

Comprehension as compression?

Università di Salerno

24 July 2026



Giovanni Sileno

g.sileno@uva.nl

University of Amsterdam

About me

- M.Sc. Electronics Engineering, Politecnico of Torino and ENSIMAG (1998–2003)



About me

- M.Sc. Electronics Engineering, Politecnico of Torino and ENSIMAG (1998–2003)
- PhD in AI & Law, Leibniz Center for Law, Law Faculty, UvA (2011–2016)



About me

- M.Sc. Electronics Engineering, Politecnico of Torino and ENSIMAG (1998–2003)
- PhD in AI & Law, Leibniz Center for Law, Law Faculty, UvA (2011–2016)
- Postdoc in Cognitive Systems, Télécom Paris and Paris Dauphine (2016–2018)



About me

- M.Sc. Electronics Engineering, Politecnico of Torino and ENSIMAG (1998–2003)
- PhD in AI & Law, Leibniz Center for Law, Law Faculty, UvA (2011–2016)
- Postdoc in Cognitive Systems, Télécom Paris and Paris Dauphine (2016–2018)
- Postdoc in Regulatory Systems, CCI/SNE, Ivi, UvA (2018–2021)



About me

- M.Sc. Electronics Engineering, Politecnico of Torino and ENSIMAG (1998–2003)
- PhD in AI & Law, Leibniz Center for Law, Law Faculty, UvA (2011–2016)
- Postdoc in Cognitive Systems, Télécom Paris and Paris Dauphine (2016–2018)
- Postdoc in Regulatory Systems, CCI/SNE, Ivi, UvA (2018–2021)
- Assistant Professor, SIAS, Ivi, UvA (2021–...)
 - Director Bachelor Information Science, FNWI/UvA (2023–...)



About me

- M.Sc. Electronics Engineering, Politecnico of Torino and ENSIMAG (1998–2003)
- PhD in AI & Law, Leibniz Center for Law, Law Faculty, UvA (2011–2016)
- Postdoc in Cognitive Systems, Télécom Paris and Paris Dauphine (2016–2018)
- Postdoc in Regulatory Systems, CCI/SNE, Ivi, UvA (2018–2021)
- Assistant Professor, SIAS, Ivi, UvA (2021–...)
 - Director Bachelor Information Science, FNWI/UvA (2023–...)

	M.Sc.	PhD	Postdoc 1	Postdoc 2	Ass. Prof
COMPUTATIONAL TRACK					
NORMATIVE SYSTEMS TRACK					
COGNITIVE MODELLING TRACK					
SOCIO-TECHNICAL SYSTEMS TRACK					

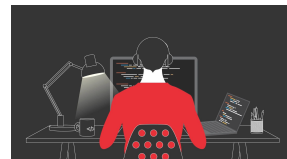
About me



business analyst (2y)



musician + freelance developer (5y)



- M.Sc. Electronics Engineering, Politecnico of Torino and ENSIMAG (1998–2003)
- PhD in AI & Law, Leibniz Center for Law, Law Faculty, UvA (2011–2016)
- Postdoc in Cognitive Systems, Télécom Paris and Paris Dauphine (2016–2018)
- Postdoc in Regulatory Systems, CCI/SNE, Ivi, UvA (2018–2021)
- Assistant Professor, SIAS, Ivi, UvA (2021–...)
 - Director Bachelor Information Science, FNWI/UvA (2023–...)

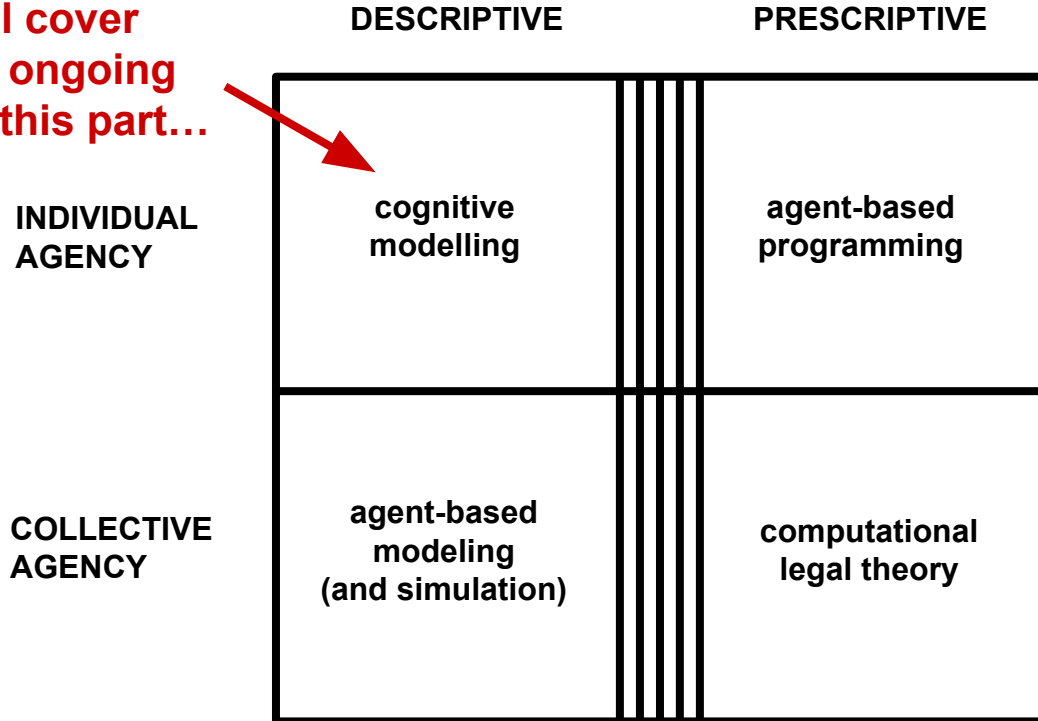
	M.Sc.	PhD	Postdoc 1	Postdoc 2	Ass. Prof
COMPUTATIONAL TRACK					
NORMATIVE SYSTEMS TRACK					
COGNITIVE MODELLING TRACK					
SOCIO-TECHNICAL SYSTEMS TRACK					

My research efforts/interests

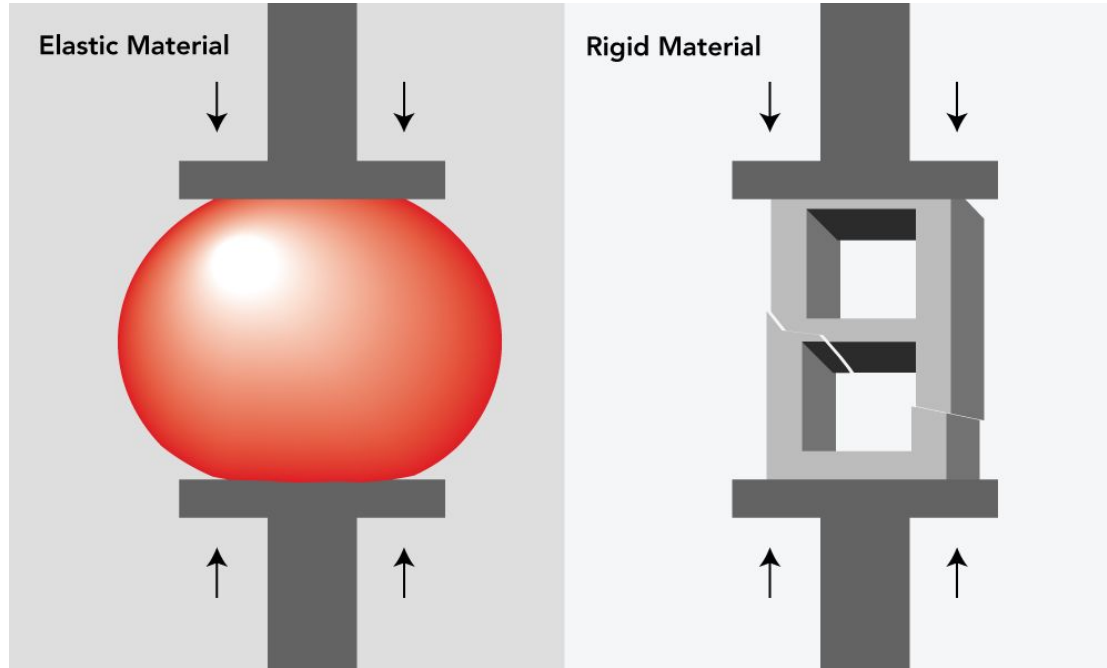
	DESCRIPTIVE	PRESCRIPTIVE
INDIVIDUAL AGENCY	cognitive modelling	agent-based programming
COLLECTIVE AGENCY	agent-based modeling (and simulation)	computational legal theory

My research efforts/interests

**This talk will cover
some of my ongoing
research in this part...**

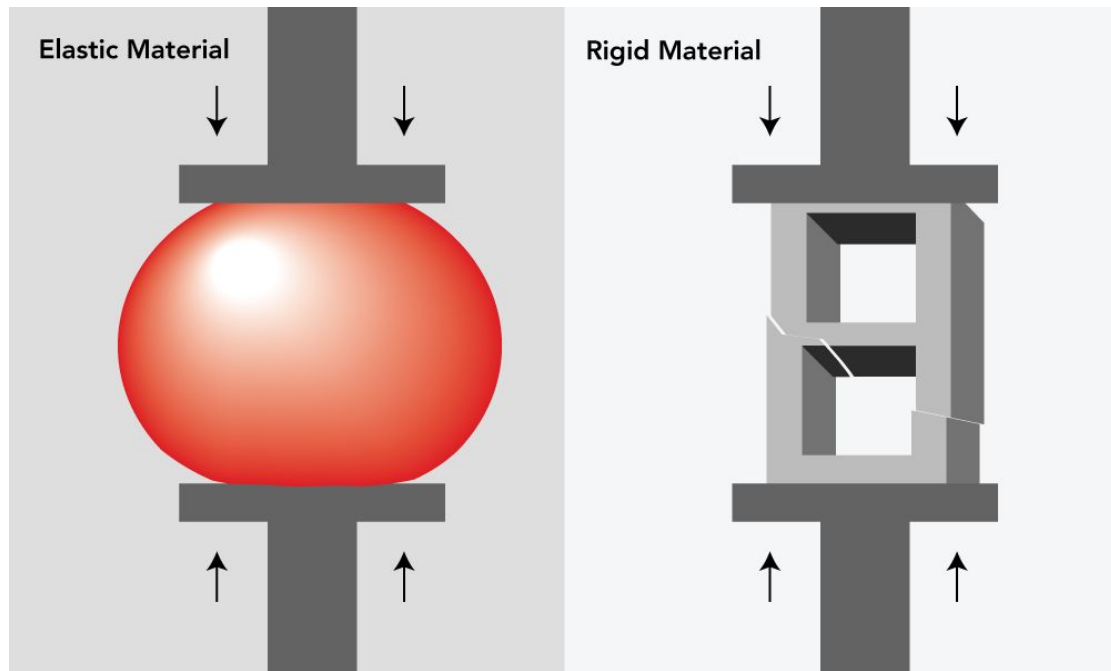


Compression at large



In physics it is about having the same material occupying less space.

Compression at large



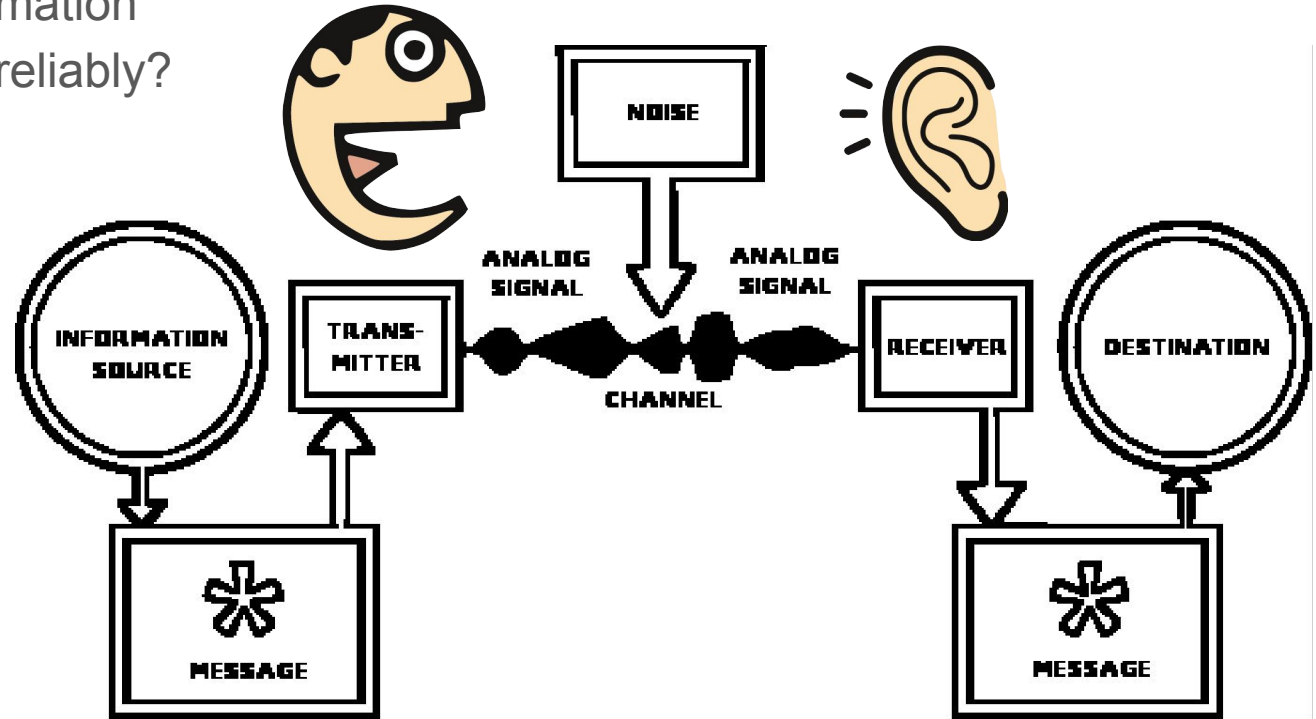
In physics it is about having the same material occupying less space.

In digital systems it is about **transmitting the same content with less symbols** (supported by adequate encoding and decoding algorithms)



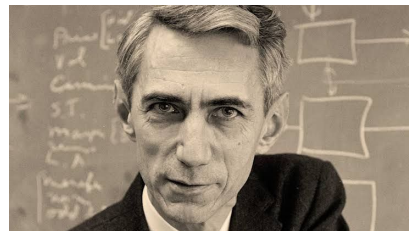
Given a certain communication channel...

- How much “information” can we transmit reliably?

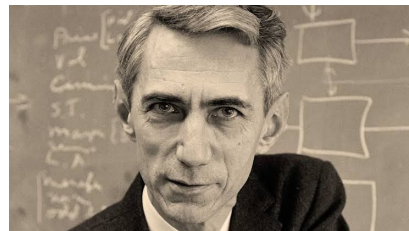


Shannon's information theory (1)

- “Information” can be quantified as as a *measure of **ignorance*** (of the observer).



