

Stefania Centrone – TUM
Cosimo Perini Brogi – IMT Lucca

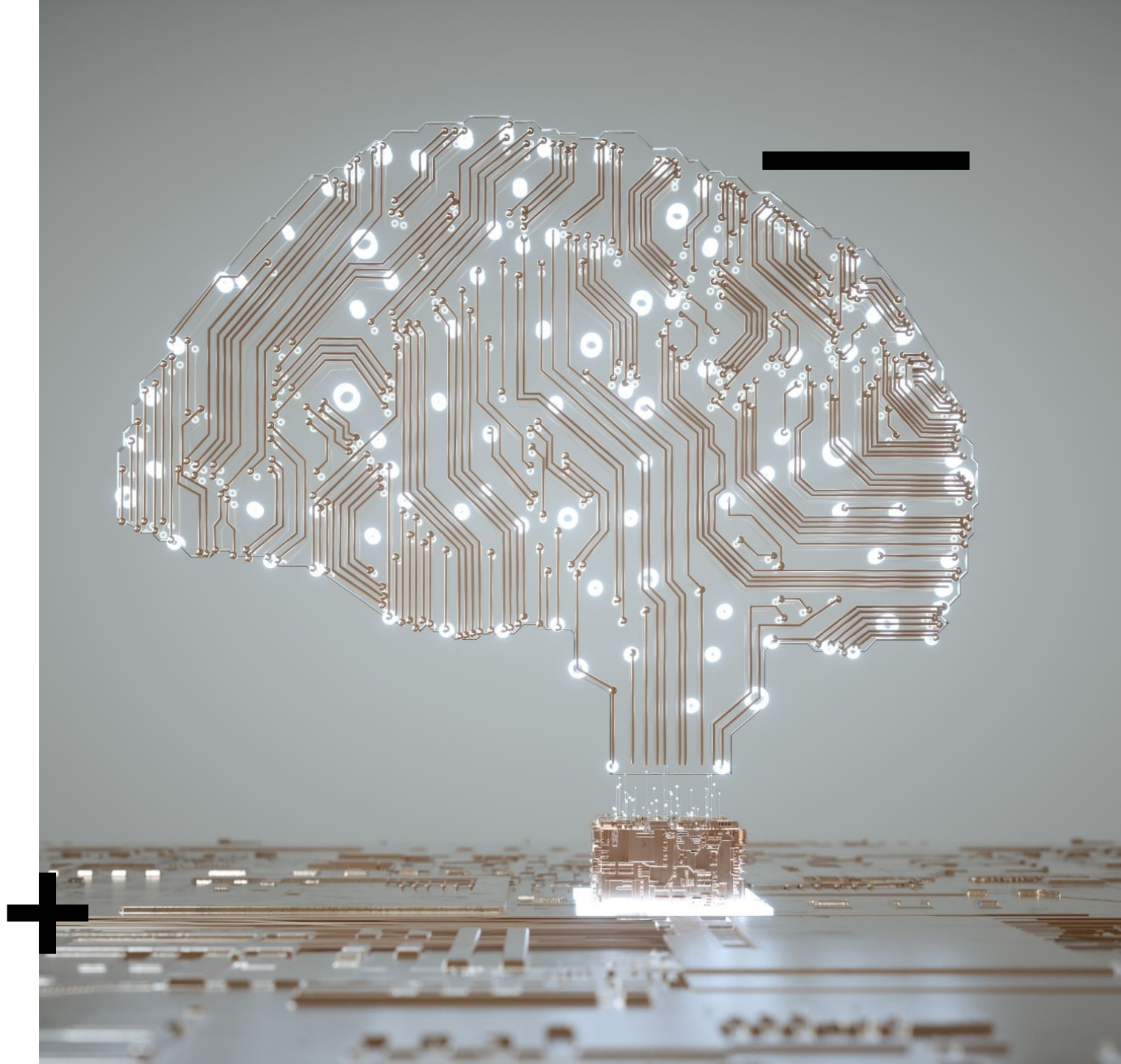


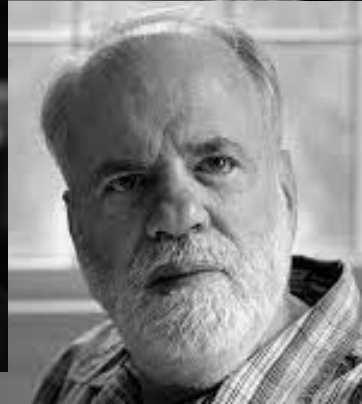
Machine Translation, Problem Solving, and Pattern Recognition:
A Historical-Phenomenological Analysis



Philosophy of Artificial Intelligence

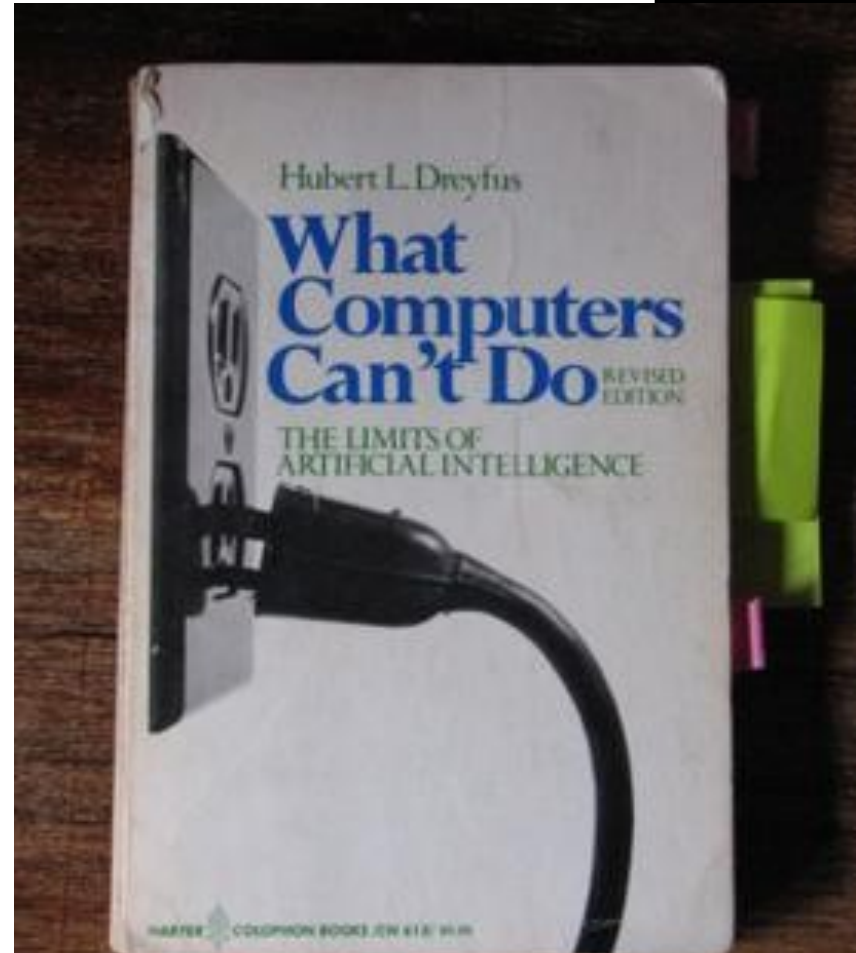
Classical Readings in the
Phenomenology of AI





Hubert Lederer Dreyfus (1929 -2017)

What Computers Can't Do
(1972; 1979; 1992)



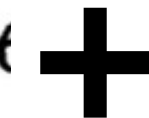
+

Part I. Ten Years of Research in Artificial Intelligence (1957–1967)

- 1. Phase I (1957–1962) Cognitive Simulation**
 - I. Analysis of Work in Language Translation, Problem Solving, and Pattern Recognition**
 - II. The Underlying Significance of Failure to Achieve Predicted Results**

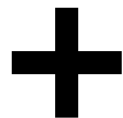
- 2. Phase II (1962–1967) Semantic Information Processing**
 - I. Analysis of Semantic Information Processing Programs**
 - II. Significance of Current Difficulties**

Conclusion



Gottfried Wilhelm Leibniz

(Leipzig 1646 – Hannover
1716)



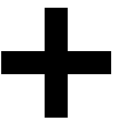
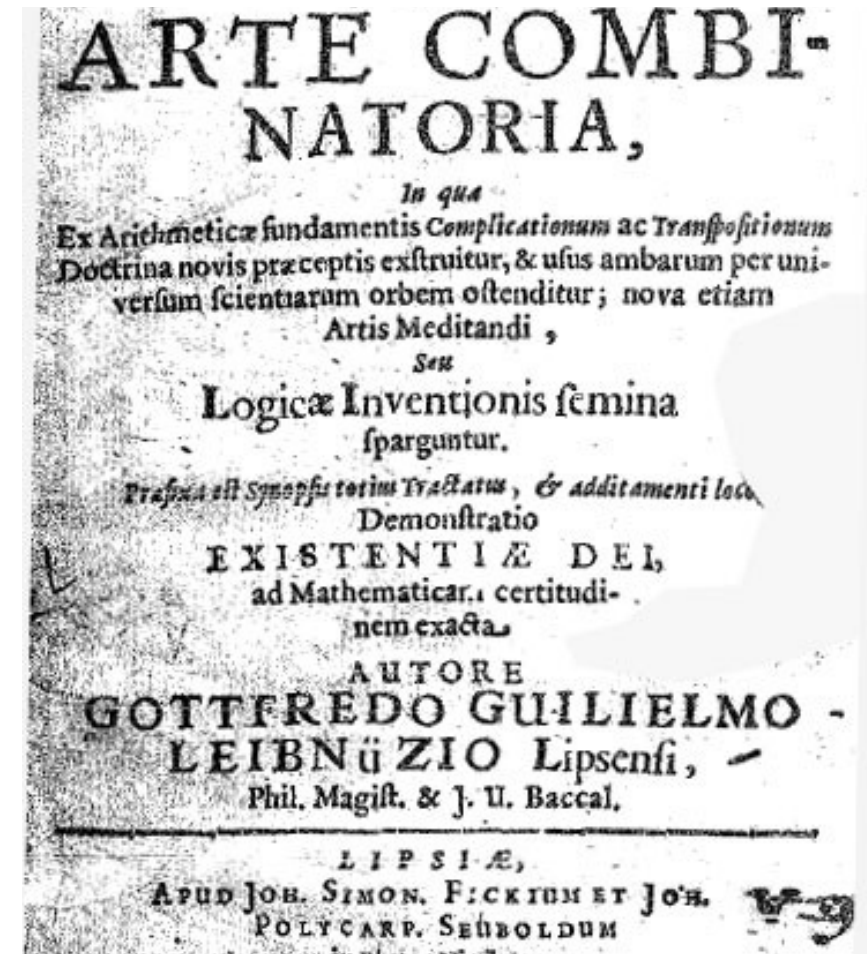
G. W. Leibniz

