Dlaczego teraz?

Ewolucja AI

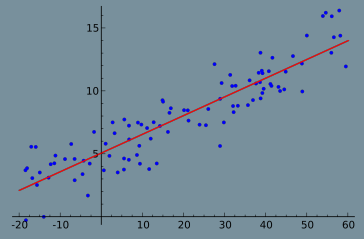
Krzysztof Joachimiak

Zanim powstały komputery

1676
Reguła łańcuchowa

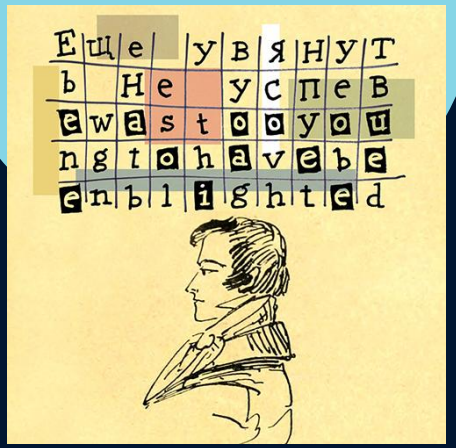
$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

1795-1805
Model liniowy

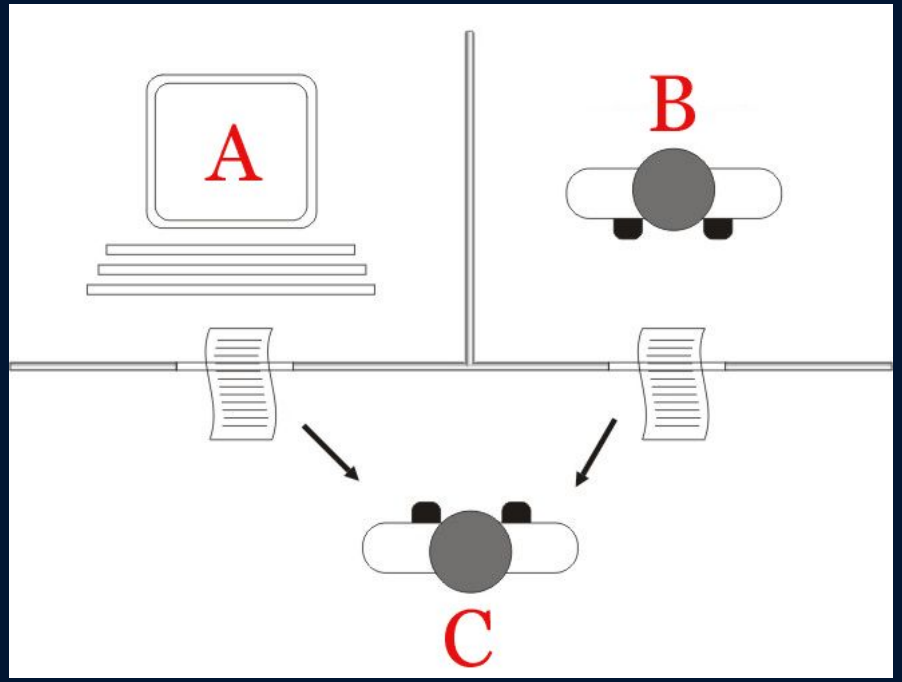


Źródło:
https://paperswithcode.com/method/linear-regression

1913
Samogłoska czy spółgłoska? (A. Markow)



Test Turinga (1950)



Prowadzimy rozmowę z człowiekiem i maszyną próbując zgadnąć, który z rozmówców jest maszyną.