Shannon's information theory (1)



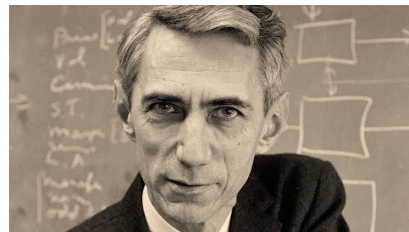
- “Information” can be quantified as as a *measure of **ignorance*** (of the observer).

For instance,

- I have chosen a card from a deck of 16 cards, ranked from 1 to 16.
- How many **yes/no** questions do you have to ask to find out what card I have in my hand?

$$\#questions = \log_2(\#symbols)$$

Shannon's information theory (1)



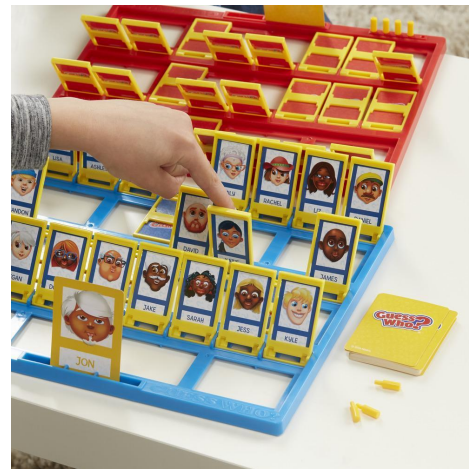
- “Information” can be quantified as as a *measure of ignorance* (of the observer).

cf. “Who’s who?” board game!

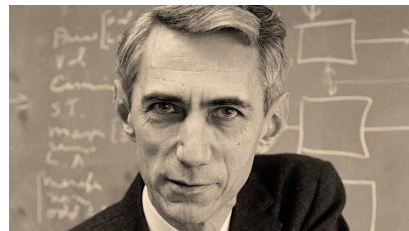
For instance,

- I have chosen a card from a deck of 16 cards, ranked from 1 to 16.
- How many **yes/no** questions do you have to ask to find out what card I have in my hand?

$$\#questions = \log_2(\#symbols)$$



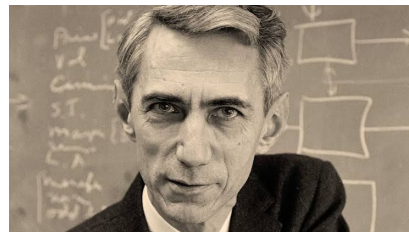
Shannon's information theory (2)



- We can connect this ignorance to the uncertainty of appearance of a certain symbol at the source:

$$I(x) = \log_2 \frac{1}{p(x)} = -\log_2 p(x)$$

Shannon's information theory (3)



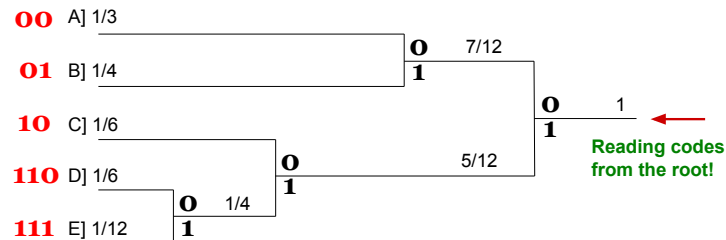
- We can connect this ignorance to the uncertainty of appearance of a certain symbol at the source:

$$I(x) = \log_2 \frac{1}{p(x)} = -\log_2 p(x)$$

- To reduce the occupation of the communication channel, we can **associate shorter codes to more frequent symbols** (less informative)

e.g. via Huffman encoding

→ **COMPRESSION!**



An interesting paradox...

According to Shannon's **theory of information**:

$$I(x) = -\log P(x)$$

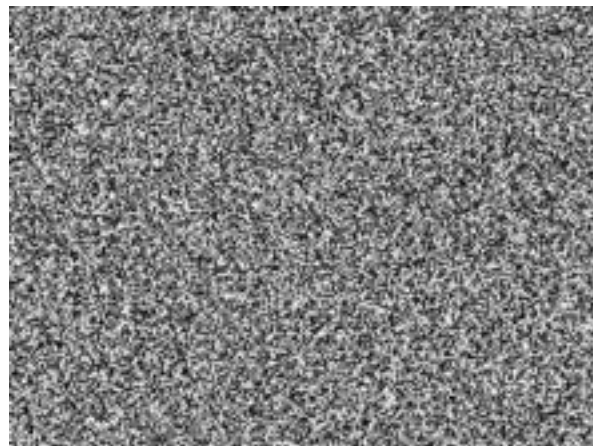
An interesting paradox...

According to Shannon's **theory of information**:

$$I(x) = -\log P(x)$$



**UNIFORM
NOISE SOURCE**
maximally informative



An interesting paradox...

According to Shannon's **theory of information**:

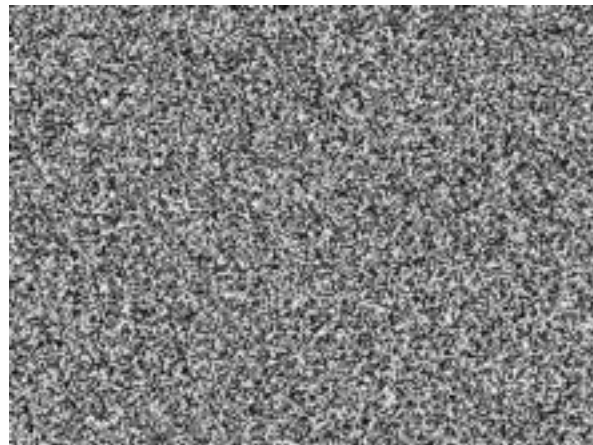
$$I(x) = -\log P(x)$$



**UNIFORM
NOISE SOURCE**
maximally informative



but this mismatches
relevance judgments
given by humans



SO WHAT?

An experiment!

- Suppose you are given for free one of two possible tickets of a lottery.

11111 11111 11111 **A**

15321 64378 81374 **B**



- Which one would you choose, before the lottery to occur?

An experiment!

- Suppose you are given for free one of two possible tickets of a lottery.

11111 11111 11111 **A**

15321 64378 81374 **B**



- After the extraction, which would be the most plausible winning ticket?

An experiment!

- Suppose you are given for free one of two possible tickets of a lottery.

A
11111 11111 11111

B
15321 64378 81374



People generally indicate A as the most improbable/implausible.

An experiment!

- Suppose you are given for free one of two possible tickets of a lottery.

A
11111 11111 11111

B
15321 64378 81374



People generally indicate A as the most improbable/implausible.

Yet, the prior/posterior probability is the same (in a fair lottery).

An experiment!

- Suppose you are given for free one of two possible tickets of a lottery.

A
11111 11111 11111

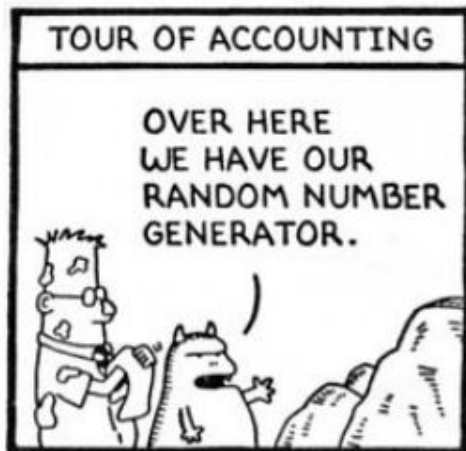
B
15321 64378 81374



People generally indicate A as the most improbable/implausible.

Yet, the prior/posterior probability is the same (in a fair lottery).

If the outcome is confirmed to be A, people think that the lottery is unfair.



10/45/01 © 2001 United Feature Syndicate, Inc.

What's the difference?

A

11111 11111 11111

B

15321 64378 81374

What's the difference?

11111 11111 11111 **A**

15321 64378 81374 **B**

Ticket A is ***descriptively*** much simpler! I need only to repeat 1 for 15 times.

What's the difference?

11111 11111 11111 **A**

Ticket A is *descriptively* much simpler! I need only to repeat 1 for 15 times.

15321 64378 81374 **B**

For Ticket B I need essentially to remember all numbers

Descriptive complexity



Andrei Kolmogorov

- In algorithmic information theory (AIT), the *complexity* of a string is the **minimal length of a program** that, given a certain optional input parameter, produces that string as an output (***Kolmogorov complexity***)

Descriptive complexity, examples



Andrei Kolmogorov

- A string like `aaaaa` is not very complex:

```
for i in range(5):  
    print("a")
```

Descriptive complexity, examples



Andrei Kolmogorov

- A string like `aaaaa` is not very complex:

```
for i in range(5):  
    print("a")
```

- A string like `4ta5J` is more complex:

```
print("4")  
print("t")  
print("a")  
print("5")  
print("J")
```

Descriptive complexity, examples



Andrei Kolmogorov

- A string like `aaaaa` is not very complex:

```
for i in range(5):  
    print("a")
```

- A string like `4ta5J` is more complex:

```
print("4")  
print("t")  
print("a")  
print("5")  
print("J")
```

- Is π complex?



Andrei Kolmogorov

Descriptive complexity, examples

- A string like `aaaaa` is not very complex:

```
for i in range(5):  
    print("a")
```

- A string like `4ta5J` is more complex:

```
print("4")  
print("t")  
print("a")  
print("5")  
print("J")
```

- Is π complex?

$$\pi/4 = 1 - 1/3 + 1/5 - 1/7 + 1/9 - \dots$$



Andrei Kolmogorov

Descriptive complexity, examples

- A string like `aaaaa` is not very complex:

```
for i in range(5):  
    print("a")
```

- A string like `4ta5J` is more complex:

```
print("4")  
print("t")  
print("a")  
print("5")  
print("J")
```

- Is π complex?

$$\pi/4 = 1 - 1/3 + 1/5 - 1/7 + 1/9 - \dots$$

*even if it is irrational number, the algorithm
producing it (to some precision) is rather simple!*



Andrei Kolmogorov

Descriptive complexity

- In algorithmic information theory (AIT), the *complexity* of a binary string is the **minimal length of a program** that, given a certain optional input parameter, produces that string as an output (*Kolmogorov complexity*)

$$K_{\phi}(x|y) = \min_p \{ |p| : p(y) = x \}$$

underlying
Turing machine

target string

additional input in support

executable program

Descriptive complexity



Andrei Kolmogorov

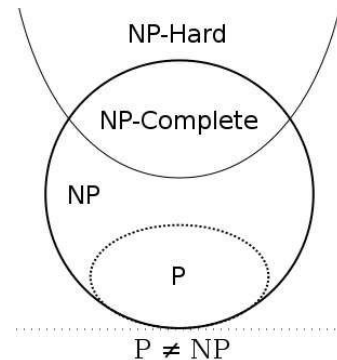
- In algorithmic information theory (AIT), the *complexity* of a string is the **minimal length of a program** that, given a certain optional input parameter, produces that string as an output (**Kolmogorov complexity**)

$$K_{\phi}(x|y) = \min \{ |p| : p(y) = x \}$$

how much information is needed for a program constructing the object (Kolmogorov complexity)

$\neq$

how much time or space is needed for running it (algorithmic or time-complexity)



relevant for *tractability analysis*

Randomness

- Are both these sequences equally random?
 - 00000000000000111111111111
 - 10010011011000111010110010



Ray Solomonoff

Randomness



Ray Solomonoff

- Are both these sequences equally random?

- 0000000000000001111111111111

- 10010011011000111010110010

$$\sum_{p:U(p)=x} \left(\frac{1}{2}\right)^{\text{length}(p)} = \sum_{p:U(p)=x} 2^{-\text{length}(p)}$$

- We can think this question in terms of **algorithmic probability**, i.e. the probability of a sequence as the sum of the probabilities of having a program with that output, when the input of that turing machine is random. *If the program is shorter it is more likely!!!*

Randomness



Ray Solomonoff

- Are both these sequences equally random?

- 0000000000000001111111111111

- 10010011011000111010110010

$$\sum_{p:U(p)=x} \left(\frac{1}{2} \right)^{\text{length}(p)}$$

- We can think this question in terms of **algorithmic probability**. The probability of a sequence as the sum of the probabilities of all programs that output that sequence, when the input of that Turing machine is random. *If the program is shorter it is more likely!!!*



OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."

Randomness and compression



Ray Solomonoff

- Are both these sequences equally random?

- 00000000000000111111111111

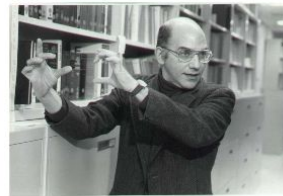
- 10010011011000111010110010

$$\sum_{p:U(p)=x} \left(\frac{1}{2}\right)^{\text{length}(p)} = \sum_{p:U(p)=x} 2^{-\text{length}(p)}$$

- We can think this question in terms of **algorithmic probability**, i.e. the probability of a sequence as the sum of the probabilities of having a program with that output, when the input of that turing machine is random. *If the program is shorter it is more likely!!!*
- A finite sequence is said to be **random** if it is **incompressible**, i.e. if its **shortest description** is the sequence itself.

Deduction

- Deduction generally works from the general to the particular



Gregory Chaitin

Deduction



Gregory Chaitin

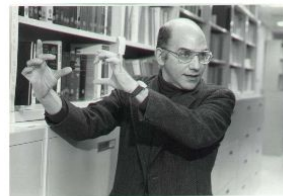
- Deduction generally works from the general to the particular

general premise and

particular premise → particular conclusion

all animals eat, fido is an animal → fido eats.

Deduction



Gregory Chaitin

- Deduction generally works from the general to the particular

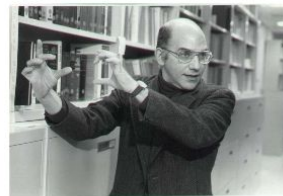
*general premise and
particular premise $\rightarrow$ particular conclusion*

all animals eat, fido is an animal $\rightarrow$ fido eats.

*general premise and
less general premises $\rightarrow$ less general conclusion*

all animals eat, cats are animals $\rightarrow$ cats eat.

Deduction



Gregory Chaitin

- Deduction generally works from the general to the particular

*general premise and
particular premise $\rightarrow$ particular conclusion*

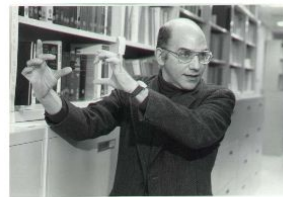
all animals eat, fido is an animal $\rightarrow$ fido eats.

*general premise and
less general premises $\rightarrow$ less general conclusion*

all animals eat, cats are animals $\rightarrow$ cats eat.