Dissertatio de Arte Combinatoria (1666)

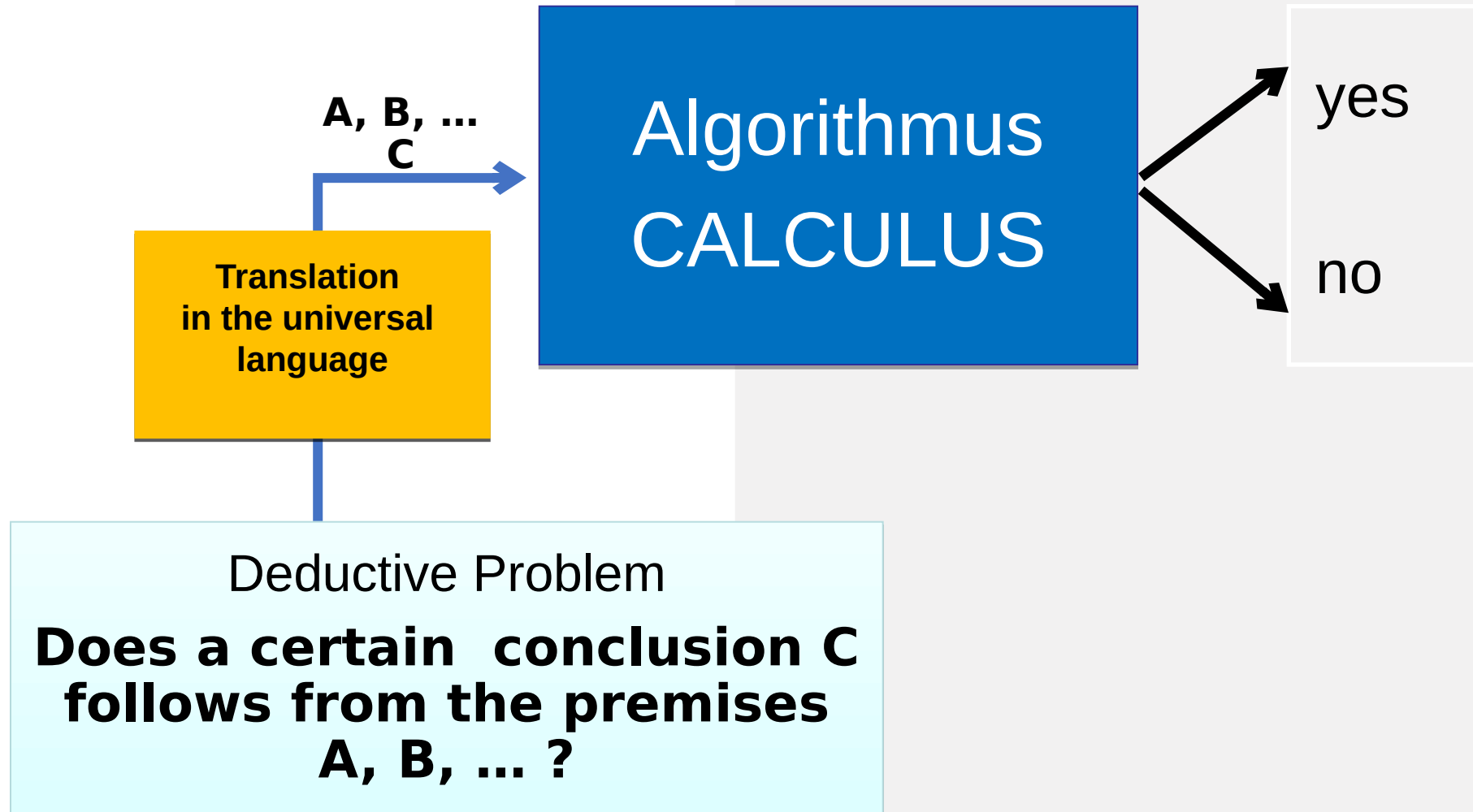
Dissertation on the Combinatorial Art

“When I was eighteen, I wrote a booklet with the Title *De Arte Combinatoria* [...] and discovered a **sure thread for thought** (*certum meditandi filum*), admirable tool of true analysis, **the corollary of which is [artificial] language or general characteristics.**”

(Letter to Tschirnhaus, 1679)



Leibniz's Dream



Partial Realizations of Leibniz'

9

Dream

- **George Boole** *The Mathematical Analysis of Logic* (1847)
An Investigation of The Laws of Thought (1854)
- **Gottlob Frege** *Begriffsschrift* (1879)
Über die Begriffsschrift des Herrn Peano und meine eigene (1897)
- **Giuseppe Peano** *I principii di geometria logicamente esposti* (1889)

⋮

1930-36: Birth of the computability theory

Reasons: *Are there problems that are not decidable?*

- Ex.: **Entscheidungsproblem**
Hilbert & Ackermann, *Grundzüge der theoretischen Logik* (1928)
→ Conjecture: “**Negative** solutions”
- **Gödel's Theorems** (1930-31) respectively, Generalisability of Gödel's Theorems



1930 – 1936: five extensional equivalent mathematical models for the informal concept of computability

- **λ -computability** [Church 1933]
- **General Computability / HG-computability** [Gödel 1934]
- **μ -computability** [Kleene < 1936 (1938)]
- **S-computability / Representability** [\approx Gödel 1936]
- **Turing-computability** [Turing 1936]



The image features a blurred background of a bookshelf filled with books. In the foreground, a stack of books is visible, with the top one open. Overlaid on this scene are various white, glowing icons and symbols, including letters (Z, X, y, W, V, C), numbers (11, 0), mathematical symbols (+, Σ), and other symbols like a lightbulb, a magnifying glass, a hand pointing, and a bar chart. The text "Language Translation" is written in a large, white, sans-serif font across the lower portion of the image.

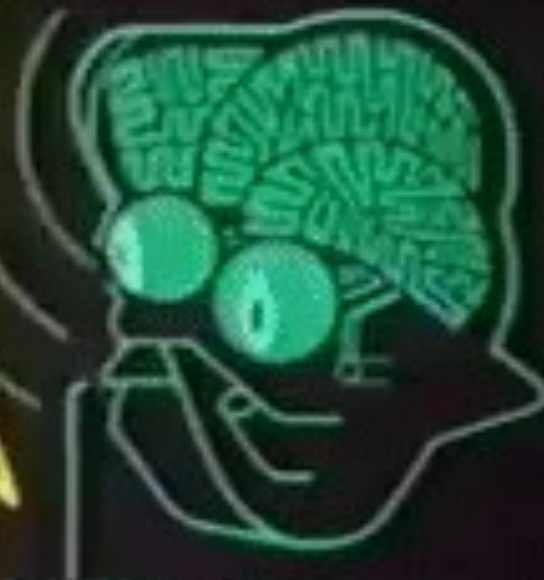
Language Translation

BABEL FISH

FREQUENCY SENS



SLURPGLURG



THE BABEL FISH IS SMALL, YELLOW, LEECH-LIKE,
AND PROBABLY THE ODDEST THING IN THE UNIVERSE.
IT FEEDS ON BRAIN WAVE ENERGY, ABSORBING ALL
UNCONSCIOUS FREQUENCIES AND THEN EXCRETING
TELEPATHICALLY A MATRIX FORMED FROM THE

makeagif.com



Y. Bar-Hillel.
*The Present
Status of
Automatic
Translation of
Languages*
(1960)

- “Let us notice that in June 1952, when the first Conference on Machine Translation convened at MIT, there was probably only one person in the world engaged more than half-time in work on machine translation, namely myself.”
- The first Conference on Machine Translation takes place in 1952. 1958 ca. i.e. only 6 years later, 250 people were working on machine translations one and one-half million dollars were spent upon research on machine translation.