Ewolucja AI

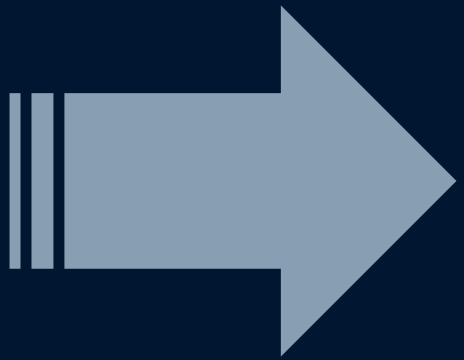
Good
old-fashioned AI

Symboliczne AI

Glass box

Wyspecjalizowanie
modele

1960



Machine Learning

Deep Learning

Black box

Multimodal AI

Artificial General
Intelligence

2020

Chatboty

1966 ELIZA

```
EEEEEEEE  L         IIIIIII  ZZZZZZZ      AAA
E         L           I          Z      A   A
E         L           I         Z      A     A
EEEEE     L           I        Z       A     A
E         L           I       Z        AAAAAAA
E         L           I      Z         A     A
EEEEEEEE  LLLLLLLL  IIIIIII  ZZZZZZ    A     A

ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... ?
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...
```

1995 AIML

```
<category>
  <pattern>JAK MASZ NA IMIĘ</pattern>
  <template>MAM NA IMIĘ <bot name="name"/>.</template>
</category>
<category>
  <pattern>JAK SIĘ NAZYWASZ</pattern>
  <template>
    <srai>JAK MASZ NA IMIĘ</srai>
  </template>
</category>
```