***reusing the most
general premises, I
can derive plenty of
less general
conclusions!!!***

Deduction and compression



Gregory Chaitin

- Deduction generally works from the general to the particular.
- ***Intuition***: A formal system (axioms AND derivation mechanisms) is a compression of the set of theorems it can prove.

understanding is compressing
comprehending is compressing

Minimum Description Length

- The **Minimum Description Length (MDL)** principle states that:

the best theory to describe **observed data** is the one which *minimizes* the sum of the description length (in bits) of:

- the **theory** description
- the **data encoded** for that theory



OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE
EXPLANATIONS, THE SIMPLER OF
THE TWO IS THE ONE MOST
LIKELY TO BE TRUE."

Minimum Description Length

- The **Minimum Description Length (MDL)** principle states that:

the best theory to describe **observed data** is the one which *minimizes* the sum of the description length (in bits) of:

- the **theory** description
- the **data encoded** for that theory

OBSERVED DATA

"aaaaa"



THEORY or ALGORITHM

```
for i in range(X):  
    print(Y)
```

DATA

```
X = 5; Y = "a"
```



OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."

General properties of Kolmogorov complexity

- **Invariance theorem**

$$|K_{\phi_1}(x) - K_{\phi_2}(x)| \leq c$$

Given two universal Turing machines and a string x , complexities will vary up to a certain constant (eg. the length of the program to translate programs for one machine to the other machine).

General properties of Kolmogorov complexity

- **Invariance theorem**

$$|K_{\phi_1}(x) - K_{\phi_2}(x)| \leq c$$

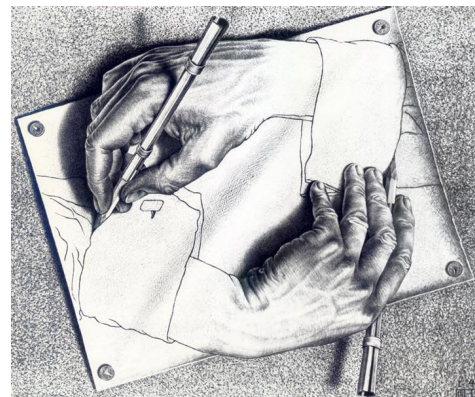
Given two universal Turing machines and a string x , complexities will vary up to a certain constant (eg. the length of the program to translate programs for one machine to the other machine).

- **Chain rule** (for binary strings, requiring terminal characters)

$$K(xy) \leq K(x) + K(y|x) + O(\log_2 K(x))$$

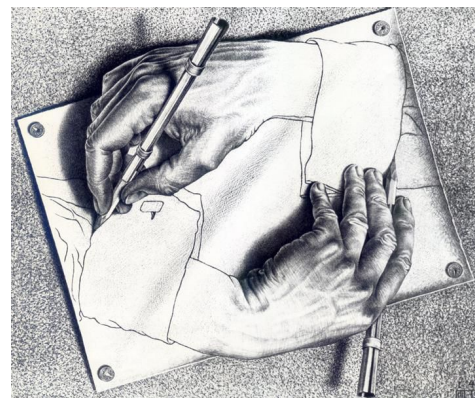
Practical problems of Kolmogorov complexity

- In the general case of a Universal Turing Machine, *Kolmogorov complexity is not computable.*



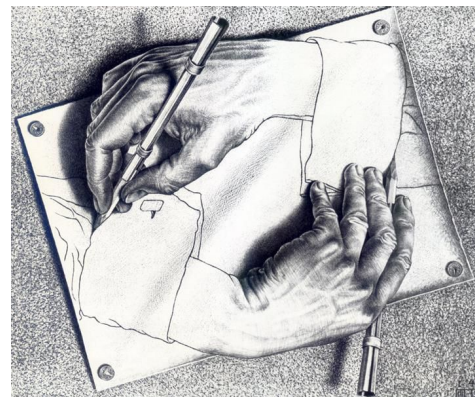
Practical problems of Kolmogorov complexity

- In the general case of a Universal Turing Machine, *Kolmogorov complexity is not computable.*
- But if the underlying machine is not universal, i.e. a number of finite operators/memory, then it is computable...



Practical problems of Kolmogorov complexity

- In the general case of a Universal Turing Machine, *Kolmogorov complexity is not computable.*
- But if the underlying machine is not universal, i.e. a number of finite operators/memory, then it is computable...
- **As a matter of fact, the human brain is not a universal Turing machine!**



Kolmogorov complexity as cognitive measure

- If we accept that minds are
 - **bounded** Turing Machines,
 - do not realize **deeply recursive algorithmic structures**

we can abstract time-complexity (all inferences will be tractable) and just focus on the simplest inferential resources producing the target object!

Question

- Is π complex?

Question

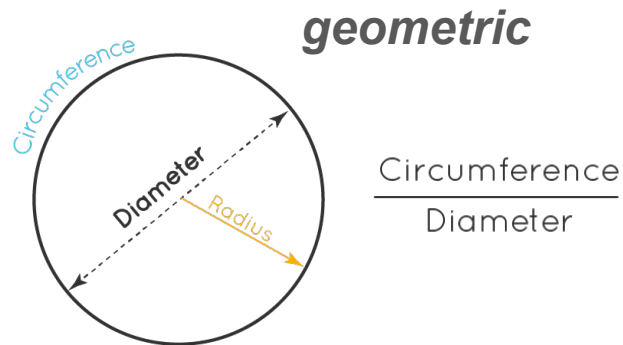


- Is π complex?
 - “referential” view on π

Question



- Is π complex?
 - “referential” view on π
 - “algorithmic” views on π



geometric

$$\pi/4 = 1 - 1/3 + 1/5 - 1/7 + 1/9 - \dots$$

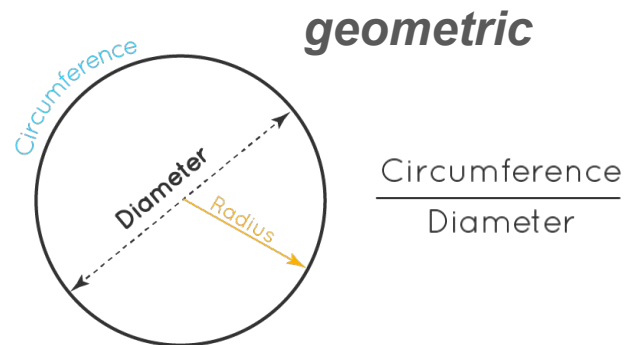
algebraic

Question



- Is π complex?

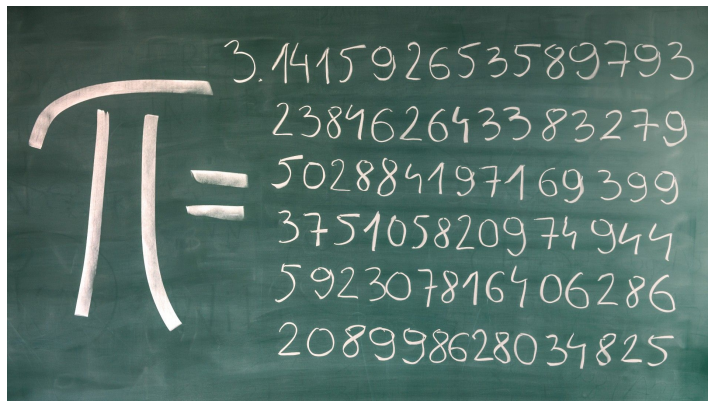
- “referential” view on π
- “algorithmic” views on π
- “value” view on π



geometric

$$\pi/4 = 1 - 1/3 + 1/5 - 1/7 + 1/9 - \dots$$

algebraic

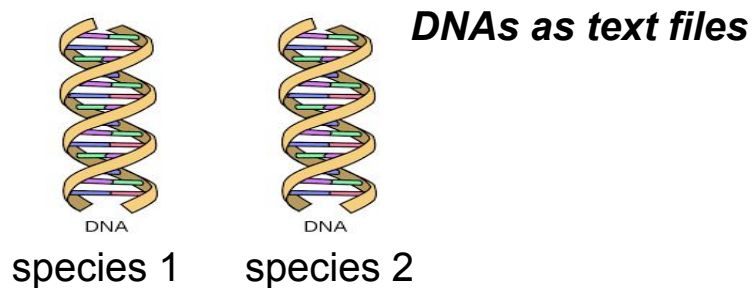


Why is this relevant for you?

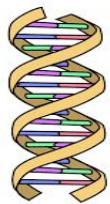
Why is this relevant for you?

Compression is key to abstraction!

Compression is key to abstraction!

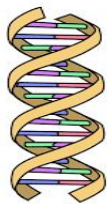


Compression is key to abstraction!



DNA

species 1

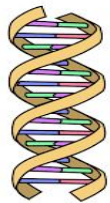


DNA

species 2

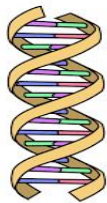
DNAs as text files

concatenate files and “zip”



DNA

species 1

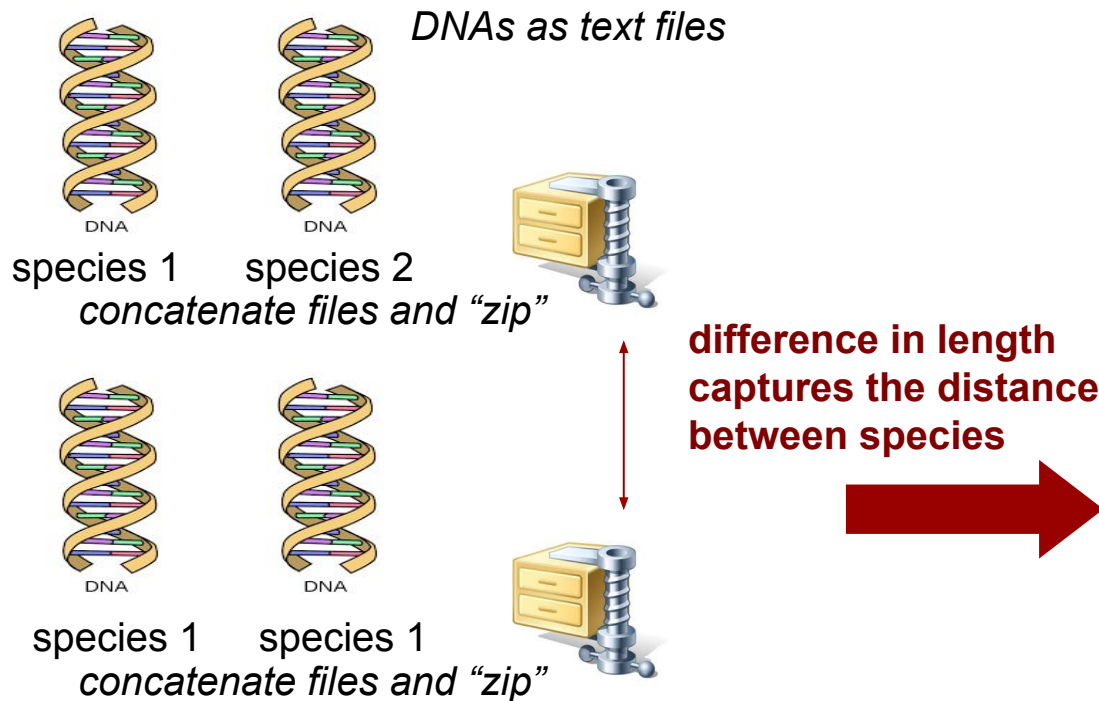


DNA

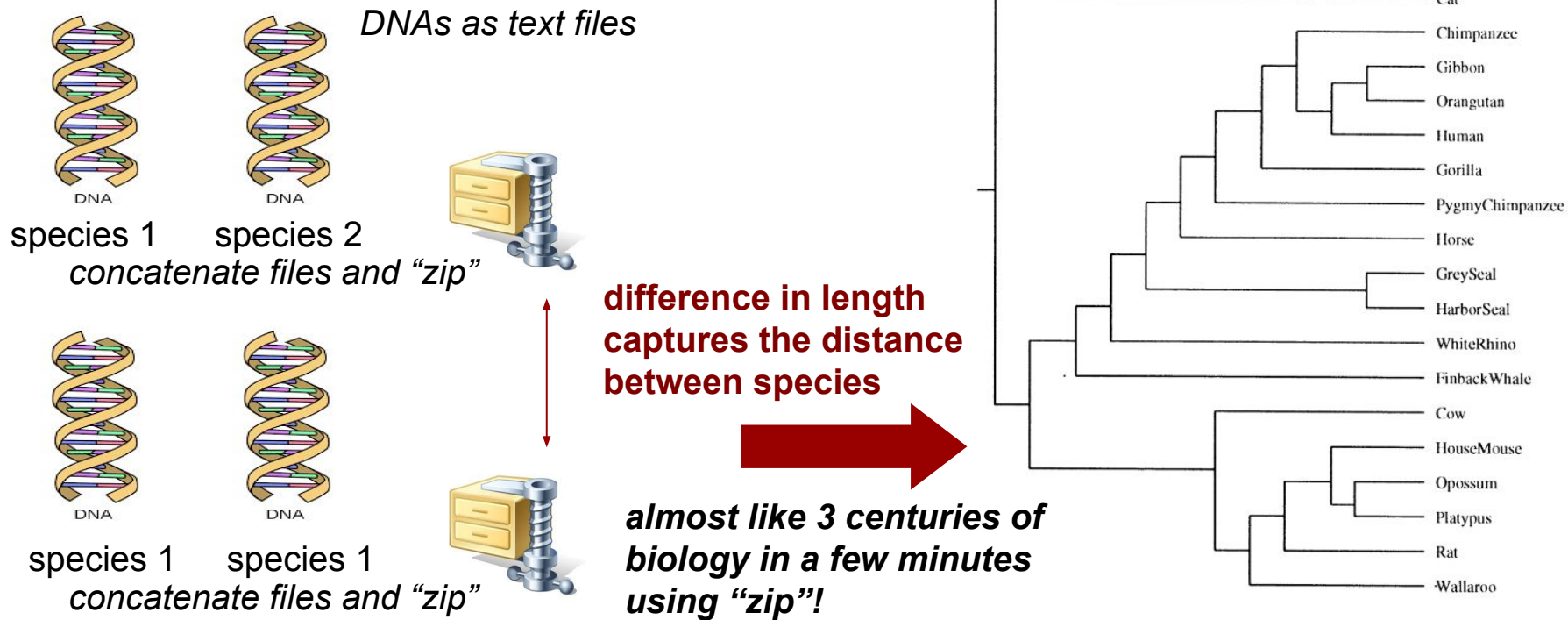
species 1

concatenate files and “zip”

