval průmyslovou revoluci.

Sadi Carnot (1824) – teorie účinnosti tepelného stroje, porozumění principům a limitům.

Maxwell, Gibbs, Boltzmann (konec 19. stol.) – rozvoj statistické fyziky. Obecné zákony nad rámec technologie.

Umělá inteligence

John Hopfield (1982) – energie a asociativní paměť → krásná demonstrace principu. Chybí konkrétní využití.

Geoffrey Hinton & kol. (1980s) – Boltzmannova mašina → první široce nasazené metody, ale neefektivní.

Deep learning & LLM (od ~2012) – klíčová inženýrská zlepšení → univerzálně použitelné, současná AI revoluce.

Porozumění možnostem a omezením? Chybí.

Obecné zákony inteligence? Zatím nemáme.