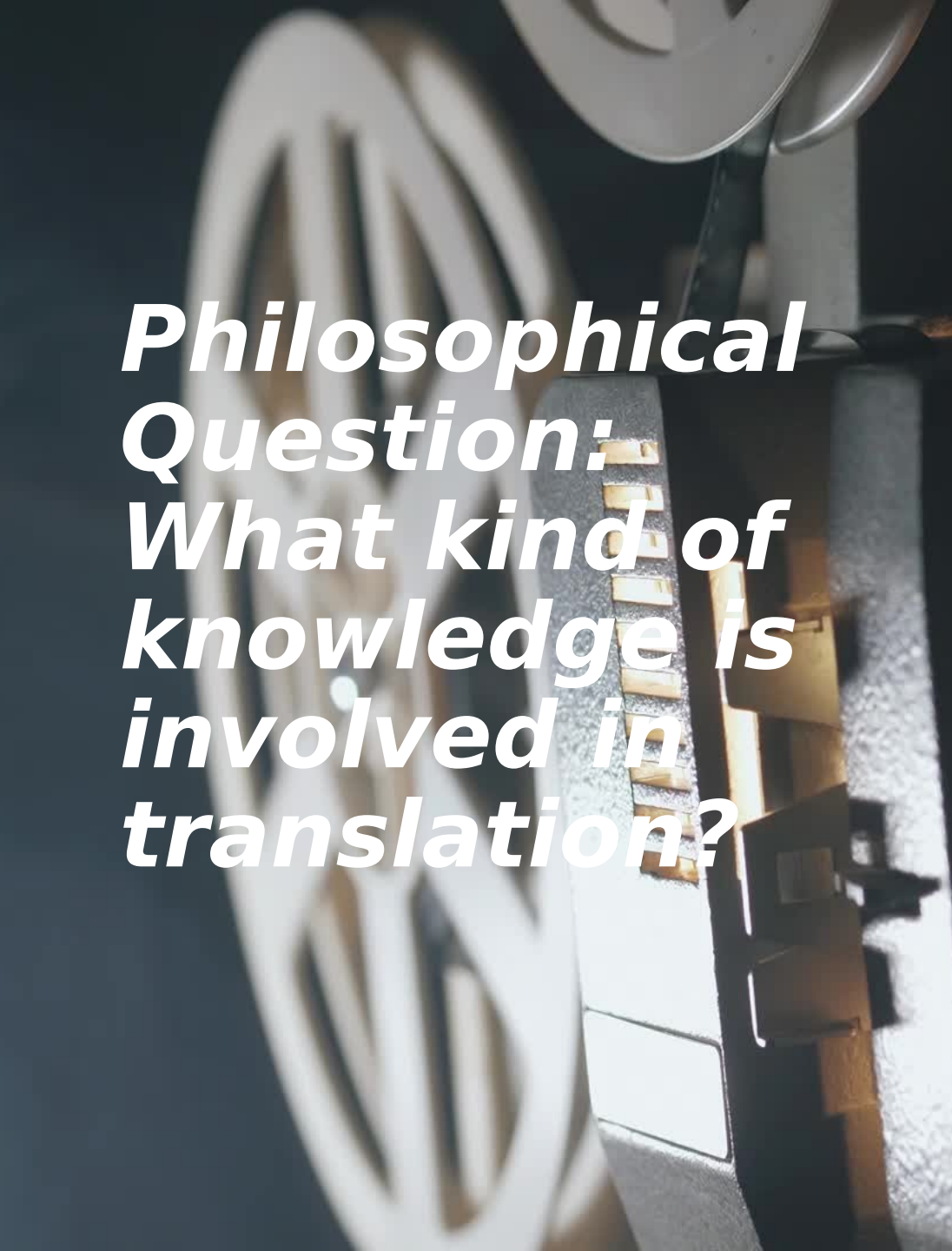


Y. Bar-Hillel.
*The Present
Status of
Automatic
Translation of
Languages*
(1960)

“During the first year of the research in machine translation, a considerable amount of progress was made ... It created among many of the workers actively engaged in this field the strong feeling that **a working system was just around the corner.** Though it is understandable that such an illusion should have been formed at the time, **it was an illusion. It was created ... by the fact that a large number of problems were rather readily solved** ... It was not sufficiently realized that the gap between such output ... and high quality translation proper was still enormous, and that the problems solved until then were indeed many but just the **simplest ones whereas the “few” remaining problems were the harder ones-** very hard indeed.”

Y. Bar-Hillel 1960.





*Philosophical
Question:
What kind of
knowledge is
involved in
translation?*

- **full automation of the translation is incompatible with high quality.**
- This is brought back by Dreyfus to a **certain particular characteristic that is peculiar to the skilled human translator** but which the machine cannot simulate, and which we might call “contextual awareness”. A good human translator translates well is able to understand nuances because he **understands the context.**

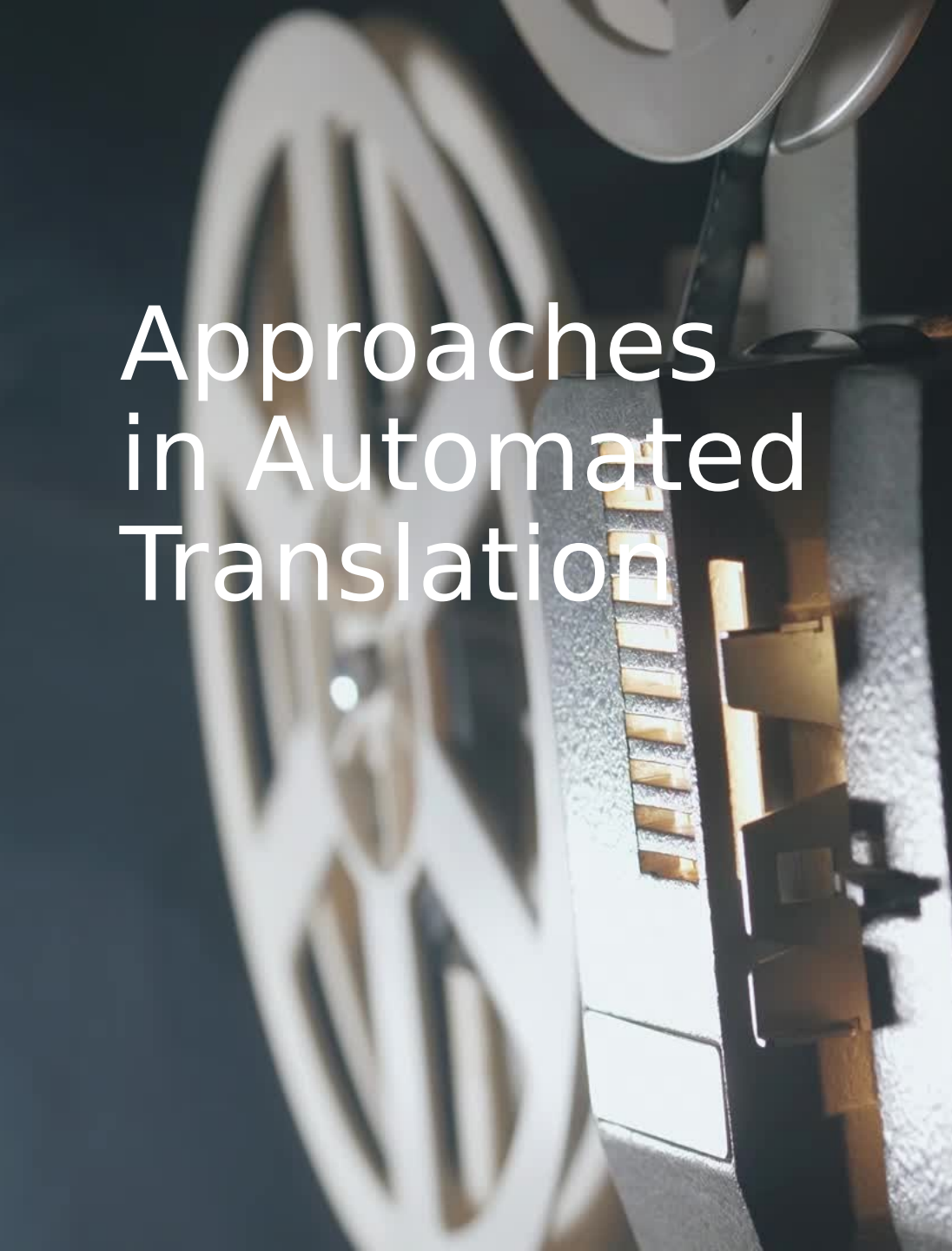




Logical Morphology... Leibniz, Husserl

- One of the first things that becomes evident when dealing with the problem of automatic translation is that computers are not aware of what we know as **logical morphology**
- Another typical human peculiarity that is difficult to teach a machine is the adaptation of words. When a translator looks at the meaning of a word, he can adapt it, for example, he knows how to obtain the plural, the corresponding verb, etc





Approaches in Automated Translation

- Should the translation be based on words, word chunks, or sentences?
- **Direct translation systems**
- **Transfer systems** □ they integrate some kind of syntactic analysis
- **Systems based on an interlingua** □ a more or less formal representation of the content to be translated
- ... Word-for-word translation is not a good praxis since even if we list in a dictionary all possible meaning of a given word, **we should have to teach the computer how to choose the right meaning.**





Statistical Approaches

- Systems for automated translations perform a sentence-by-sentence translation, actually they translate fragments of sentences





Statistical Approaches

- Statistical approaches use large quantities of text available on the web
- One idea that is at the basis of translation systems is that simultaneous translators often translates **semiautonomous groups of words** without having heard the full sentence.
- Statistical systems do not perform a deep analysis **but identify groups of words that work together.**
- Statistical approaches take meaning to corresponds **to the ways words are used!!!**
- Statistical approaches discover regularities in the way the words are used

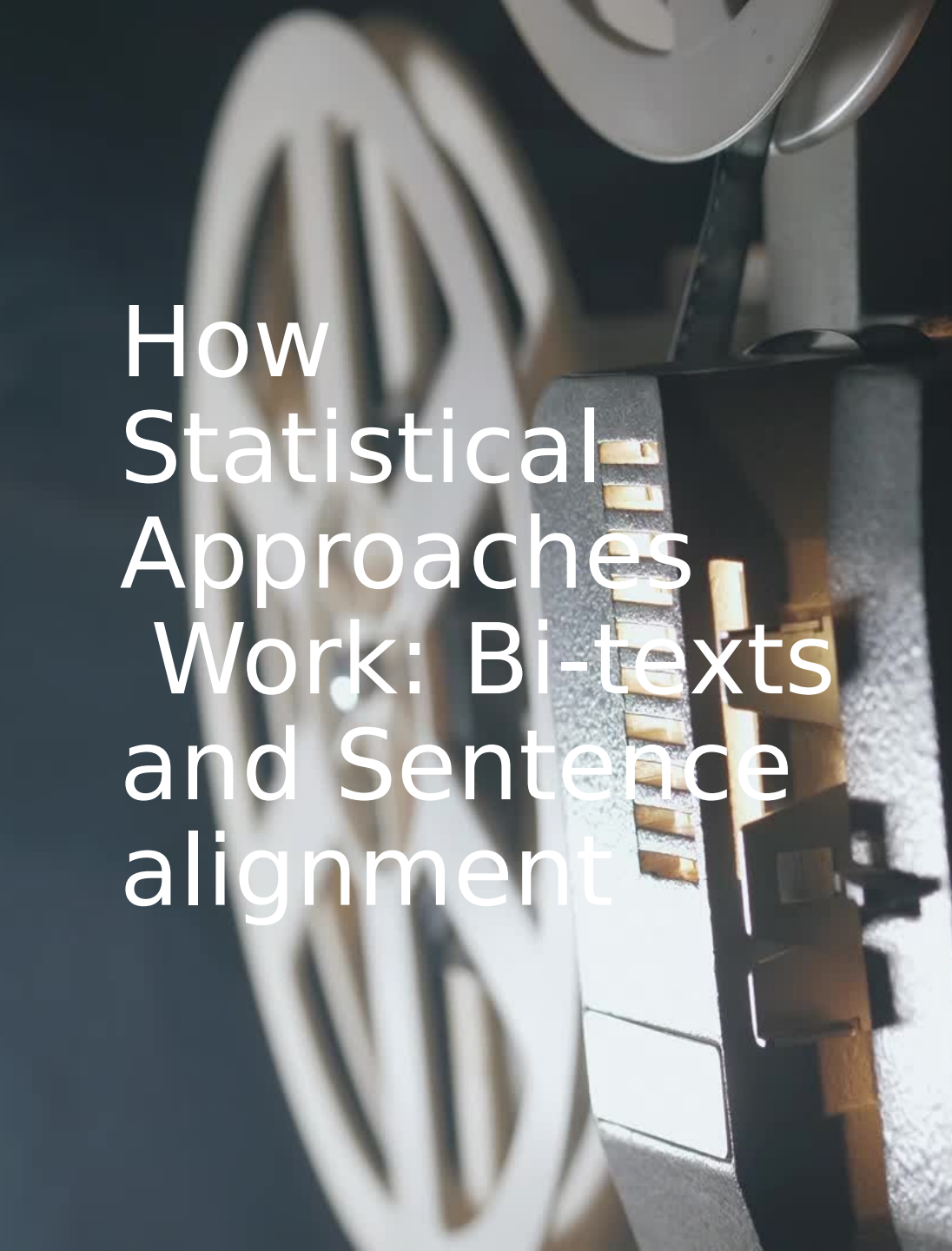


A close-up photograph of a film reel and its mechanical components. The reel is on the left, and the film strip is visible in the center. The background is dark, and the lighting highlights the metallic and plastic parts of the machine.