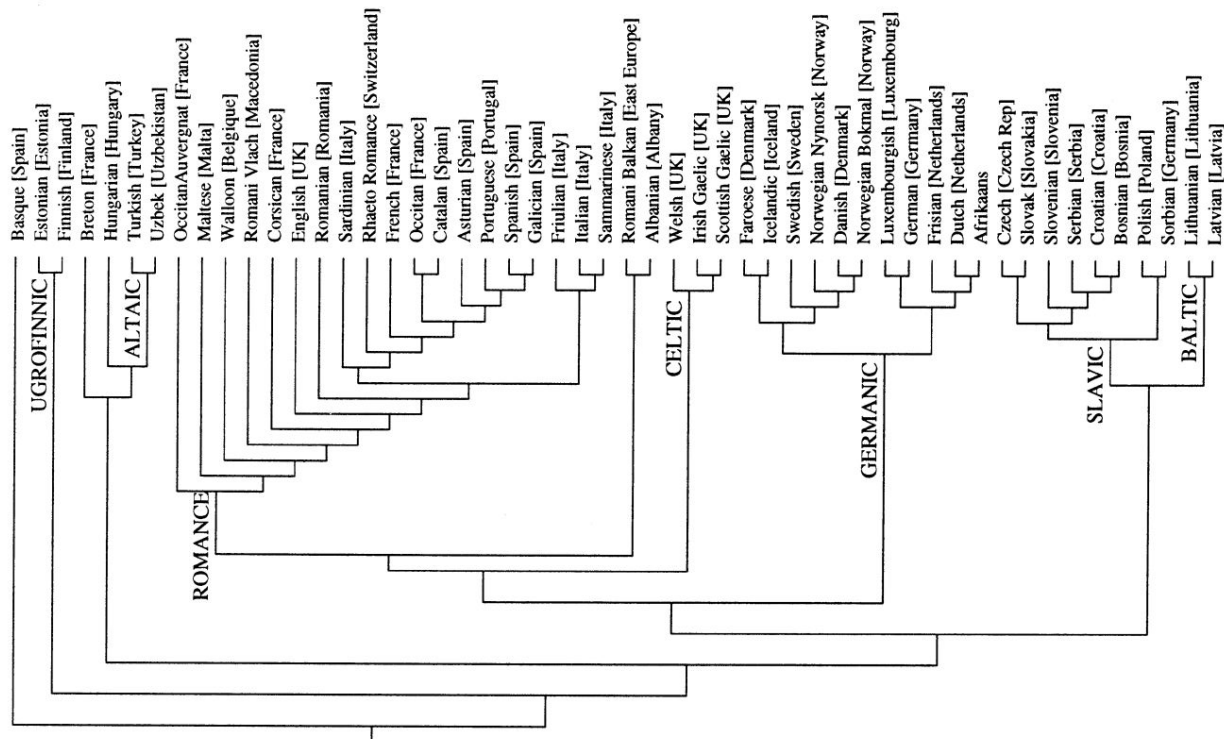
Compression is key to abstraction!



Compression is key to abstraction!

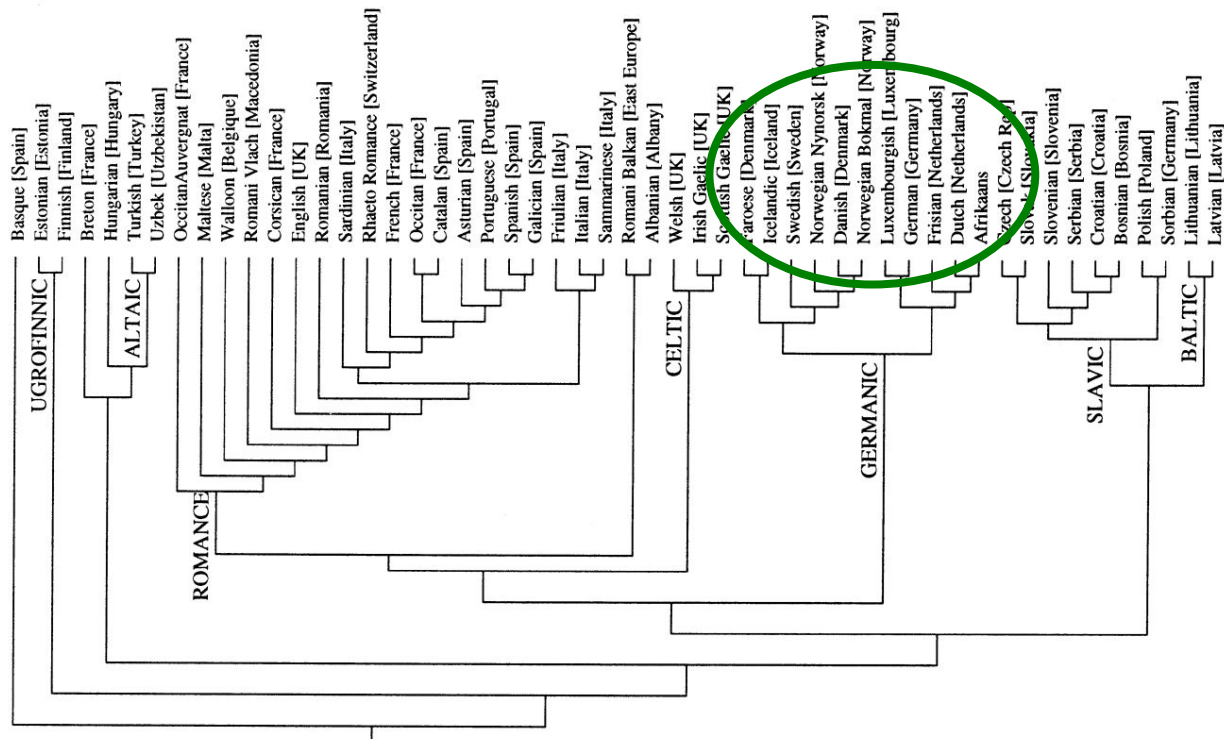


Compression is key to abstraction!



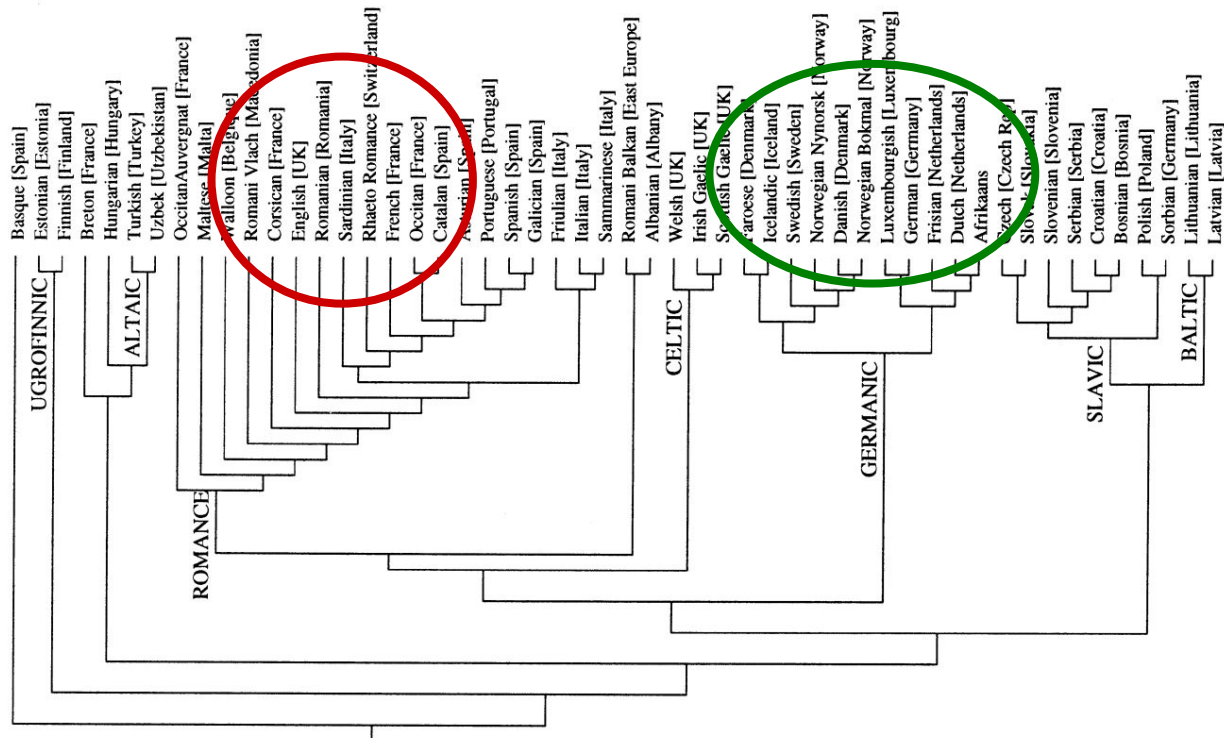
They also extracted a **“tree” of languages** (less impressive, but still very interesting), doing the same with translations of the Declaration of Human Rights!

Compression is key to abstraction!



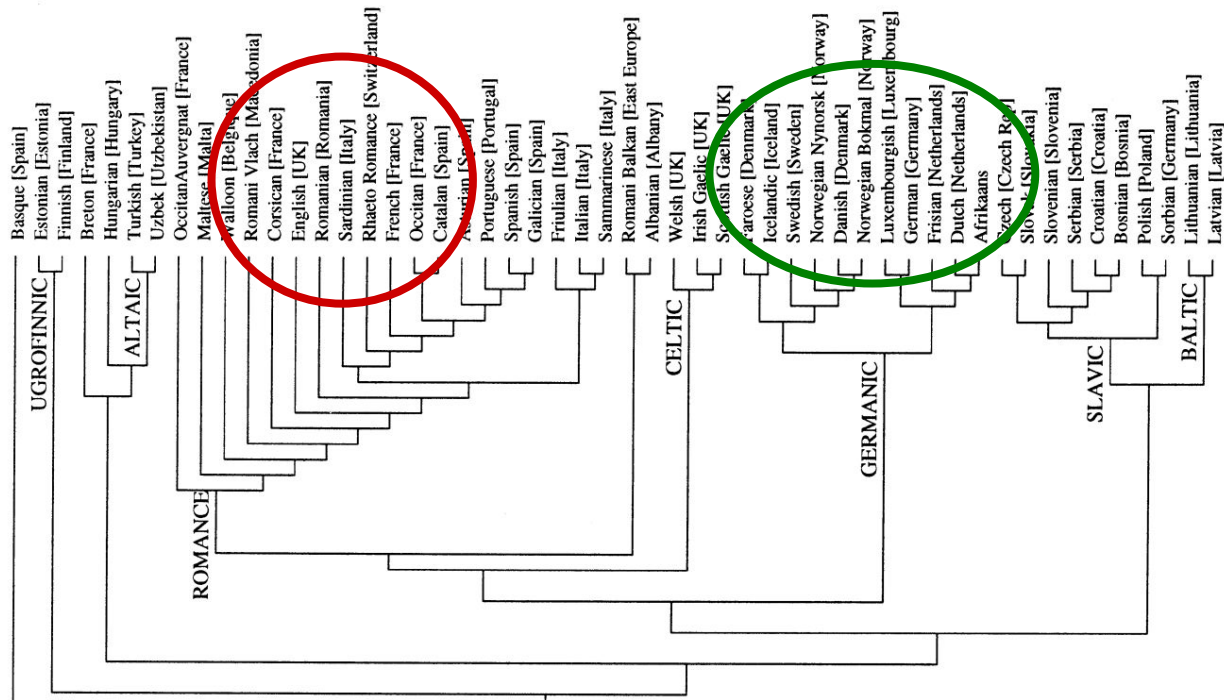
They also extracted a **“tree” of languages** (less impressive, but still very interesting), doing the same with translations of the Declaration of Human Rights!

Compression is key to abstraction!



They also extracted a **“tree” of languages** (less impressive, but still very interesting), doing the same with translations of the Declaration of Human Rights!

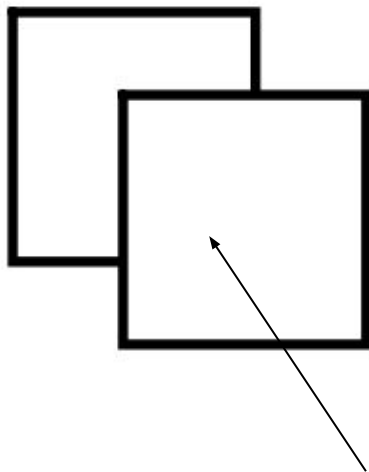
Compression is key to abstraction!



They also extracted a **“tree” of languages** (less impressive, but still very interesting), doing the same with translations of the Declaration of Human Rights!

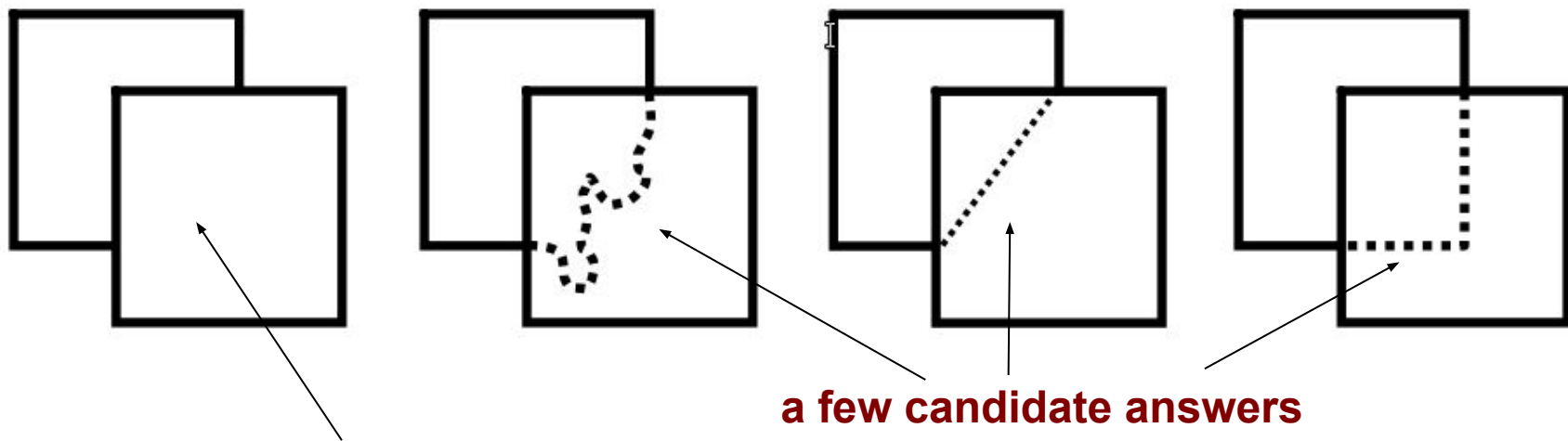
but which compressor to use?

A related problem in perception...