1997 Cleverbot

Question Answering

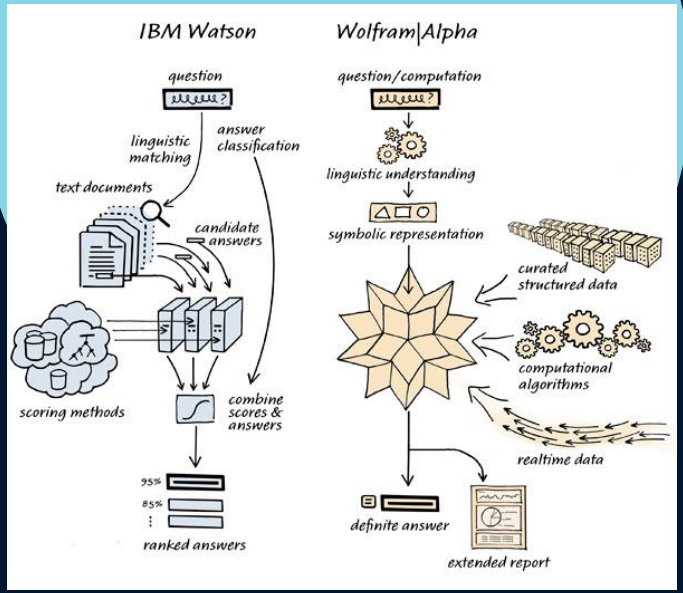
1961
BASEBALL

[How many games] did

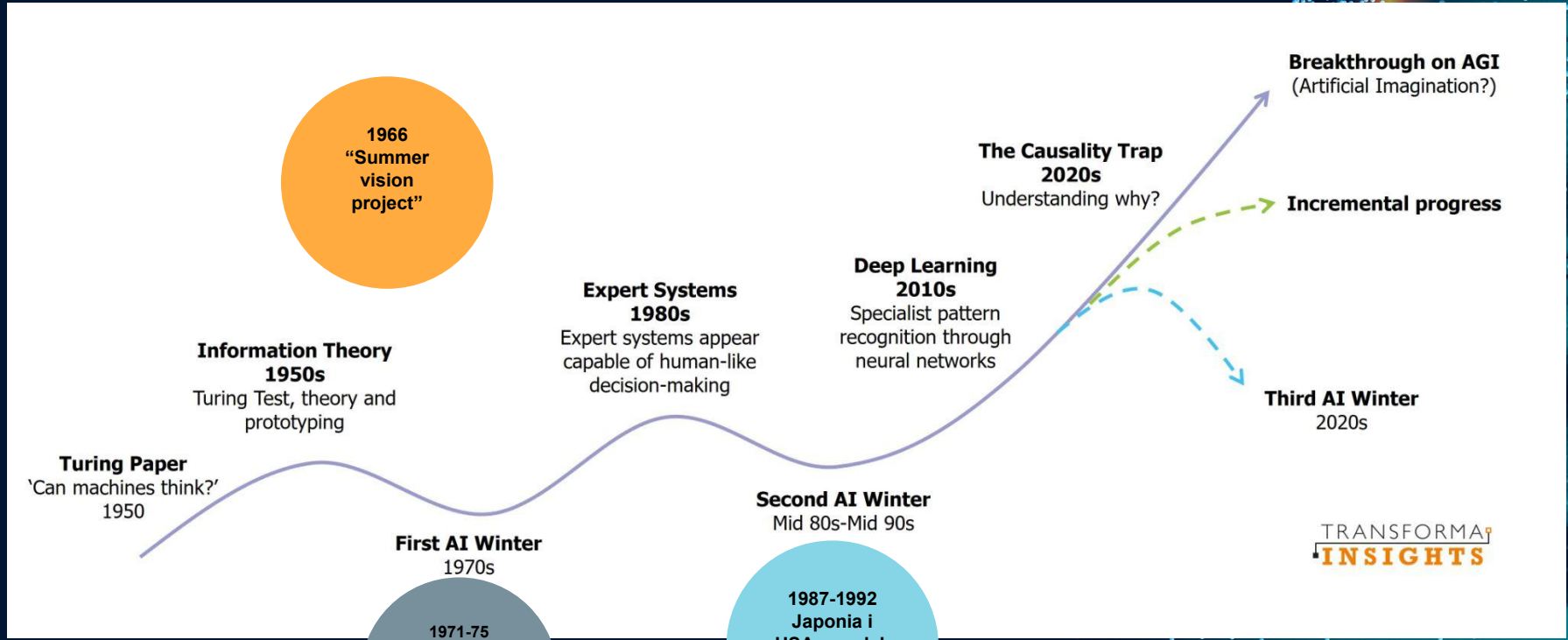
[the Yankees] play (in [July])?

When the question has been bracketed, any un-bracketed preposition is attached to the first noun phrase in the sentence, and prepositional brackets added. For example, "Who did the Red Sox lose to on July 5?" becomes "(To [who]) did [the Red Sox] lose (on [July 5])?"

2008-2011
Wolfram Alpha
IBM Watson



Zimy sztucznej inteligencji



“(Jack) Schwartz believed that DARPA was using a swimming model—setting a goal, and paddling toward it regardless of currents or storms. DARPA should instead be using a surfer model— waiting for the big wave, which would allow its relatively modest funds to surf gracefully and successfully toward that same goal.”

— *Machines Who Think*, Pamela McCorduck

Czynniki wzrostu (lata 90 - teraz)