Wittgenstein

- Wittgenstein *tells an alternative story in which the words of our language are something we operate with and not as something standing for certain objects.*
- At the end of § 1 of *Philosophical Investigations* he considers the number-word “five” and writes: “But what is the meaning of the word »five«? – No such thing was in question here, only how the word »five« is used.”
- What Wittgenstein is conveying here is the thought that what we call “reference” **is secondary in comparison with use.**





How Statistical Approaches Work: Bi-texts and Sentence alignment

French Text	English Text
J'ai fait cette comparaison et je tiens à m'arrêter sur ce point.	I have looked at this and I want to talk about it for a second.
L'article 11 du projet de loi crée tellement d'exceptions qu'il va bien au-delà de l'article 21 de la convention, au point de carrément compromettre l'objet même de celle-ci.	Clause 11 in the bill creates so many exceptions that it goes well beyond article 21 of the treaty and basically completely undercuts the intention of the convention itself.
Je cite l'article 21 de la convention.	I will read what article 21 says.
C'est assez simple:	It is pretty straightforward:
Chaque État partie encourage les États non parties à la présente Convention à la ratifier, l'accepter, l'approuver ou y adhérer [...]	Each State Party shall encourage States not party to this Convention to ratify, accept, approve or accede to this Convention [...]
Chaque État notifie aux gouvernements de tous les États non parties à la présente Convention.	Each State Party shall notify the governments of all States not party to this Convention.

Each cell is a sentence.
Each number refers to the number of words in the sentence.

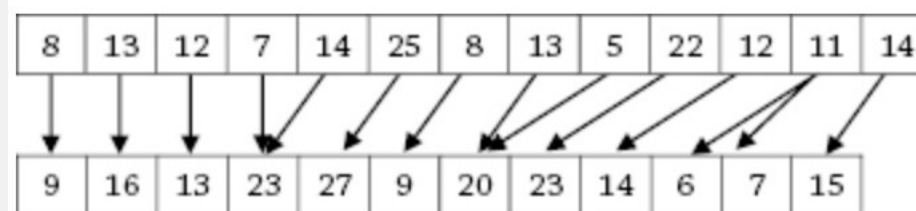
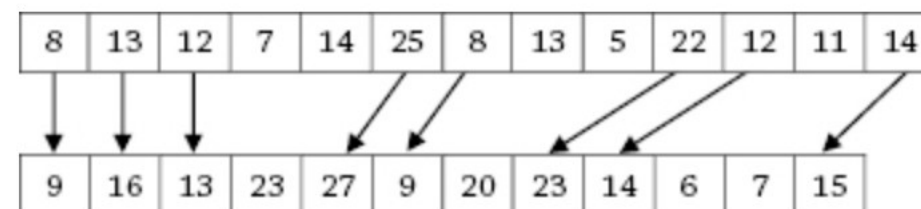
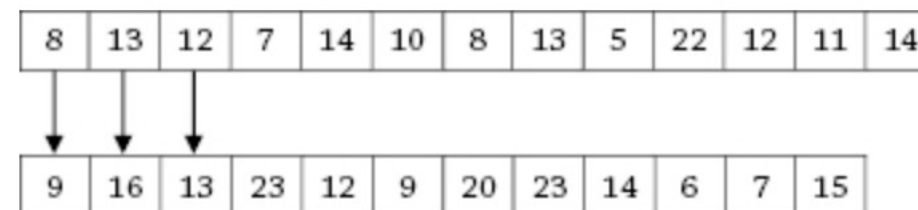
The system sees that the first three sentences have a relatively similar number of words and can therefore be linked together.

So, it begins the alignment based on sentence length. The same applies for some specific patterns in the text

Finally, the system tries to “bridge the gaps” by establishing links between the source and the target text so as to obtain a fully connected bi-text in the end.

8	13	12	7	14	10	8	13	5	22	12	11	14
---	----	----	---	----	----	---	----	---	----	----	----	----

9	16	13	23	12	9	20	23	14	6	7	15
---	----	----	----	----	---	----	----	----	---	---	----





Problem Solving

The General Problem Solver (1957) and

the Logic Theorist (1955)

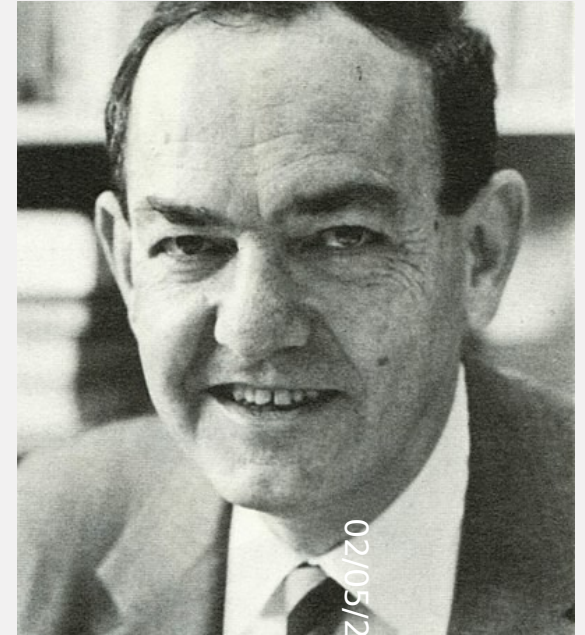
General Problem Solver (GPS)

Any formalized
solved, in princi
theorem



Herbert Alexander Simon

(June 15, 1916 - February 9, 2001)



02/05/2024

Allen Newell

(March 19, 1927 - July 19, 1992)



44

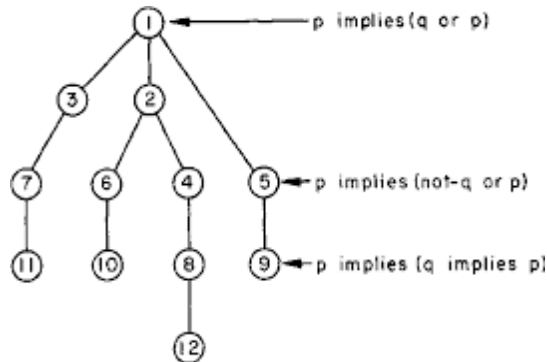
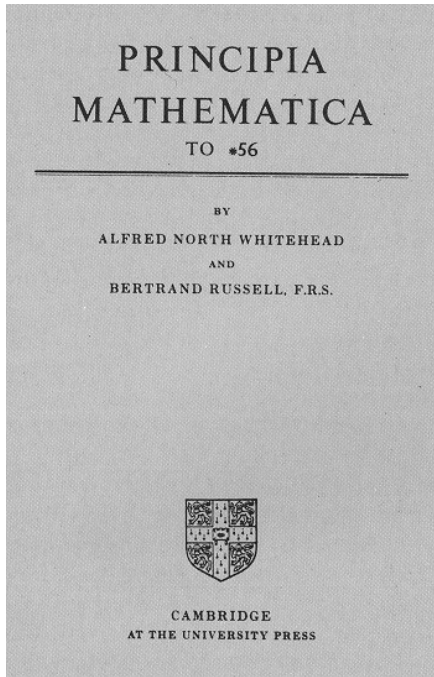


Fig. 5.—Proof tree of proof of 2.02 for Whitehead's system.

The Logic Theorist



- The ***Logic Theorist***, was the first program conceived to perform automated reasoning. It was able to prove 38 of the first 52 theorems in chapter two of Whitehead and Russell's ***Principia Mathematica*** and found new and shorter proofs for some of them.
- The Logic Theorist is a program that performs logical operations on logical expressions.



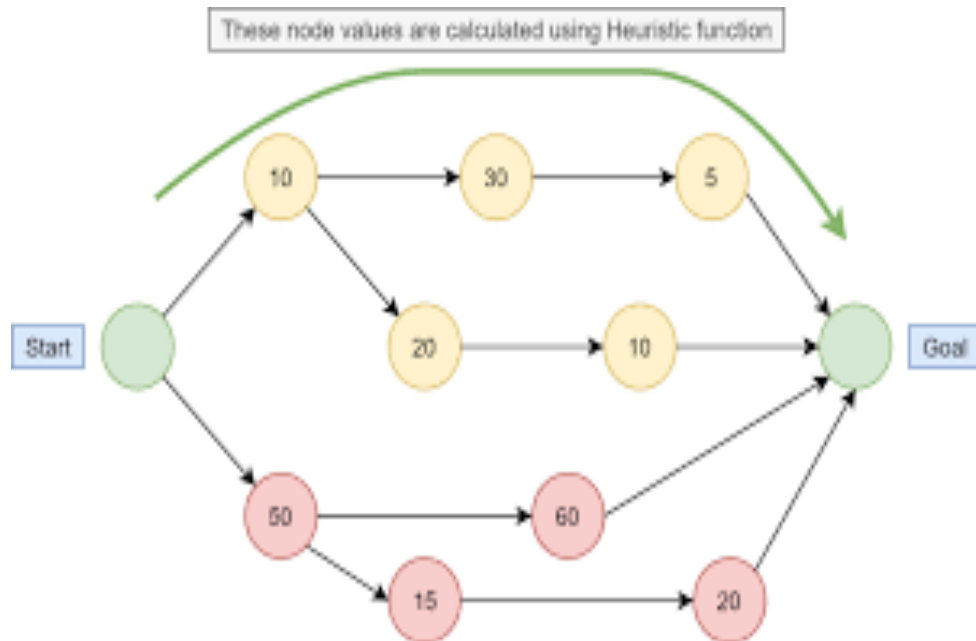
The Logic Theorist

- When the Logic Theorist applies a transformation rule to a statement it generates one or more new statements.
- The information content of the resulting statement is compared to the original statement
- If the resulting statements has lower information content this indicates progress toward the proof.
- To calculate the information content the logic theorist uses a heuristic measure that assigns values to different types of transformations.
- For example, a transformation rule that simplifies the statements significantly would be assigned higher values than a rule that makes a minor modification.
- The information content measure serves as a guide for the logic theorist search algorithm.