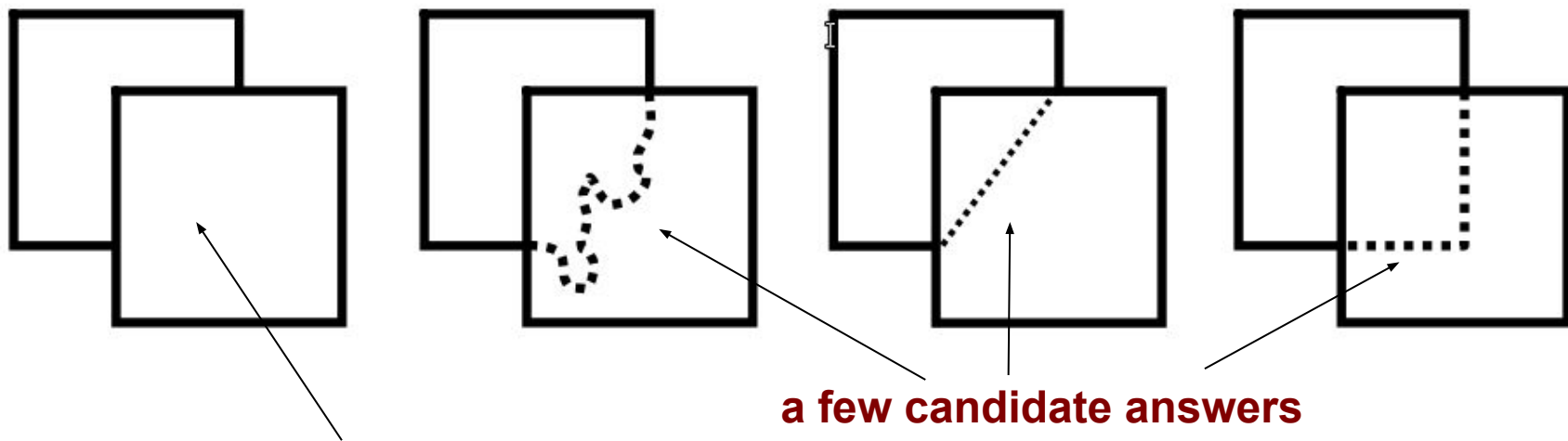
What is behind the square at the top?

A related problem in perception...



What is behind the square at the top?

A related problem in perception...



What is behind the square at the top?

AS WELL IN OTHER DOMAINS...

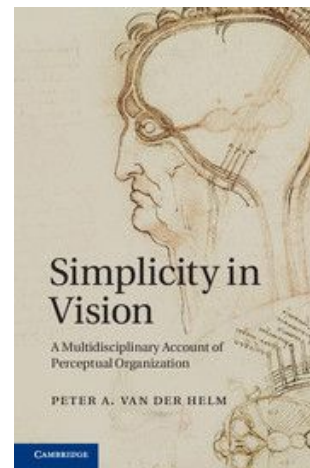
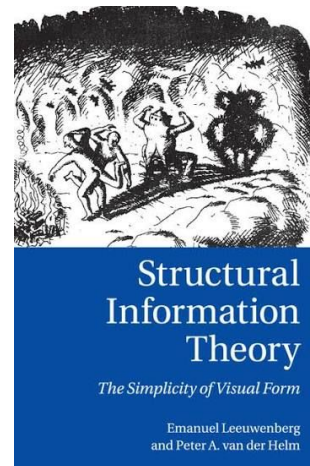
Structural Information Theory

- A perceptual theory about *Gestalt* effects (emerging organization), non-probabilistic, symbol-based

Input: strings

- SIT coding language based on 3 operators:
 - *Iteration* $3^*(A)$ $\Rightarrow$ **AAA**
 - *Symmetry* $S[(A)(B),(C)]$ $\Rightarrow$ **ABCBA**
 - *Alternation* $<(A)>/<(B)(C)>$ $\Rightarrow$ **ABAC**

Leeuwenberg, E., & van der Helm, P. A. (2013). Structural Information Theory: The Simplicity of Visual Form. Cambridge University Press.

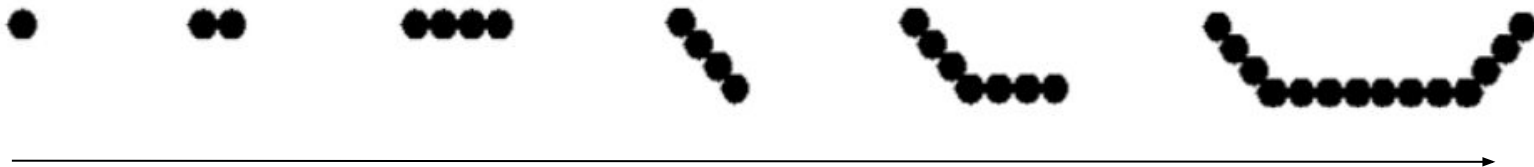


Generative Theory of Shape

- Perceptual theory about perceptual organization, based on generative grammars, non-probabilistic

Input: 2D images

- Operators: *translation*, *rotation* and *mirroring/reflection*



Borsa, D., Graepel, T., & Gordon, A. (2015). The Wreath Process: A totally generative model of geometric shape based on nested symmetries. Michael Leyton. A generative theory of shape. Springer, 2003.

Process Mining

- A number of efficient techniques developed to create process models out of event logs

Input: set of lists of strings (traces)

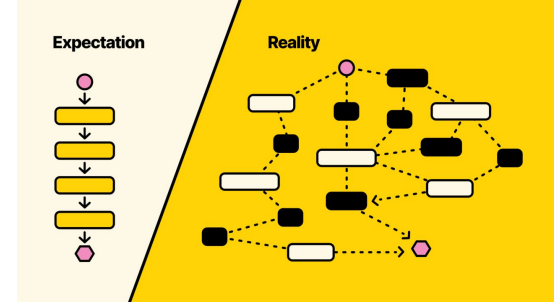
- Operators: *sequential, alternative, parallel branchings, loop*

ALT(a, b) => a; b PAR(a, b) => ab; ba

SEQ(a, b) => ab LOOP(a) => a; aa; aaa; aaaa; ...

SEQ(a, PAR(c, d), b) => acdb; adcb

SEQ(SEQ(a, b, c), LOOP(SEQ(b, d))) => abcbdbd; abcbd; abcbdbdbd; ...

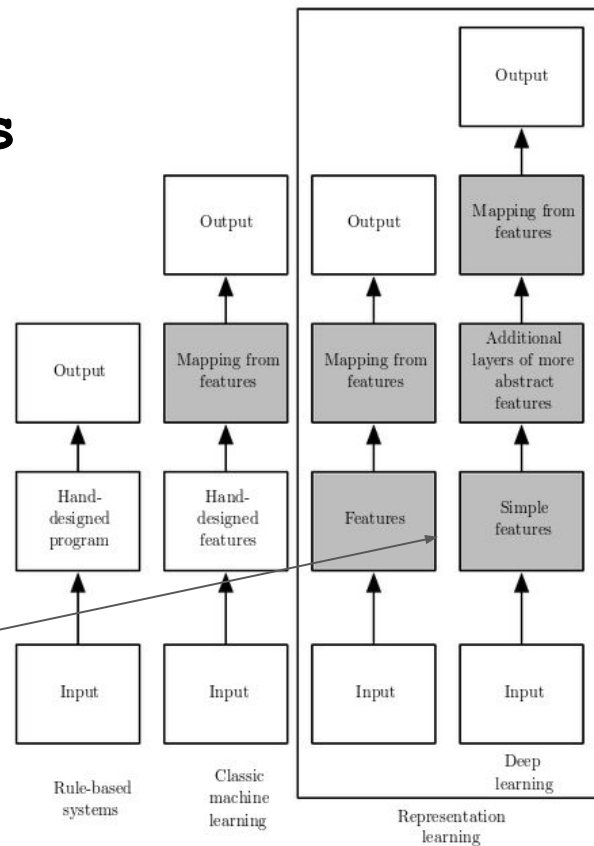


van der Aalst, W. M. P. (2010). Process Discovery: Capturing the Invisible. IEEE Computational Intelligence Magazine, 5(1), 28-41.
Weijters, A. J. M. M., Van der Aalst, W. M. P., & Alves de Medeiros, A. K. (2006). Process mining with the heuristics miner-algorithm. Technische Universiteit Eindhoven, Department of Technology Management.
Leemans, S. J., Fahland, D., & Van Der Aalst, W. M. (2013). Discovering block-structured process models from event logs—a general framework. In Business Process Management: 11th International Conference, BPM 2013.
van Der Aalst, W. M. P. (2016). Process Mining: Data Science in Action

Ok, Giovanni, but this sounds
all old-fashioned symbolic
stuff, what about machine
learning methods?

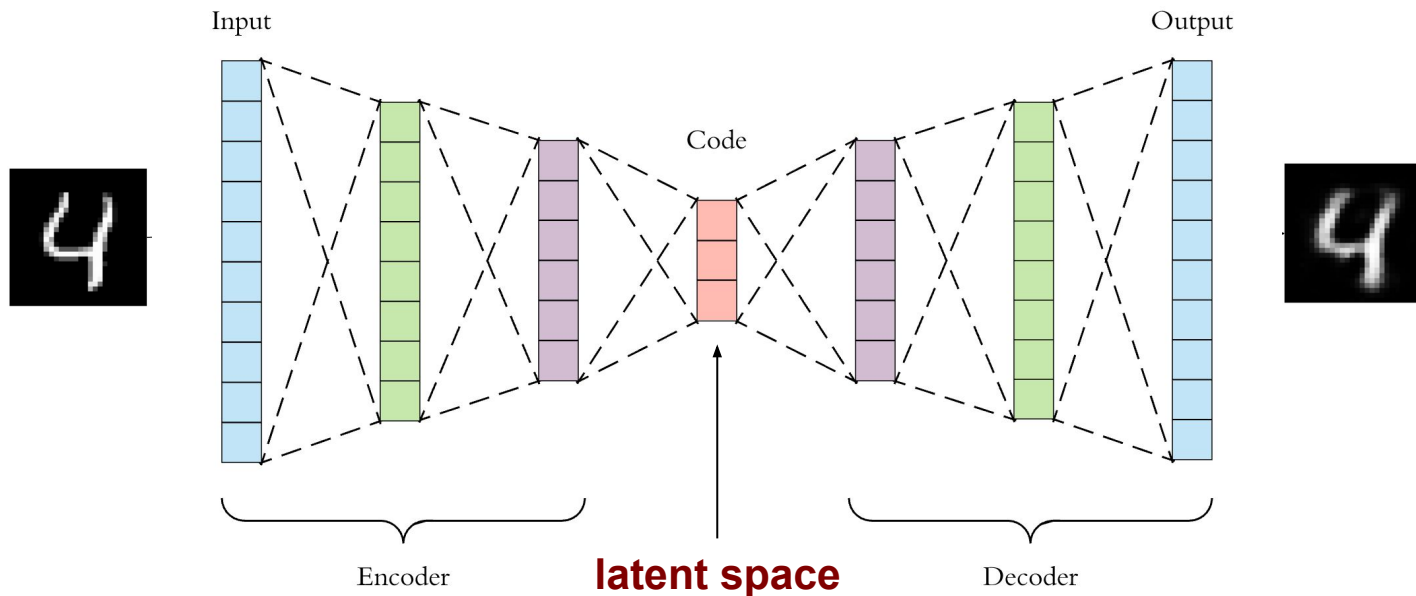
Ok, Giovanni, but this sounds all old-fashioned symbolic stuff, what about machine learning methods?

Well, a core part of contemporary ML techniques is **representation learning**, i.e. identifying features – the “latent space” – relevant to perform the task



Prototypical case: autoencoders!

- Autoencoders function explicitly as unsupervised “compressors”: what are the features that make the output as much similar as possible to the input



And obviously, LLMs!

Published as a conference paper at ICLR 2024

LANGUAGE MODELING IS COMPRESSION

Grégoire Delétang^{*1} Anian Ruoss^{*1} Paul-Ambroise Duquenne² Elliot Catt¹

Tim Genewein¹ Christopher Mattern¹ Jordi Grau-Moya¹ Li Kevin Wenliang¹

Matthew Aitchison¹ Laurent Orseau¹ Marcus Hutter¹ Joel Veness¹

ABSTRACT

It has long been established that predictive models can be transformed into lossless compressors and vice versa. Incidentally, in recent years, the machine learning community has focused on training increasingly large and powerful self-supervised (language) models. Since these large language models exhibit impressive predictive capabilities, they are well-positioned to be strong compressors. In this work, we advocate for viewing the prediction problem through the lens of compression and evaluate the compression capabilities of large (foundation) models. We show that large language models are powerful general-purpose predictors and that the compression viewpoint provides novel insights into scaling laws, tokenization, and in-context learning. For example, Chinchilla 70B, while trained primarily on text, compresses ImageNet patches to 43.4% and LibriSpeech samples to 16.4% of their raw size, beating domain-specific compressors like PNG (58.5%) or FLAC (30.3%), respectively. Finally, we show that the prediction-compression equivalence allows us to use any compressor (like gzip) to build a conditional generative model.

And obviously, LLMs!

although there are caveats:

Demystifying Verbatim Memorization in Large Language Models

Jing Huang, Diyi Yang, Christopher Potts

April 26, 2025

What is verbatim memorization?

In December 2023, The New York Times filed a lawsuit against OpenAI, alleging unauthorized use of its articles to train GPT models. The lawsuit included a 127-page exhibit demonstrating how GPT-4 could reproduce substantial portions of copyrighted articles verbatim when prompted with the opening sentences, as shown in Figure 1.

OpenAI from GPT-4:
The lawsuit included a 127-page exhibit demonstrating how GPT-4 could reproduce substantial portions of copyrighted articles verbatim when prompted with the opening sentences, as shown in Figure 1.

Rethinking LLM Memorization through the Lens of Adversarial Compression

Avi Schwarzschild¹
schwarzschild@cmu.edu
Carnegie Mellon University

Zhili Feng²
zhilif@andrew.cmu.edu
Carnegie Mellon University

Pratyush Maini¹
pratyushmaini@cmu.edu
Carnegie Mellon University

Zachary C. Lipton
Carnegie Mellon University

J. Zico Kolter
Carnegie Mellon University

Abstract

Large language models (LLMs) trained on web-scale datasets raise substantial concerns regarding permissible data usage. One major question is whether these models “memorize” all their training data or they integrate many data sources in some way more akin to how a human would learn and synthesize information. The answer hinges, to a large degree, on how we define memorization. In this work, we propose the Adversarial Compression Ratio (ACR) as a metric for assessing memorization in LLMs. A given string from the training data is considered memorized if it can be elicited by a prompt (much) shorter than the string itself—in other words, if these strings can be “compressed” with the model by computing adversarial prompts of fewer tokens. The ACR overcomes the limitations of existing

Published as a conference paper at ICLR 2024

LANGUAGE MODELING IS COMPRESSION

Grégoire Delétang^{*1} Anian Ruoss^{*1} Paul-Ambroise Duquenne² Elliot Catt¹

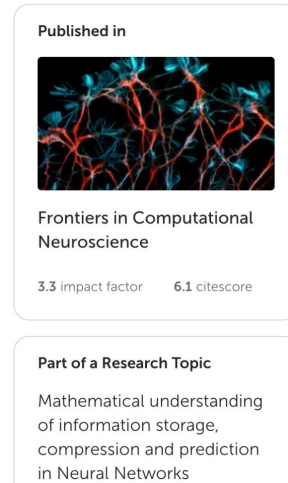
Tim Genewein¹ Christopher Mattern¹ Jordi Grau-Moya¹ Li Kevin Wenliang¹

Matthew Aitchison¹ Laurent Orseau¹ Marcus Hutter¹ Joel Veness¹

ABSTRACT

It has long been established that predictive models can be transformed into lossless compressors and vice versa. Incidentally, in recent years, the machine learning community has focused on training increasingly large and powerful self-supervised (language) models. Since these large language models exhibit impressive predictive capabilities, they are well-positioned to be strong compressors. In this work, we advocate for viewing the prediction problem through the lens of compression and evaluate the compression capabilities of large (foundation) models. We show that large language models are powerful general-purpose predictors and that the compression viewpoint provides novel insights into scaling laws, tokenization, and in-context learning. For example, Chinchilla 70B, while trained primarily on text, compresses ImageNet patches to 43.4% and LibriSpeech samples to 16.4% of their raw size, beating domain-specific compressors like PNG (58.5%) or FLAC (30.3%), respectively. Finally, we show that the prediction-compression equivalence allows us to use any compressor (like gzip) to build a conditional generative model.