Internet

- Ogromne wolumeny danych
- Przyspieszenie obiegu wiedzy
- Cloud computing

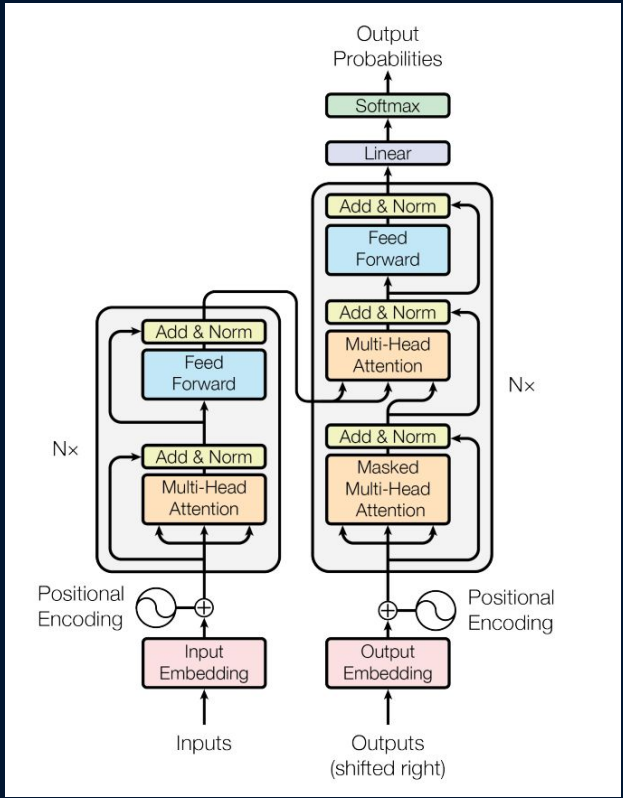
Sprzęt

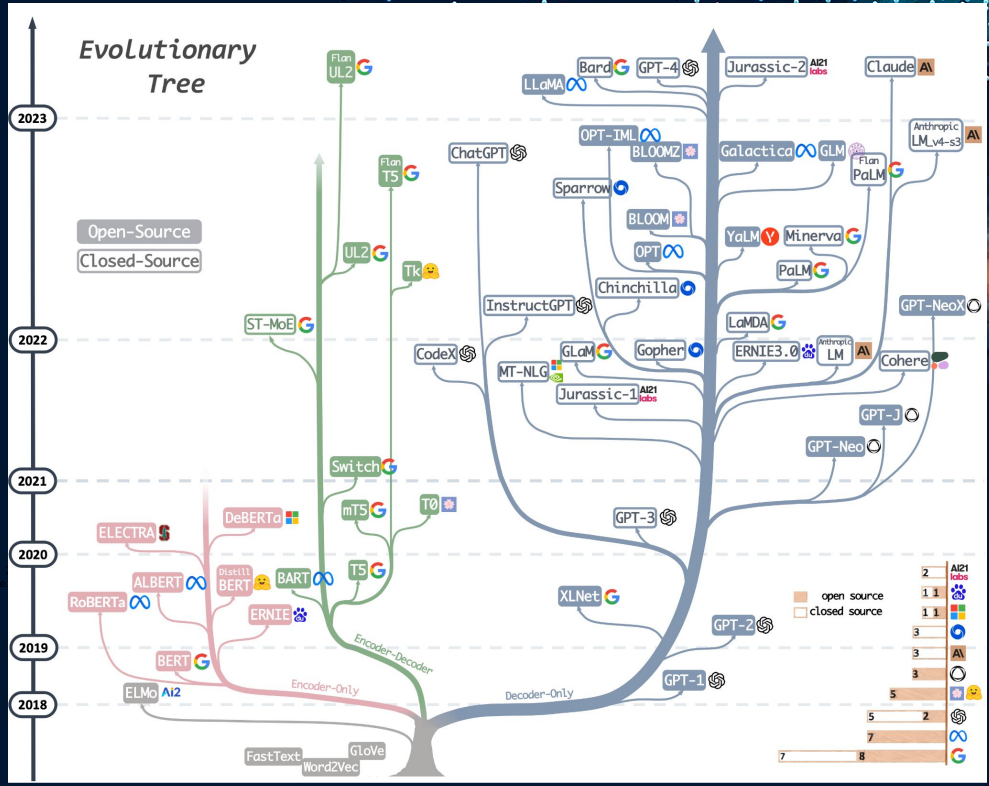
- Prawo Moore'a - wykładniczy wzrost mocy obliczeniowej
- Karty graficzne