- 
- When the Logic Theorist was being built, the field of AI did not exist and was just coined by John McCarthy during the summer of 1956 for a conference organized at Dartmouth College in collaboration with Marvin Minsky.
 - At the conference, Simon and Newell presented their invention, which was accepted. Later on, Simon confides: “They didn’t want to hear from us, and we sure didn’t want to hear from them: we had something to show them! In a way, it was ironic because we already had done the first example of what they were after; and second, they didn’t pay much attention to it”.
 - So, while McCarthy and his team were brainstorming the possibility of Artificial Intelligence, Simon, Newell, and Shaw were already recording breakthroughs in the same field
- 

Heuristic Search Techniques



- What is Heuristics?

- Heuristics is a problem-solving method that

i. Uses shortcuts /calculated guesses to provide good enough solutions.

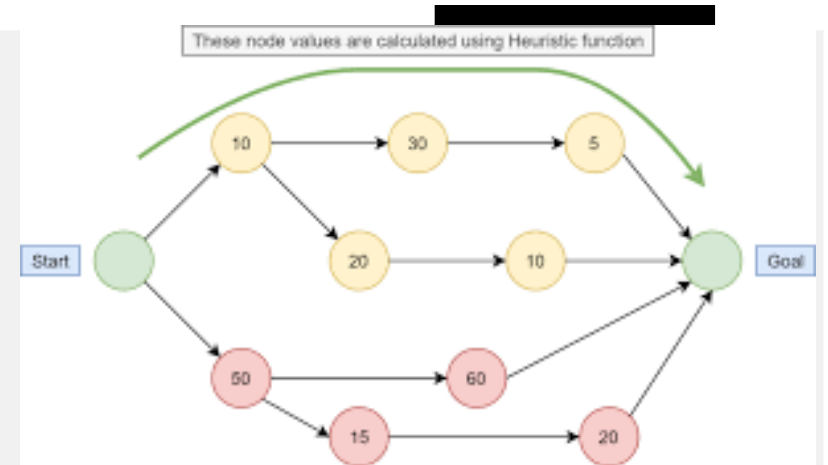
ii.Reduces time complexity to reach solution

iii. May not give best /optimal solution



Heuristic Search Techniques

What is Heuristics?



- This a state-space representation
- the numbers stand for the “costs” in terms of the time that it takes to go from one state to another state.
- We want to go from the start to the goal.
- The idea is to verify all paths and calculate the sum of the seconds employed in reaching the goal for each path. The algorithm should choose the shortest path in terms of time.
- The shortest path is the optimal solution.



The image features a stack of books on a wooden surface, with a blurred bookshelf in the background. Overlaid on the scene is a collection of glowing white icons representing various mathematical and scientific concepts, including numbers (1, 11, 2, 0, x), letters (y, z), symbols (+, -, Σ, ∫, ∞), and geometric shapes. The text 'Pattern Recognition' is written in a large, white, sans-serif font across the middle of the image, partially overlapping the books.

Pattern Recognition

What is Pattern Recognition?

“That something besides the sensations actually is immediately given follows (independently of mathematics) from the fact that even our idea referring to physical objects contain constituents qualitatively different from sensations or mere combinations of sensations, e.g., the idea of object itself ... Evidently, the “given” underlying mathematics is closely related to the abstract elements contained in our empirical ideas. It by no means follows, however, that the data of this second kind, because they cannot be associated with actions of certain things upon our sense organs, are something purely subjective, as Kant asserted. Rather they, too, may represent an aspect of objective reality, but, as opposed to the sensations, their presence may be due to another kind of relationship between ourselves and reality. “


Kurt Gödel, Supplement to “What is Cantor’s continuum problem?” when it was reprinted in Benacerraf and Putnam, *Philosophy of Mathematics: Selected Readings* in 1964,





- Suppose we stand in front of a triangularly shaped tree, we may be interested in its triangularity, rather than its front side, its branches or leaves, etc. It does not matter that the tree is replaced by another thing of the same shape; as long as the thing is triangular, the object of our interest remains the same.
- A crucial difference between a shape and a physical object is that in the case of a shape, that shape could be had by another object than the one that happens to have it now, so the shape is a general notion, and if we are concerned with a shape, it does not matter whether we replace the particular physical object that instantiates it with another object that instantiates the same shape. The same holds of arithmetic features, couples, triples, etc., we can focus on these general features rather than focusing on the individual physical objects.



- 
- If we can take the physical object that we are facing away and put it in another one, and still say that we are experiencing the same object now as we did before, for example, circularity or triangularity, we are not experiencing a physical object, but what Husserl calls an *eidos* or essence (*Wesen*).
 - Mathematical entities are typical of the kind of entities we are dealing with when we are focusing on essence. For Husserl, mathematics is a typical example of an eidetic science.



Gestalt Psychology

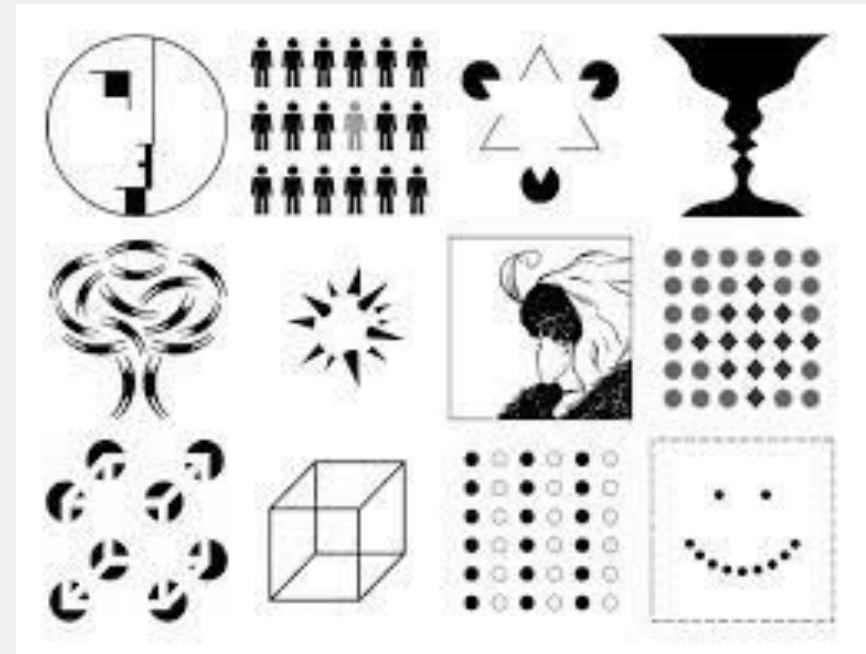
One speaks, for example, of a file of soldiers, of a heap of apples, of a row of trees, of a covey of hens, of a flight of birds, of a gaggle of geese, and so on.

These are *quasi- qualitative* moments that allows apprehension of large groups of things of the same kind

In each of these examples we speak of a sensible group of objects like each other, which are also named in terms of their kind. But not this alone is expressed. Rather there is expressed **a certain characteristic property of the unitary total intuition of the group, which can be grasped at once** and which in its well-distinguished forms constitutes the most essential part of the signification of those expressions introducing the plural: “file,” “heap,” “row,” “covey,” “flight,” “gaggle,” etc.



Gestalt is both a school of psychology and a theory perception that emphasises the processing of entire patterns and configurations, instead of individual components. It emerged in the early twentieth century in Austria and Germany



Pattern Recognition

By DENIS RUTOVITZ

Medical Research Council

[Read before the ROYAL STATISTICAL SOCIETY on Wednesday, May 18th, 1966,
the President, Mr L. H. C. TIPPETT, in the Chair]

1. INTRODUCTION

DURING the past 10 years about 200 articles and several books have appeared, dealing with machine recognition of optical and other patterns (mainly alphabetic characters and numerals). About half of these have described methods not linked to a specific machine; usually, a general-purpose digital computer was programmed to carry out or simulate the operation of the system. But perhaps equally many of the papers have included accounts of machines built to implement the proposed method of recognition, and no doubt many other machines have been constructed which have not featured in the literature. These devices range from fairly inexpensive realizations of small parts of a complex system to large computers, with special attachments for pattern processing, in the million-pound class. I would like to indicate some of the many reasons for this volume of interest and to review the main trends and preoccupations in this rapidly expanding area.

The subject may be said to have begun in about 1955 when banks began to be seriously interested in automatic cheque sorting. The character recognition problem

Hand-written or unstructured data



Organised and searchable data



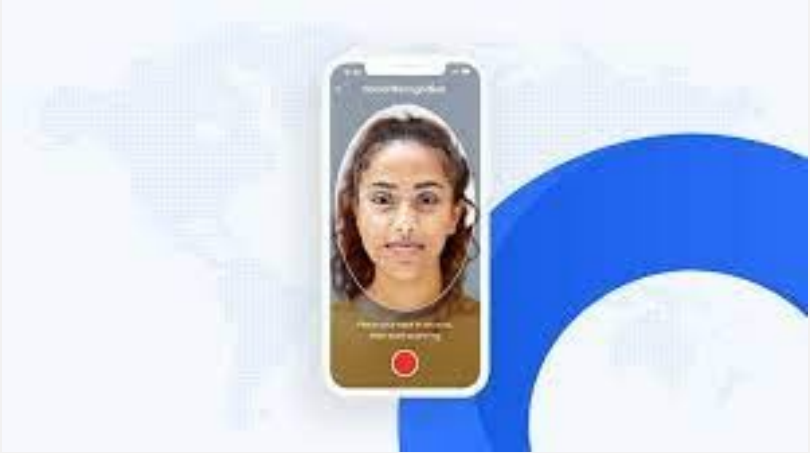
Optical Character Recognition

Optical character recognition designed to bank cheque sorting (although with very low resolution)

Aerial photo recognition for military application. What is the common feature of all these old applications?



What about today's applications?



Big Data + Deep Learning + GPUs

Today ...