A recent study confirms the connection between neural networks and Kolmogorov complexity...

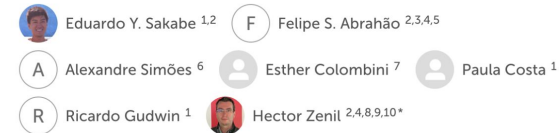


ORIGINAL RESEARCH article

Front. Comput. Neurosci., 29 May 2026

Volume 20 - 2026 | <https://doi.org/10.3389/fncom.2026.1791546>

Binarized neural networks converge toward algorithmic simplicity: empirical support for the learning-as-compression hypothesis



Sakabe EY, Abrahão FS, Simões A, Colombini E, Costa P, Gudwin R and Zenil H (2026) Binarized neural networks converge toward algorithmic simplicity: empirical support for the learning-as-compression hypothesis. Frontiers in Computational Neuroscience

And several people working on AGI is building upon these theories...

JMLR: Workshop and Conference Proceedings vol 40:1–16, 2015

Universal Intelligence: A Definition of Machine Intelligence

Shane Legg

IDSIA, Galleria 2, Manno-Lugano CH-6928, Switzerland
shane@vetta.org www.vetta.org/shane

Marcus Hutter

RSISE@ANU and SML@NICTA, Canberra, ACT, 0200, Australia
marcus@hutter1.net www.hutter1.net

December 2007

Abstract

A fundamental problem in artificial intelligence is that nobody really knows what intelligence is. The problem is especially acute when we need to consider artificial systems which are significantly different to humans. In this paper we approach this problem in the following way: We take a number of well known informal definitions of human intelligence that have been given by experts, and extract their essential features. These are then mathematically formalised to produce a general measure of intelligence for arbitrary machines. We believe that this equation formally captures the concept of machine intelligence in the broadest reasonable sense. We then show

Bad Universal Priors and Notions of Optimality

Jan Leike

Australian National University

JAN.LEIKE@ANU.EDU.AU

Marcus Hutter

Australian National University

MARCUS.HUTTER@ANU.EDU.AU

Abstract

A big open question of algorithmic information theory is the choice of the universal Turing machine (UTM). For Kolmogorov complexity and Solomonoff induction we have invariance theorems: the choice of the UTM changes bounds only by a constant. For the universally intelligent agent AIXI (Hutter, 2005) no invariance theorem is known. Our results are entirely negative: we discuss cases in which unlucky or adversarial choices of the UTM cause AIXI to misbehave drastically. We show that Legg-Hutter intelligence and thus balanced Pareto optimality is entirely subjective, and that every policy is Pareto optimal in the class of all computable environments. This undermines all existing optimality properties for AIXI. While it may still serve as a gold standard for AI, our results imply that AIXI is a *relative* theory, dependent on the choice of the UTM.

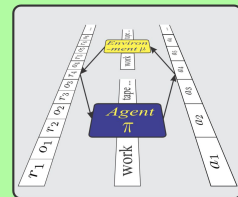
Keywords: AIXI, general reinforcement learning, universal Turing machine, Legg-Hutter intelligence, balanced Pareto optimality, asymptotic optimality.

An Introduction to Universal Artificial Intelligence

Marcus Hutter

David Quarel

Elliot Catt



(2024)

And several people working on AGI is building upon these theories...

JMLR: Workshop and Conference Proceedings vol 40:1–16, 2015

Universal Intelligence: A Definition of Machine Intelligence

Shane Legg

IDSIA, Galleria 2, Manno-Lugano CH-6928, Switzerland
shane@vetta.org www.vetta.org/shane

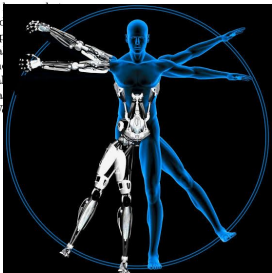
Marcus Hutter

RSISE@ANU and SML@NICTA, Canberra, ACT, 0200, Australia
marcus@hutter1.net www.hutter1.net

December 2007

Abstract

A fundamental problem in artificial intelligence is that nobody really knows what intelligence is. The problem is especially acute when we need to consider systems which are significantly different to humans. In this paper we approach the problem in the following way: We take a number of well known informal theories of human intelligence that have been given by experts, and extract their common features. These are then mathematically formalised to produce a general definition of intelligence for arbitrary machines. We believe that this equation formalises the concept of machine intelligence in the broadest reasonable sense. We



Bad Universal Priors and Notions of Optimality

Jan Leike

Australian National University

Marcus Hutter

Australian National University

JAN.LEIKE@ANU.EDU.AU

MARCUS.HUTTER@ANU.EDU.AU

Abstract

A big open question of algorithmic information theory is the choice of the universal Turing machine (UTM). For Kolmogorov complexity and Solomonoff induction we have invariance theorems: the choice of the UTM changes bounds only by a constant. For the universally intelligent agent AIXI (Hutter, 2005) no invariance theorem is known. Our results are entirely negative: we discuss cases in which unlucky or adversarial choices of the UTM cause AIXI to misbehave drastically. We show that Legg-Hutter intelligence and thus balanced Pareto optimality is entirely subjective, and that every policy is Pareto optimal in the class of all computable environments. This undermines all existing optimality properties for AIXI. While it may still serve as a gold standard for AI, our results imply that AIXI is a *relative* theory, dependent on the choice of the UTM.

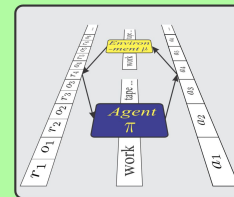
Keywords: AIXI, general reinforcement learning, universal Turing machine, Legg-Hutter intelligence, balanced Pareto optimality, asymptotic optimality.

An Introduction to Universal Artificial Intelligence

Marcus Hutter

David Quarell

Elliot Catt



(2024)

with some rather arguable “twist”...

And several people working on AGI is building upon these theories...

JMLR: Workshop and Conference Proceedings vol 40:1–16, 2015

Universal Intelligence: A Definition of Machine Intelligence

Shane Legg

IDSIA, Galleria 2, Manno-Lugano CH-6928, Switzerland
shane@vetta.org www.vetta.org/shane

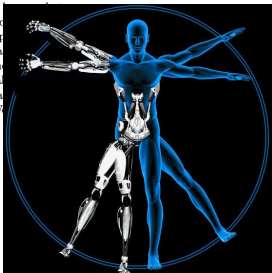
Marcus Hutter

RSISE@ANU and SML@NICTA, Canberra, ACT, 0200, Australia
marcus@hutter1.net www.hutter1.net

December 2007

Abstract

A fundamental problem in artificial intelligence is that nobody really understands intelligence. The problem is especially acute when we need to consider systems which are significantly different to humans. In this paper we approach the problem in the following way: We take a number of well known informal theories of human intelligence that have been given by experts, and extract their common features. These are then mathematically formalised to produce a general definition of intelligence for arbitrary machines. We believe that this equation formalises the concept of machine intelligence in the broadest reasonable sense. We



Bad Universal Priors and Notions of Optimality

Jan Leike

Australian National University

Marcus Hutter

Australian National University

JAN.LEIKE@ANU.EDU.AU

A big open question of algorithmic information theory (ITM). For Kolmogorov complexity (K), the choice of the universal Turing machine (UTM) is arbitrary. We discuss cases in which this choice is not optimal. We show that Legg-Hutter's definition of universal prior is not optimal, and that every policy that is optimal for AI, our results imply that it is not optimal for general intelligence, balanced Pareto optimality.

An Introduction to Universal Artificial Intelligence

Marcus Hutter

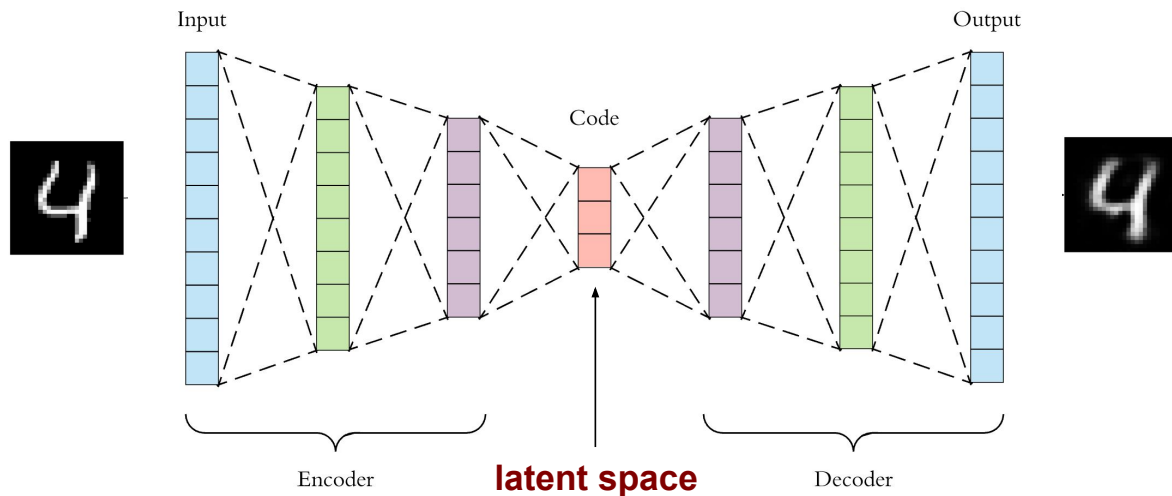
*Dedicated to all transhumans
and descendants of AIXI*

(2024)

with some rather arguable "twist"...

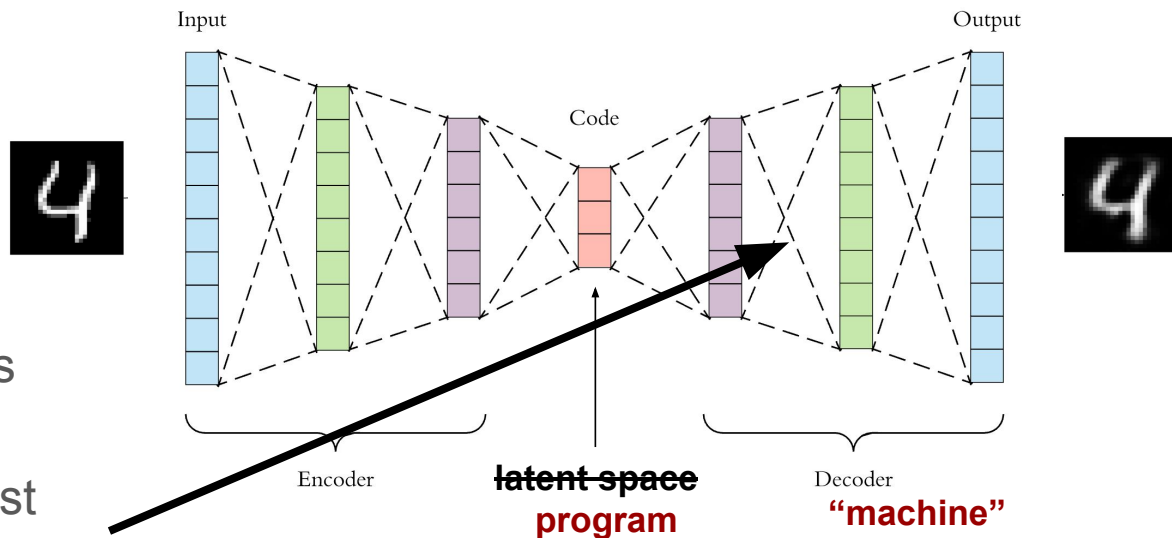
Yet, there is a difference!

- In ML we do not talk about explicitly about “operations” or “programs”, the latent space is about objects (parameters), all transformations are encoded within the network...



Yet, there is a difference!

- In ML we do not talk about explicitly about “operations” or “programs”, the latent space is about objects (parameters), all transformations are encoded within the network...



- Or rather, ML methods create the “Turing machine” which is most appropriate to the task

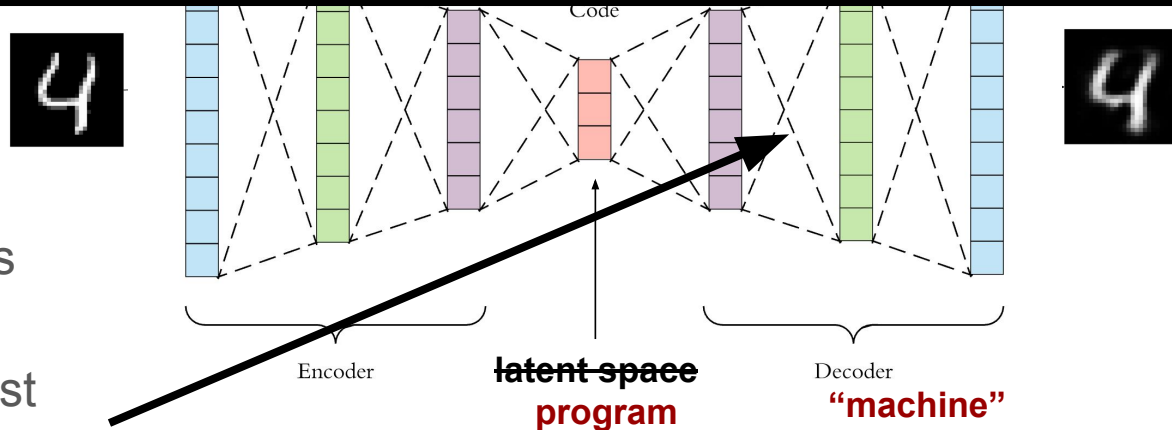
Yet, there is a difference!

- In ML we do not talk about explicitly about “operations” or “programs”, the latent space is about objects (parameters) all transformations are

YET, no alignment with human or natural cognition! (unless emergent, for evolutionary reasons, but *no guarantees it will follow the same path*).