Oprogramowanie

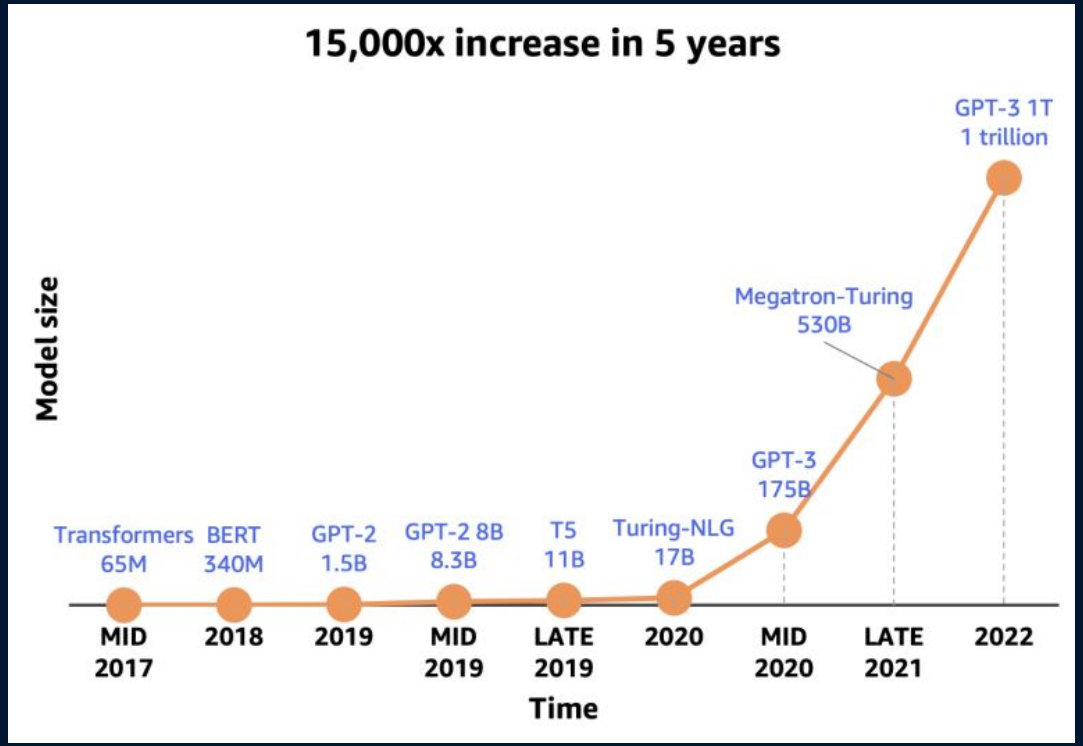
- Python (1991)
- CUDA (2007)
- Keras, Tensorflow (2015)
- PyTorch (2016)

Ewolucja architektury





Zwiększenie rozmiaru modeli



- Można w miarę precyzyjnie prognozować przyrosty jakości modelu
- W coraz większych modelach występuje efekt ulepszeń-niespodzianek
- Parametry, dane, czas treningu

Scaling Laws for Neural Language Models

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

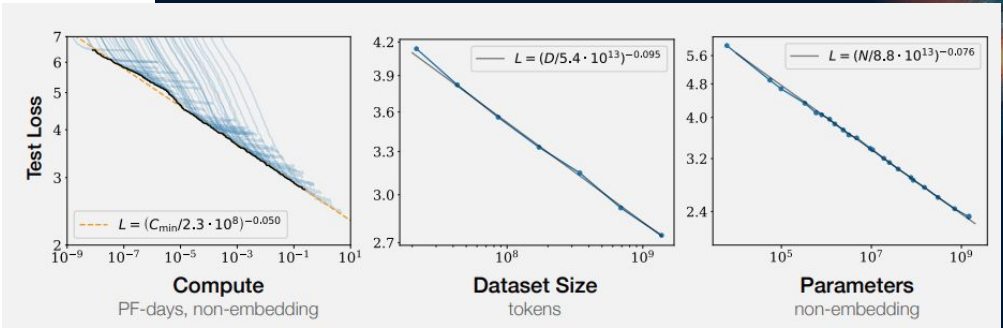
Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com

Abstract

We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These relationships allow us to determine the optimal allocation of a fixed compute budget. Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.



Trend - zmniejszanie modeli

Stanford Alpaca

- Najpopularniejsze techniki:
- Destylowanie wiedzy z większych modeli
- Kwantyzacja większych modeli

llama.cpp

CI passing license MIT

Inference of LLaMA model in pure C/C++

Hot topics:

- Quantization formats `Q4` and `Q8` have changed again (19 May) - (info)
- Quantization formats `Q4` and `Q5` have changed - requantize any old models (info)
- Roadmap May 2023

Dziękuję za uwagę!

Wykorzystano szablon prezentacji z:

slidesgo