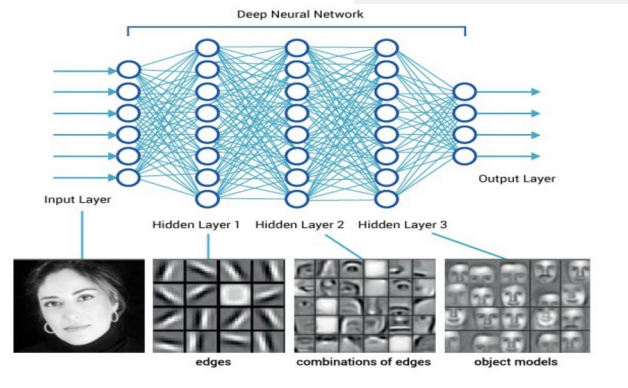
After 50 years of research, the main stream is **Big Data + Deep Learning + GPUs**

BIG DATA

Facebook 350
millions of images
per day

YouTube 300
hours of videos
per minute

DEEP
LEARNING



GPU



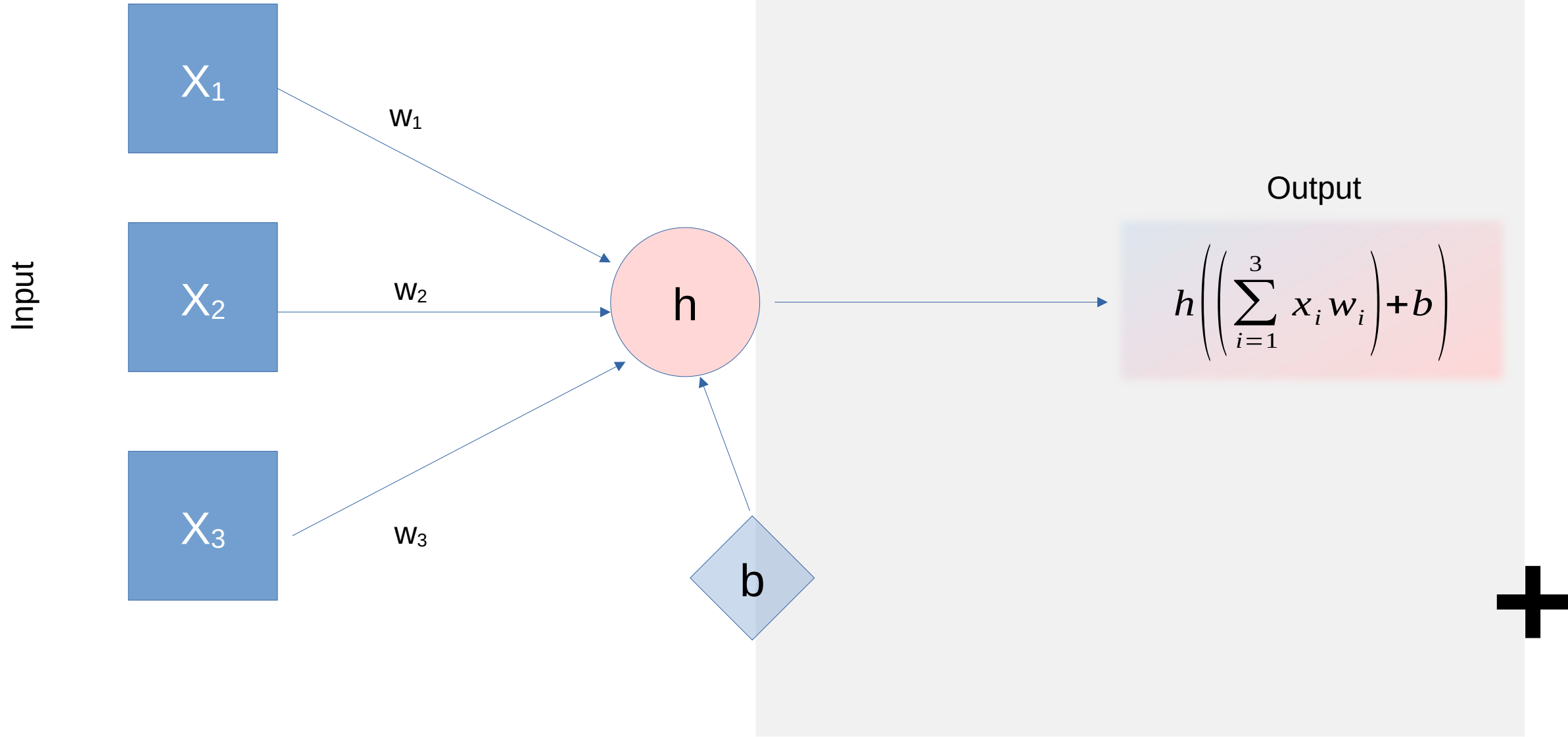
(Courtesy by F. Roli)



- In his popular text *How to solve it?* Pólya distinguishes different phases in the heuristics process
- (i) understanding the problem as a whole
- (ii) establishing a program that leads to the solution
- According to Dreyfus, computer scientists have only focused on the second point while neglecting the first. But it is precisely the first point that is important and must be introduced into the program. And this is precisely the difficult thing because it is possible that the human mind in understanding a problem as a whole does not follow rules.
- Problem Solving, language translation, pattern recognition all grounds in this essentially human way of processing information.



Neural Networks



Large Language Models

General Artificial Intelligence



Narrow Artificial Intelligence



LLMs

Main Role:

Machine language manipulation and generation of sentences in different (human) languages

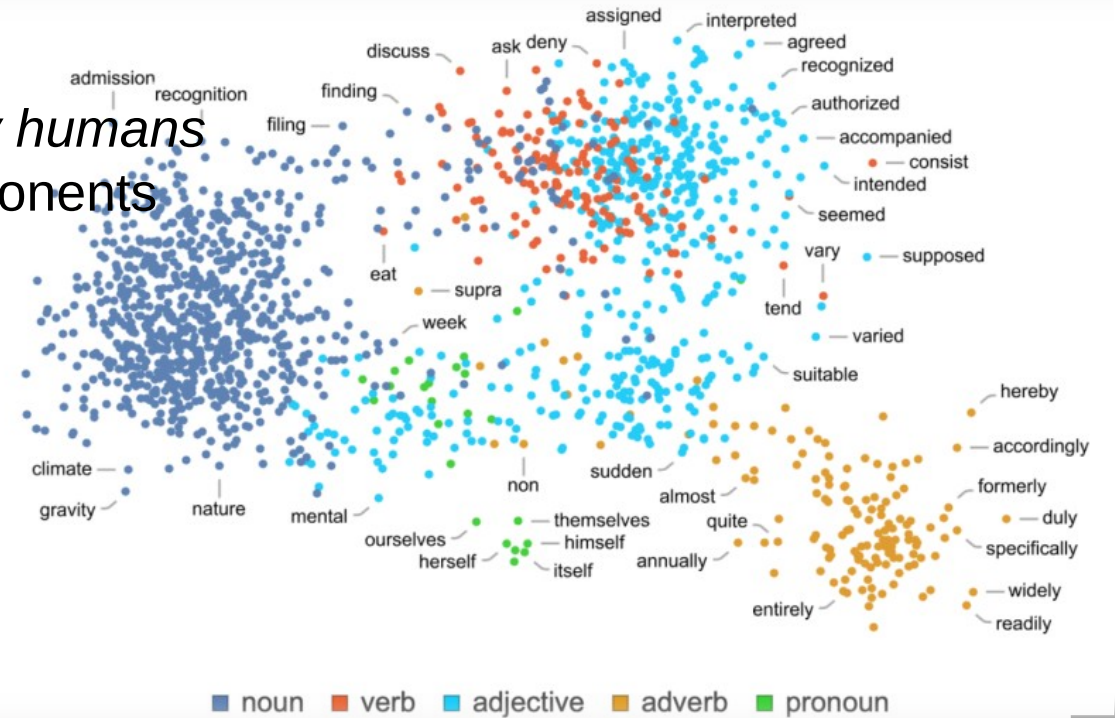
BERT-style: complete missing words from a sentence

GPT-style: Add next word



Ingredients for LLMs

- ◆ Huge text corpora (mainly from the web) *produced by humans*
- ◆ Vectorial encoding of semantic proximity of text components
- ◆ Neural/sub-symbolic computational architecture
- ◆ Matrix operations on (encoding of) linguistic data
- ◆ GPUs used for parallel computing



Recipe for LLMs

These are statistical encodings of *other words* that are “called by” or “calling for” the given word

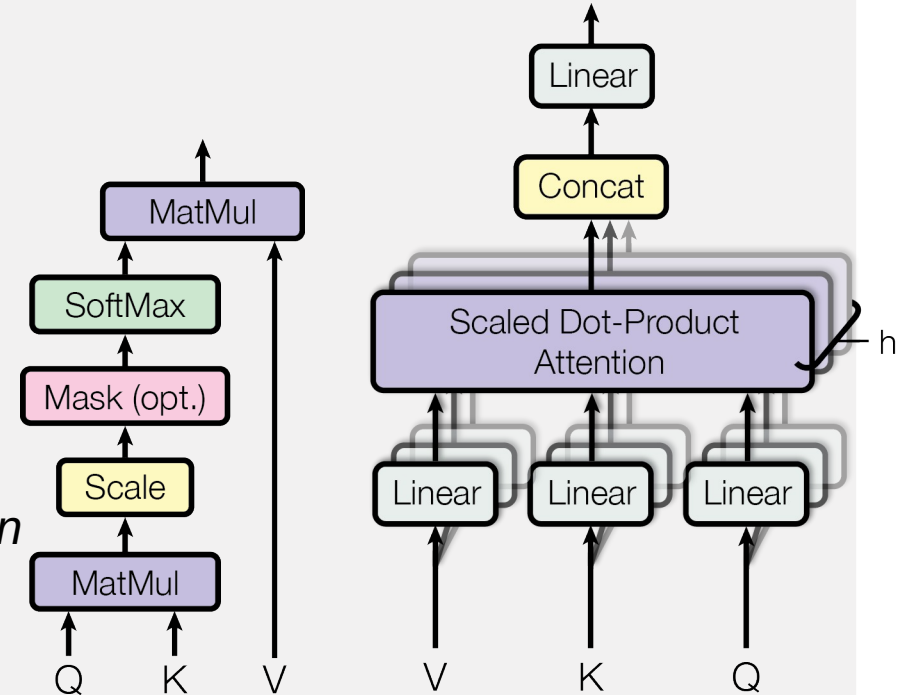
1. Consider a string of linguistic items in a text
2. Assign to each of them 3 matrices: *Queries*; *Keys*; *Values*
3. Pair Q and K matrices to compute probability weights
4. Generate average values using these probability weights
- ✓ Complete BERT- or GPT-style tasks¹, including *language translation*