- Or rather, ML methods create the “Turing machine” which is most appropriate to the task



Why natural intelligence is key

- Striving to find theories on how humans (but also other living entities) thinks and acts is an important endeavours for at least two reasons:
 - **efficiency reasons**: we do not want to waste enormous amount of energy to “re-perform” evolution at every training



Why natural intelligence is key

- Striving to find theories on how humans (but also other living entities) thinks and acts is an important endeavours for at least two reasons:
 - **efficiency reasons**: we do not want to waste enormous amount of energy to “re-perform” evolution at every training
 - **governance reasons**: if we do *understand* how artificial systems operate, then we as humans will keep controlling them, rather than being controlled by them.

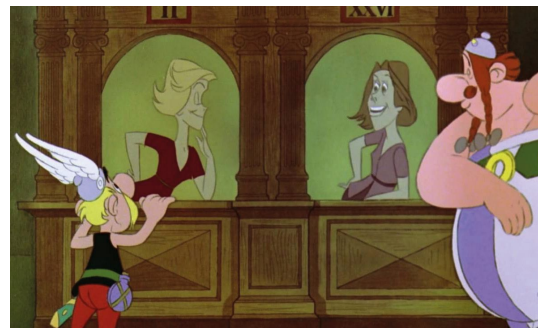
The Law of Requisite Variety: A control system must have as many possible states as the system it wants to control.



Why natural intelligence is key

- Striving to find theories on how humans (but also other living entities) thinks and acts is an important endeavours for at least two reasons:
 - **efficiency reasons**: we do not want to waste enormous amount of energy to “re-perform” evolution at every training
 - **governance reasons**: if we do *understand* how artificial systems operate, then we as humans will keep controlling them, rather than being controlled by them.

NOTE: the same applies to *legal systems*!



So, what would be examples of crucial mechanisms appearing in natural intelligence?

So, what would be examples of crucial mechanisms appearing in natural intelligence?

I will now pinpoint to 2:

surprisal (related to **unexpectedness**)

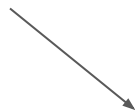
contrast (core descriptive operation)

So, what would be examples of crucial mechanisms appearing in natural intelligence?

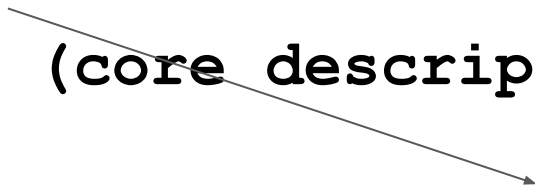
I will now pinpoint to 2:

surprisal (related to **unexpectedness**)

contrast (core descriptive operation)



in the context of bias, fairness



through Simplicity Theory

(I) Surprisal and Unexpectedness

surprising

A few examples of unexpected events (1)

- Remarkable lottery draws

e.g. 11111 is more unexpected than 64178,
even if the lottery is fair.

surprising

A few examples of unexpected events (1)

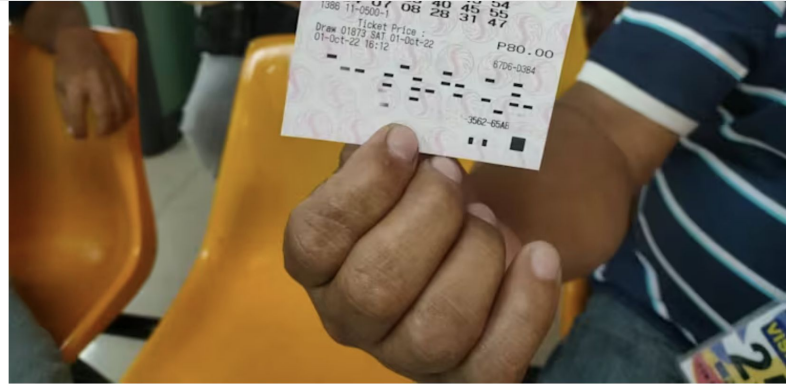
- Remarkable lottery draws

e.g. 11111 is more unexpected than 64178, even if the lottery is fair.



Asia

400 people win Philippines lottery jackpot with 'strange and unusual' number sequence



The winning numbers were multiples of nine. (Photo: Facebook/Philippine Charity Sweepstakes Office)

SINGAPORE: A lottery in the Philippines has sparked calls for a probe after more than 400 people won with a number combination that was multiples of nine.

The numbers were 9, 18, 27, 36, 45 and 54.

Held on Saturday night (Oct 1) by the Philippine Charity Sweepstakes Office (PCSO), the Grand Lotto jackpot of 236 million Philippine peso (US\$4 million) was split among 433 winners, Bloomberg reported.

Philippines' jackpot extraction in 2022:
9, 18, 21, 36, 45, 54

surprising

A few examples of unexpected events (2)

- **Coincidences**

e.g. meeting by chance an old friend (or a known person) abroad is more unexpected than meeting there any random unknown person.



surprising A few examples of unexpected events (2)

- Coincidences

e.g. meeting by chance an old friend (or a known person) abroad is more unexpected than meeting there any random unknown person.



my colleague,
director
master IS at UvA



Do Children Trust AI, and Should They? Designing and Validating a Child-Centred K-AI Trust Scale for Intelligent Systems

Grazia Ragone
University of Bari Aldo Moro
Bari, Italy
grazia.ragone@uniba.it

Judith Good
University of Amsterdam
Amsterdam, Netherlands
j.a.good@uva.nl

Abstract

Most trust metrics for intelligent systems are developed for adults, relying on complex reasoning and language that do not align with the cognitive and emotional capacities of children. This paper introduces a child-centred K-AI Trust Scale for Intelligent Systems, designed to systematically assess and validate trust in AI technologies for children.



me two days ago at
a birthday party
of the daughter of an
old friend of mine in
Puglia

Paolo Buono
University of Bari Aldo Moro
Bari, Italy
paolo.buono@uniba.it

Rosa Lanzilotti
University of Bari Aldo Moro
Bari, Italy
rosa.lanzilotti@uniba.it

1 Introduction

As Artificial Intelligence (AI) technologies become increasingly embedded in everyday human activities, the capacity to systematically assess and validate trust in AI technologies for children is becoming increasingly important.

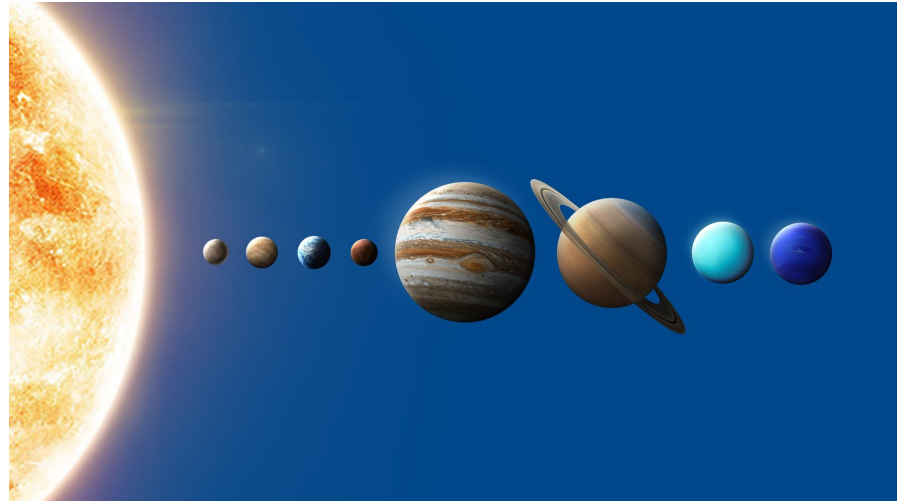
**best paper
CHI 2026**

surprising

A few examples of unexpected events (3)

- **Deterministic yet unexpected events**

e.g. a lunar eclipse or a planet alignment

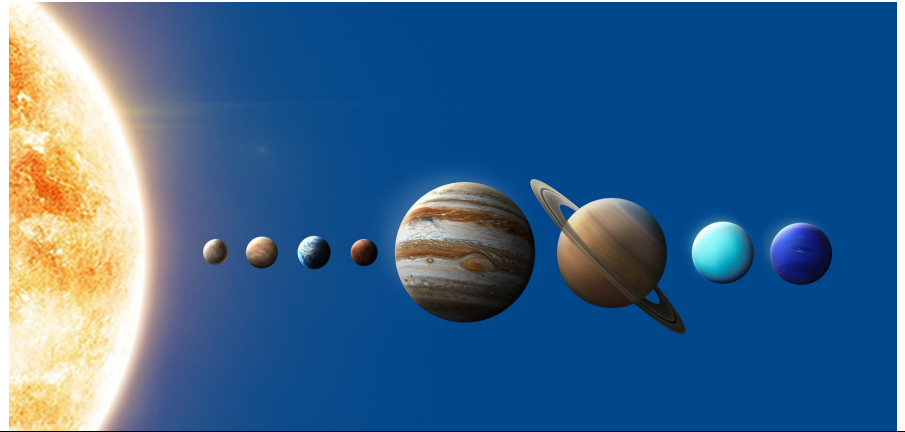


surprising

A few examples of unexpected events (3)

- Deterministic yet unexpected events

e.g. a lunar eclipse or a planet alignment



Neither **probability** nor **description complexity** can explain all these surprisals!

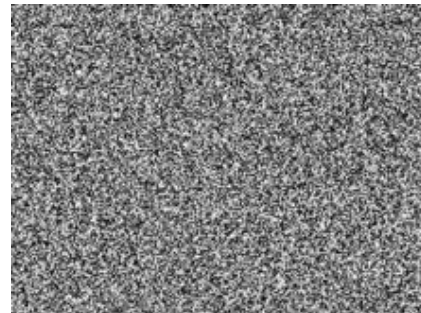
Unexpectedness in Simplicity Theory

- **Simplicity Theory** (ST) is a computational model of cognition introduced by Jean-Louis Dessalles, whose investigation started by observing the “informativity” mismatch.

UNIFORM
NOISE SOURCE
maximally informative



but this mismatches
relevance judgments
given by humans



SO WHAT?

Unexpectedness in Simplicity Theory

- Simplicity Theory stems from the observation that humans are highly susceptible to *complexity drops*, ie. for them

situations are *relevant* if they are *simpler* **to describe** than **to explain**

Unexpectedness in Simplicity Theory

- Simplicity Theory stems from the observation that humans are highly susceptible to *complexity drops*, ie. for them

situations are *relevant* if they are *simpler* **to describe** than **to explain**

*(mentally simulate
their generation in
the world)*

Unexpectedness in Simplicity Theory

- Simplicity Theory stems from the observation that humans are highly susceptible to *complexity drops*, ie. for them

situations are *relevant* if they are *simpler* **to describe** than **to explain**

- Formally, this is captured by the formula of **unexpectedness**, expressed as divergence of Kolmogorov complexities computed on two distinct machines

$$U(s) = C_W(s) - C_D(s)$$

Unexpectedness in Simplicity Theory

- Simplicity Theory stems from the observation that humans are highly susceptible to *complexity drops*, ie. for them

situations are *relevant* if they are *simpler* **to describe** than **to explain**

- Formally, this is captured by the formula of **unexpectedness**, expressed as divergence of Kolmogorov complexities computed on two distinct machines

$$U(s) = C_W(s) - C_D(s)$$

causal complexity
via world machine

description complexity
via description machine

C_W
world $\rightarrow$ situation

C_D
situation $\leftarrow$ mind

Unexpectedness in Simplicity

- Simplicity Theory stems from the observation that humans are susceptible to *complexity drops*, ie. for them

situations are *relevant* if they are *simpler* to describe than

- Formally, this is captured by the formula of **unexpectedness** as the divergence of Kolmogorov complexities computed on two

$$U(s) = C_W(s) - C_D(s)$$

causal complexity
via world machine

description complexity
via description machine

C_W
world $\rightarrow$ situation

C_D
situation $\leftarrow$ mind

*“Dolphin rock”
in Sardinia*



Unexpectedness in Simplicity

- Simplicity Theory stems from the observation that humans are susceptible to *complexity drops*, ie. for them **situations** are *relevant* if they are *simpler to describe* than the actual situation.
- Formally, this is captured by the formula of **unexpectedness** as the divergence of Kolmogorov complexities computed on two different machines:

$$U(s) = C_W(s) - C_D(s)$$

causal complexity
via world machine

description complexity
via description machine

C_W
world $\rightarrow$ situation

C_D
situation $\leftarrow$ mind

*“Dolphin rock”
in Sardinia*

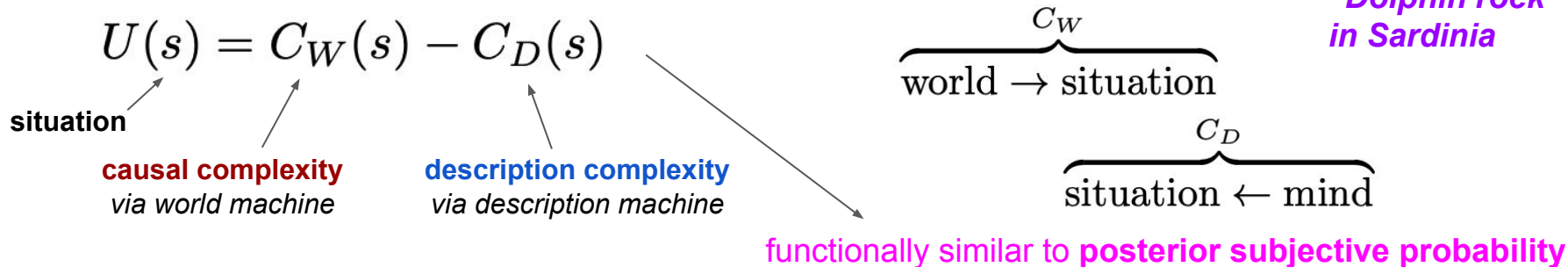


Unexpectedness in Simplicity

- Simplicity Theory stems from the observation that humans are susceptible to *complexity drops*, ie. for them

situations are *relevant* if they are *simpler* to describe than

- Formally, this is captured by the formula of **unexpectedness** as the divergence of Kolmogorov complexities computed on two



But why this formula works?

Three Conjectures on Unexpectedness

Giovanni Sileno¹ and Jean-Louis Dessalles²

¹ University of Amsterdam, Amsterdam, the Netherlands

² Télécom Paris, Institut Polytechnique de Paris, Paris, France
g.sileno@uva.nl, dessalles@telecom-paris.fr

Abstract. Unexpectedness is a central concept in Simplicity Theory, a theory of cognition relating various inferential processes to the computation of Kolmogorov complexities, rather than probabilities. Its predictive power has been confirmed by several experiments with human subjects, yet its theoretical basis remains largely unexplored: *why* does it work? This paper lays the groundwork for three theoretical conjectures. First, unexpectedness can be seen as a generalization of Bayes' rule. Second, the frequentist core of unexpectedness can be connected to the function of tracking ergodic properties of the world. Third, unexpectedness can be seen as constituent of various measures of divergence between the entropy of the world (environment) and the variety of the observer (system). The resulting framework hints to research directions that go beyond the division between probabilistic and logical approaches, potentially bringing new insights into the extraction of causal relations, and into the role of descriptive mechanisms in learning.