Q and K matrices are a *statistical proxy* for syntactic relationship

Syntax is re-constructed from its semantic mirror

Huge parallel computations (in GPUs) for searching syntactic structures within semantic encoding

Attention architecture



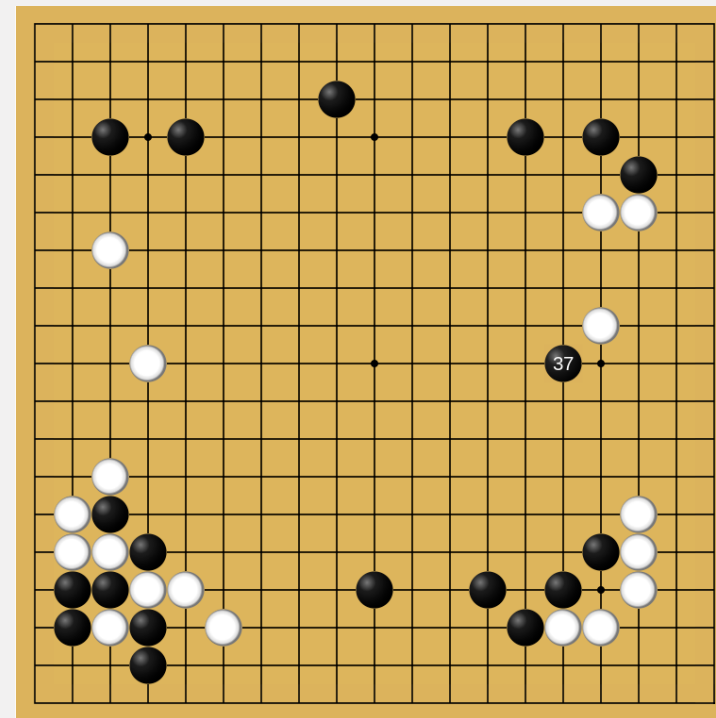
¹ Modulo parameter refinements/training

Problem solving, before LLMs

“

I thought AlphaGo was based on probability calculation and that it was merely a machine. But when I saw this move, I changed my mind. Surely, AlphaGo is creative.

LEE SEDOL
WINNER OF 18 WORLD GO TITLES



Go, with its complex rules and vast possibilities, has long been considered one of the ultimate tests of human problem solving...

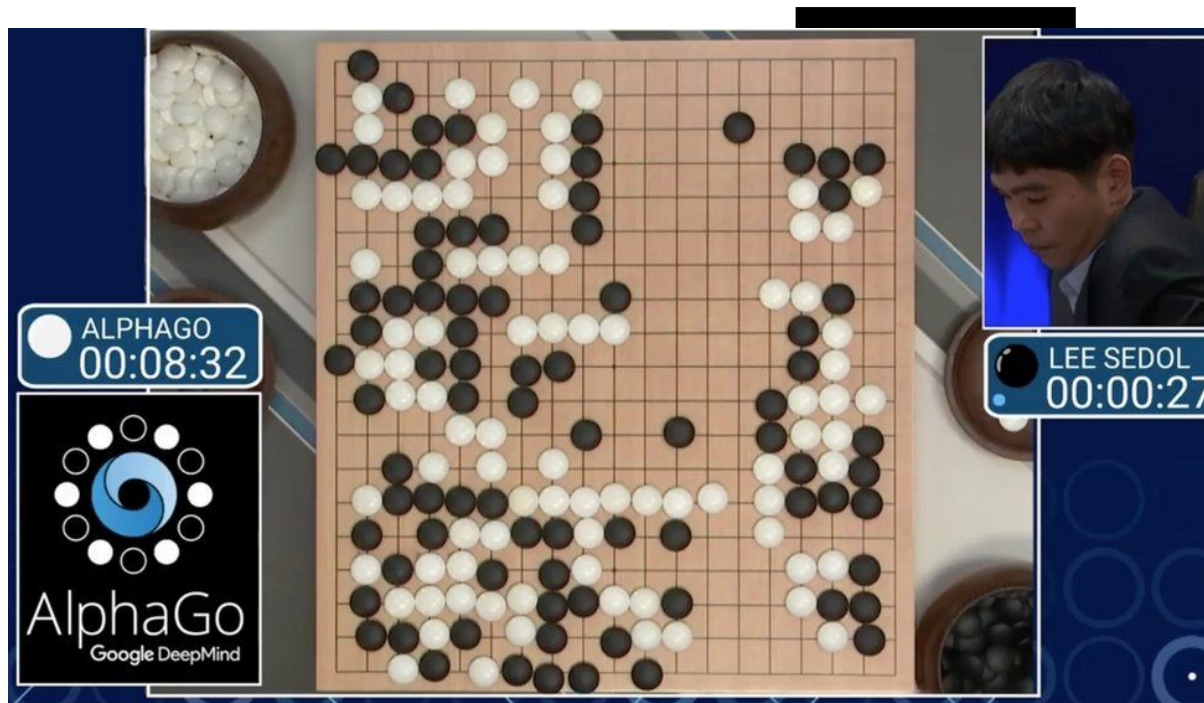


Problem solving, before LLMs

Match Overview:

- The match consisted of a series of five games played in Seoul, South Korea.
- AlphaGo won the first three games, showcasing its *unprecedented* strength and *strategic capabilities*.
- Lee Sedol managed to win the fourth game, demonstrating human resilience and adaptability.
- AlphaGo ultimately won the series **4-1**, marking a historic milestone in AI development.

The match sparked widespread discussion about the implications of AI advancements on various aspects of society, including employment and ethics. It has prompted reflections on the nature of intelligence, creativity, and intuition

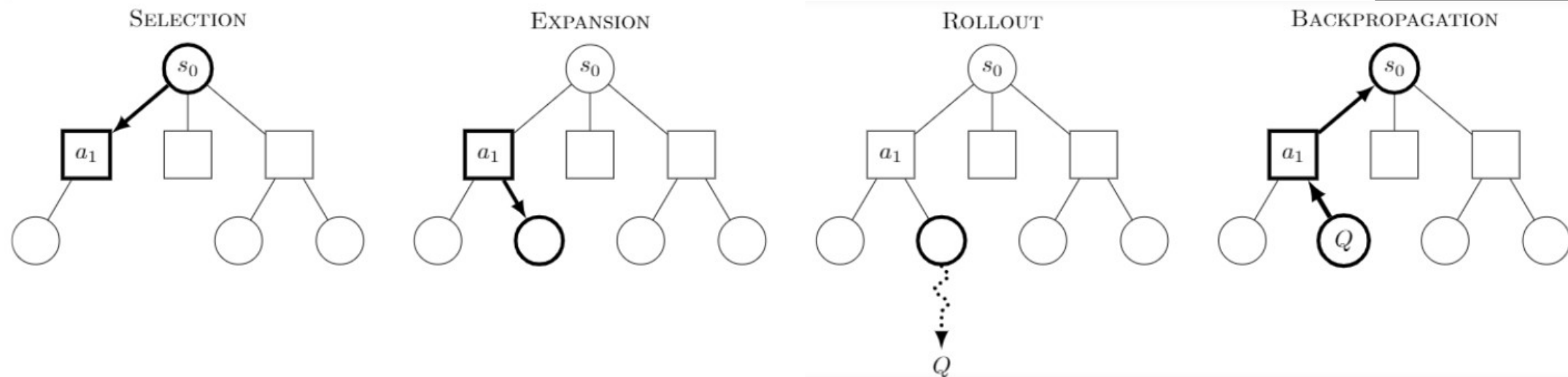


... *BUT*, in March 2016, Lee Sedol, a legendary Go player with multiple world championships, faced off against AlphaGo, an AI program developed by DeepMind.



Monte Carlo Tree Search

An alternative attention mechanism in AlphaGo

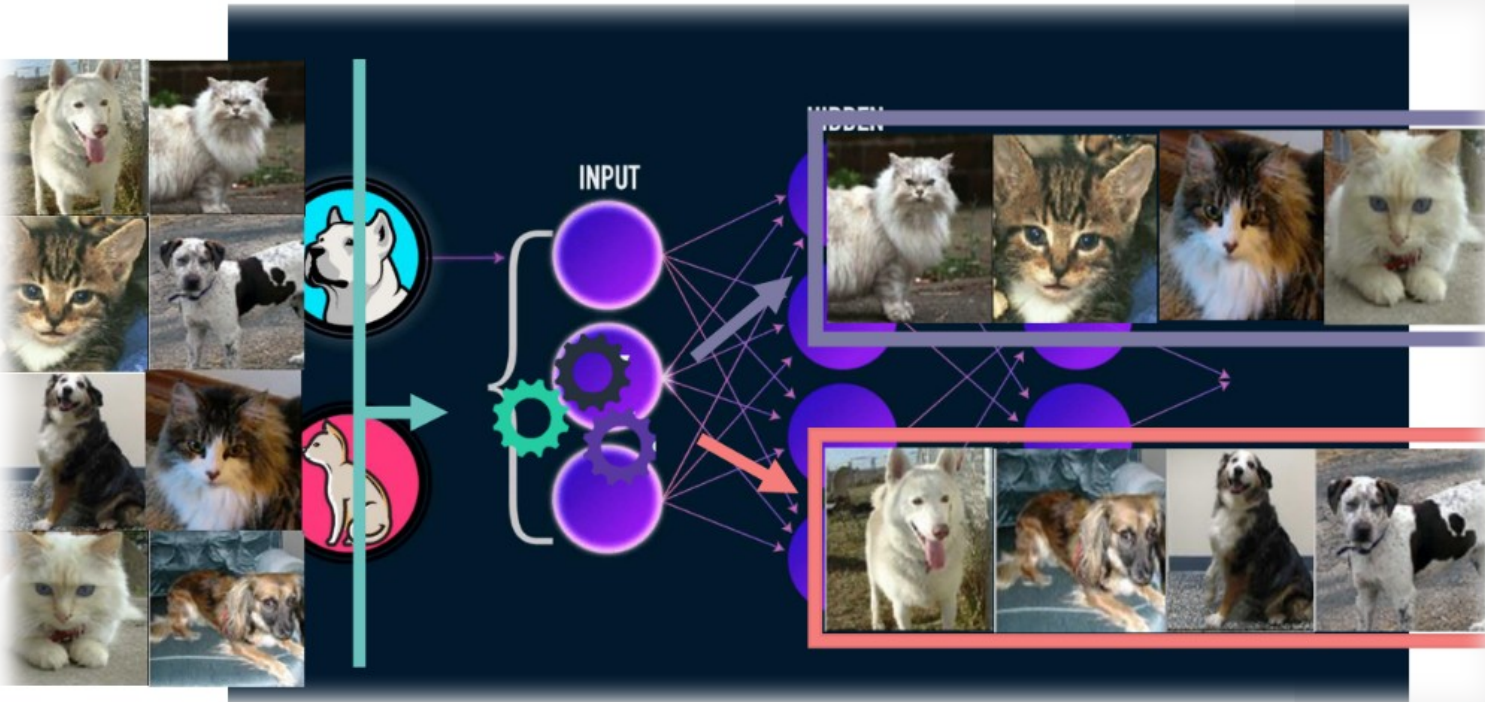


- 1. Selection:** Starting from the root node, traverse the tree using a *selection policy* until reaching a leaf node.
- 2. Expansion:** Expand the selected leaf node by adding one or more child nodes representing possible actions or states.
- 3. Simulation (Rollout):** Perform a simulated play-out from the newly added node until reaching a terminal state or a predefined depth, using a *default policy* (random or heuristic-based).
- 4. Backpropagation:** Update the statistics of visited nodes along the path from the expanded node to the root based on the outcome of the simulation.



Pattern recognition

Using algorithms and current AI techniques to detect regularities or similarities within datasets, images, signals, or any other form of structured information¹



- Identifying and interpreting patterns in data is crucial for various tasks such as image recognition, speech recognition, natural language processing, and predictive modeling.
- Recognising and categorising regularities within datasets enables AI technologies to “make sense” of the world.

¹ Modern version incepted by Kunihiko Fukushima’s “Neocognitron paper” in 1979





Pattern recognition

Types of Algorithms

Facial features like eyes, ears, mouth, and nose form a feature vector.

Facial recognition software uses this vector to compare and identify new data by matching it against stored feature vectors.

- **Supervised Algorithms (Classification)** → Two-stage methodology:
 - Stage 1:** Development and construction of the model;
 - Stage 2:** Predicting new or unseen objects
- **Unsupervised Algorithms** → "Group by" approach
 - ≈ Identifies data patterns and groups them based on similarity for predictions



Convolutional Neural Networks

The diagram illustrates a convolution operation. On the left is an 8x8 input grid of pixel values. A 3x3 sub-region of this grid is highlighted with a solid border, representing the kernel's current position. The values in this sub-region are 60, 60, 68 in the first row; 44, 60, 60 in the second row; and 68, 76, 76 in the third row. The value 60 in the second row, second column of this sub-region is also enclosed in a dashed box. To the right of the input grid is a 3x3 kernel matrix with values: $\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$. This is followed by an equals sign, then a 3x3 output grid: $\begin{bmatrix} -60 & -120 & -68 \\ 0 & 0 & 0 \\ 68 & 152 & 76 \end{bmatrix}$. The value 48 is shown in a dashed box to the right of the output grid, representing the sum of the products of the kernel and the input sub-region.

- ✓ Slide a small square, called a kernel, over the image.
- ✓ Movement occurs from top to bottom and left to right.
- ✓ At each position, pixel values within the kernel are multiplied by corresponding kernel values.
- ✓ The products are summed to produce a single output pixel value for that position.

Convolutional Neural Networks

Input/Model

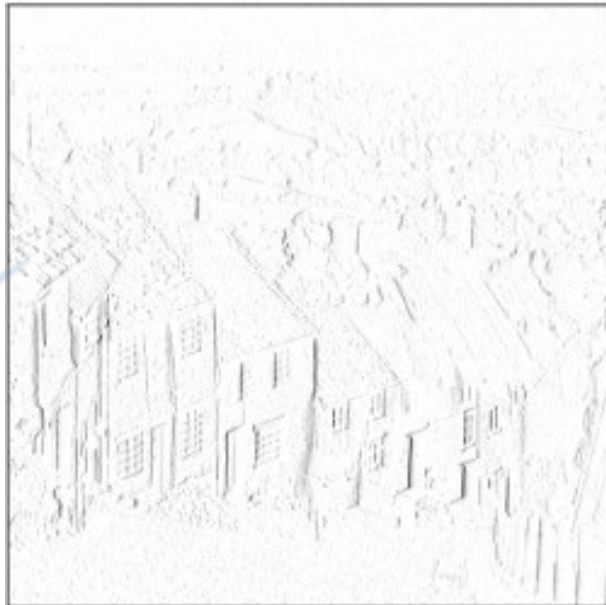


Blurring kernel

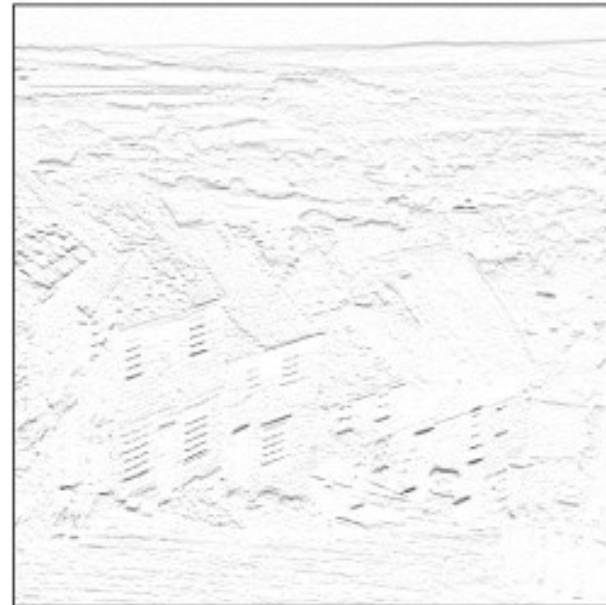


Gold Hill in Shaftesbury,
England

Vertical kernel



Horizontal kernel



Convolutional Neural Networks

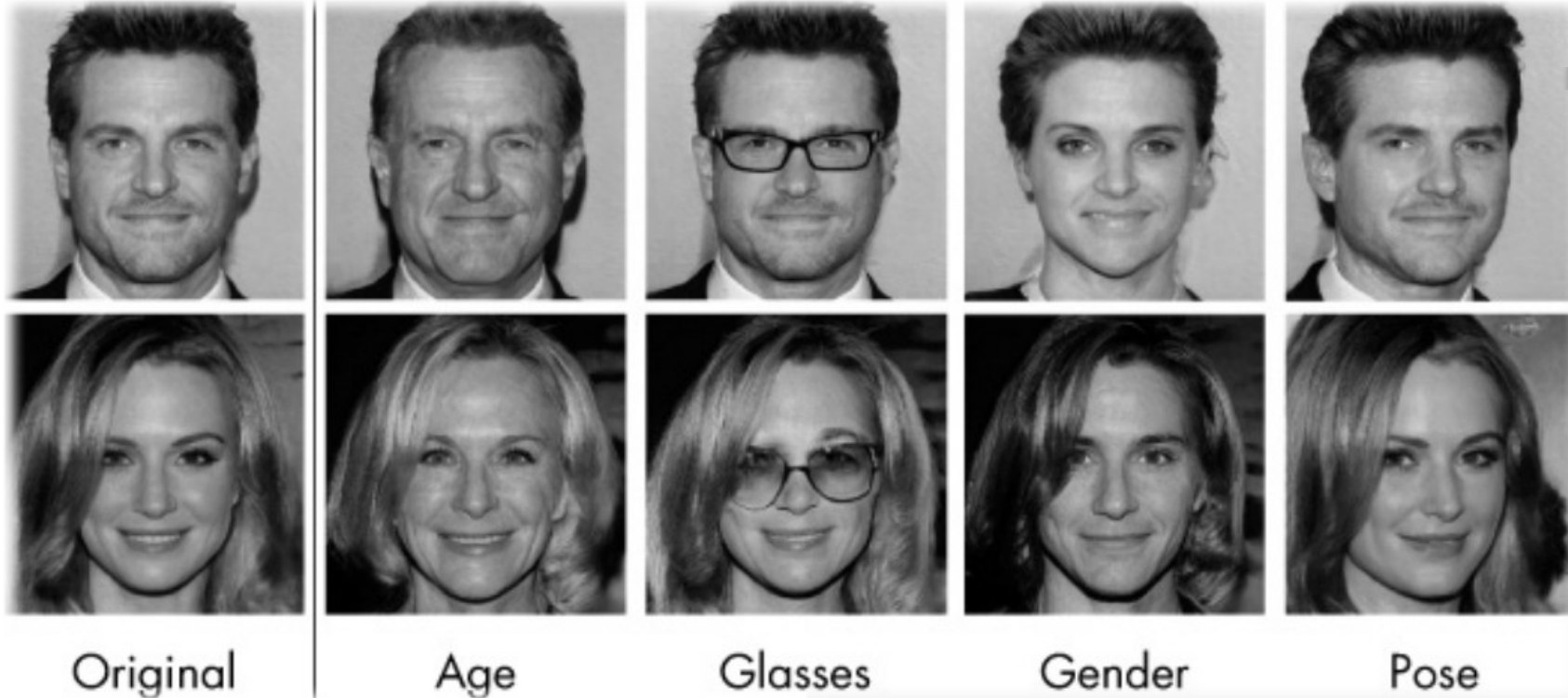
- ◆ Convoluting an image with different kernels highlights different aspects of the image
- ◆ During end-to-end training, CNNs apply a set of kernels in order to extract structural information that is relevant for classifying the input image
- ◆ The mechanism is inspired by the way our brain detects edges, textures, and colours

N.B.

CNNs apply different kinds of layers combined in various ways, to classify and/or localise inputs. Their design can then be quite complex. TensorFlow, PyTorch and others are *FLOSS* toolkits for implementing very efficient CNNs, handled as computational graphs.



Beyond Classical Pattern Recognition



Generative Adversarial Networks have two parts: a **generator** and a **discriminator**.

The generator creates *fake data*, and the discriminator tries to tell real from fake.

They compete in training, improving the generator's ability to create realistic data, like images or text.

In GANs, movement along specific directions in the noise/random/input vector space can predictably alter features of the generated output, facilitating control over desired attributes.



Many thanks for listening!

`cosimo.perinibrogi@imtlucca.it`