<https://arxiv.org/pdf/2311.08768>

First conjecture on unexpectedness

Bayes' rule is a specific instantiation of a more general template captured in ST by Unexpectedness

Reminder: Bayes' rule

- From the definition of conditional probability:

$$p(O \cap M) = p(M|O) \cdot p(O) = p(M) \cdot p(O|M)$$

we can obtain the formula of Bayes simply:

$$p(M|O) = \frac{p(M \cap O)}{p(O)} = \frac{p(O|M) \cdot p(M)}{p(O)}$$

model

observation



often informally rewritten as:

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

First conjecture on unexpectedness

- Bayes' rule is a specific instantiation of ST's Unexpectedness that:
 - makes a **candidate "cause" explicit** and does **not select** automatically **the best one**
 - takes a **frequentist-like** approach for **encoding observables**.

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

$$U(s) = \min_c \underbrace{[C_W(c * s) - C_D(s)]}_{\text{posterior}} = \min_c \underbrace{[C_W(s||c)]}_{\text{likelihood}} + \underbrace{C_W(c)}_{\text{prior}} - \underbrace{C_D(s)}_{\text{evidence}}$$

Example

- Consider the estimation of the likelihood that the wall changes colour if I close the door:



Example

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$$C_D \approx 0$$

because these
elements are just
in front of me

$$C_W \gg 0$$

because this never
occurred



Example

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$$C_D \approx 0$$

because these
elements are just
in front of me

$$C_W \gg 0$$

because this never
occurred



$$U \approx C_W \gg 0$$

it is implausible
(if it occurred)

$$C_W^U = U + C_D \gg 0$$

it is improbable
(to occur)



Example

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$$C_D \approx 0$$

because these
elements are just
in front of me

$$C_W \gg 0$$

because this never
occurred



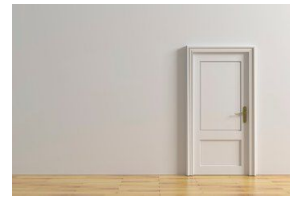
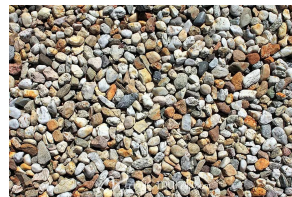
$$U \approx C_W \gg 0$$

it is implausible
(if it occurred)

$$C_W^U = U + C_D \gg 0$$

it is improbable
(to occur)

- The likelihood that a stone in the world moves if I close the door:



Example

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$$C_D \approx 0$$

because these
elements are just
in front of me

$$C_W \gg 0$$

because this never
occurred



$$U \approx C_W \gg 0$$

it is implausible
(if it occurred)

$$C_W^U = U + C_D \gg 0$$

it is improbable
(to occur)

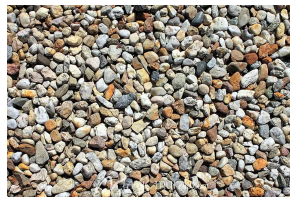
- The likelihood that a stone in the world moves if I close the door:

$$C_D \gg 0$$

because I need to
specify of which
stone I am talking

$$C_W \gg 0$$

because this never
occurred



Example

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$$C_D \approx 0$$

because these
elements are just
in front of me

$$C_W \gg 0$$

because this never
occurred



$$U \approx C_W \gg 0$$

it is implausible
(if it occurred)

$$C_W^U = U + C_D \gg 0$$

it is improbable
(to occur)

- The likelihood that a stone in the world moves if I close the door:

$$C_D \gg 0$$

because I need to
specify of which
stone I am talking

$$C_W \gg 0$$

because this never
occurred

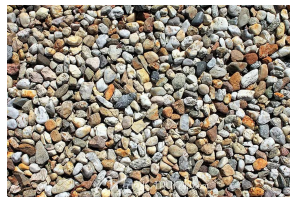


$$U \approx 0$$

it is plausible
(if it occurred)

$$C_W^U = U + C_D \gg 0$$

it is improbable
(to occur)



Example

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$$C_D \approx 0$$

because these
elements are just
in front of me

$$C_W \gg 0$$

because this never
occurred



$$U \approx C_W \gg 0$$

it is implausible
(if it occurred)

$$C_W^U = U + C_D \gg 0$$

it is improbable
(to occur)

- The likelihood that a stone in the world moves if I close the door:

$$C_D \gg 0$$

because I need to
specify of which
stone I am talking

$$C_W \gg 0$$

because this never
occurred



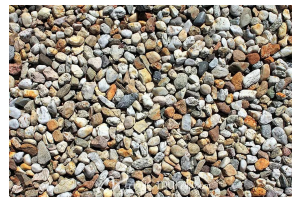
$$U \approx 0$$

it is plausible
(if it occurred)

$$C_W^U = U + C_D \gg 0$$

it is improbable
(to occur)

*NOTE: If the stone e.g. is in the room or
was already described, we return to the first case!*



Example

- What do you deem more plausible to occur, meeting a pink elephant or a black elephant?



Example

- What do you deem more plausible to occur, meeting a pink elephant or a black elephant?



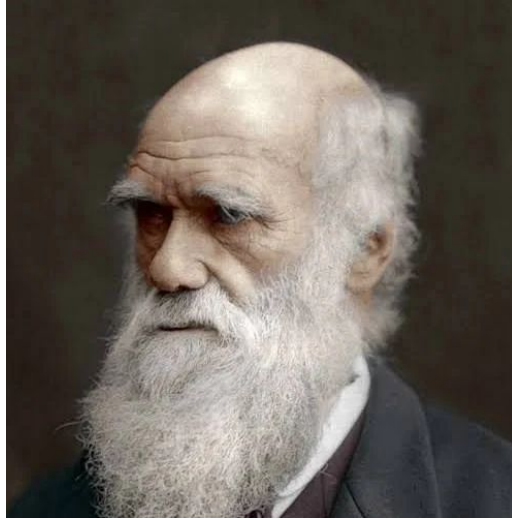
both are never seen, but one is descriptively more accessible from the usual case.

(II) Contrast (and Fairness)

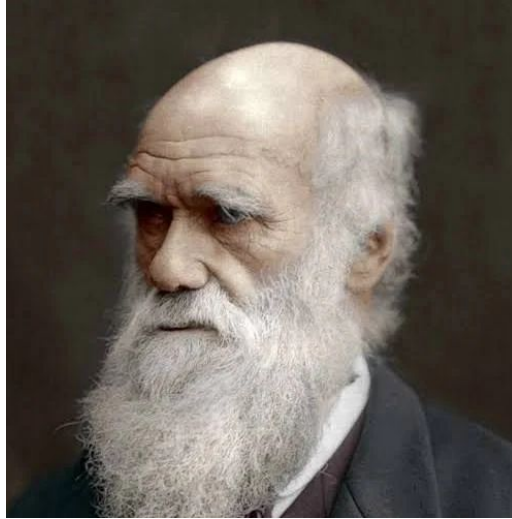
Fundamental principle of perception:

We see differences against common grounds

Fundamental principle of perception:
We see differences against common grounds



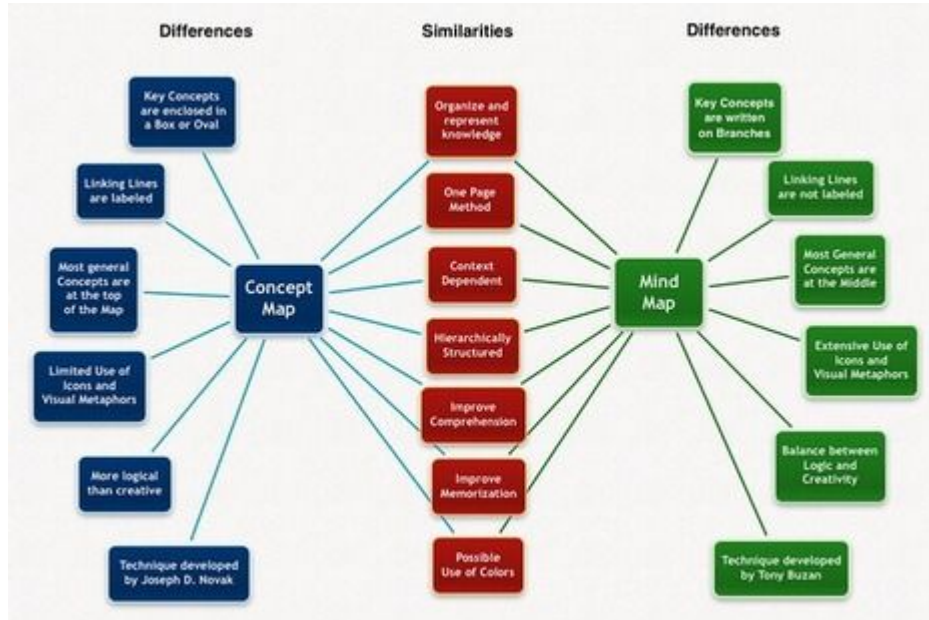
**Fundamental principle of perception:
We see differences against common grounds**



finding what is in common is an
operation of compression

Fundamental principle of ~~perception~~ conceptualization

We see differences against common grounds

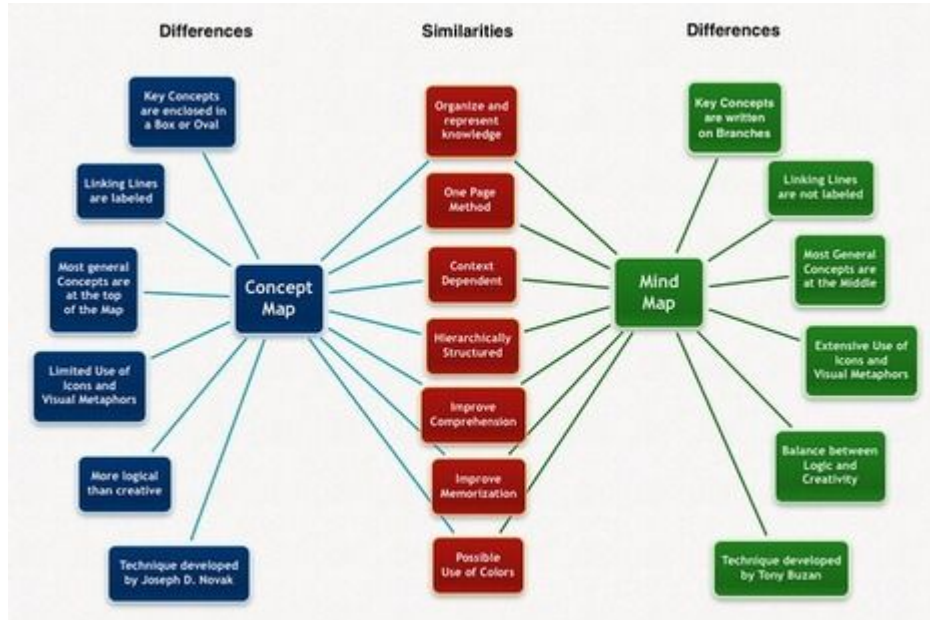


finding what is in common is an operation of compression

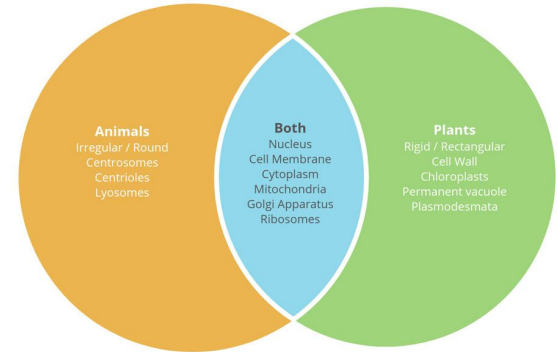
comparing and **contrasting** is at the basis of *knowledge construction*

Fundamental principle of ~~perception~~ conceptualization

We see differences against common grounds



finding what is in common is an operation of compression



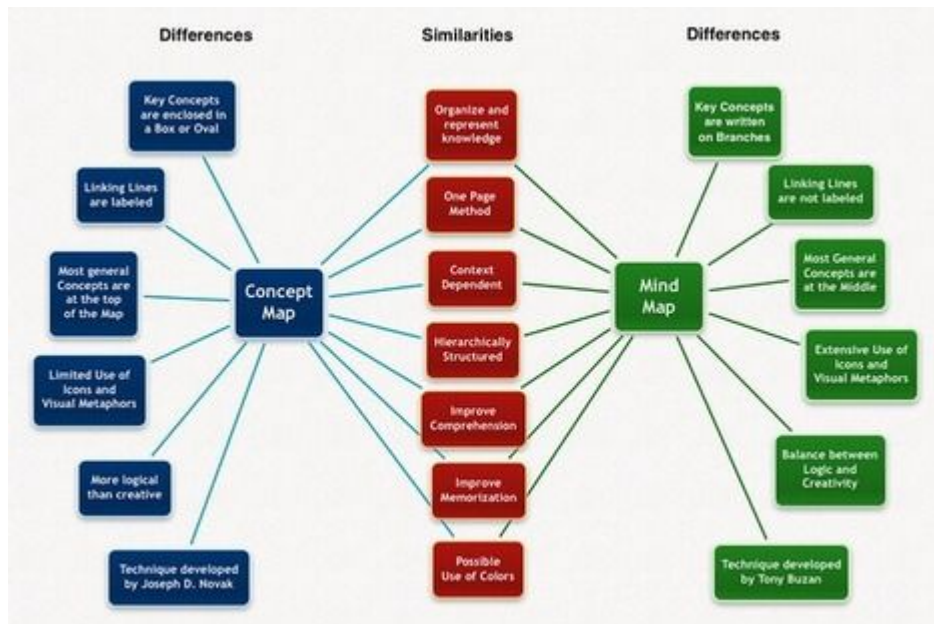
eg. to construct taxonomies

comparing and **contrasting** is at the basis of *knowledge construction*

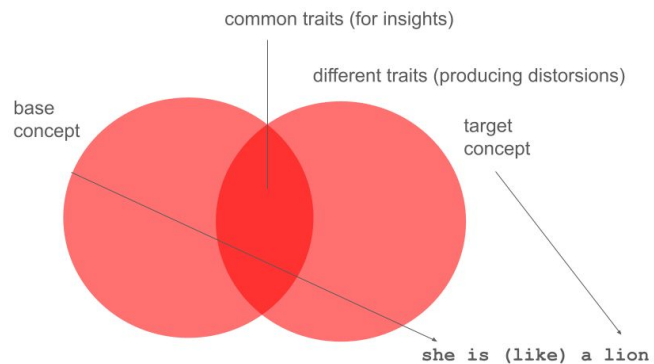
conceptualization

Fundamental principle of ~~perception~~:

We see differences against common grounds



finding what is in common is an operation of compression

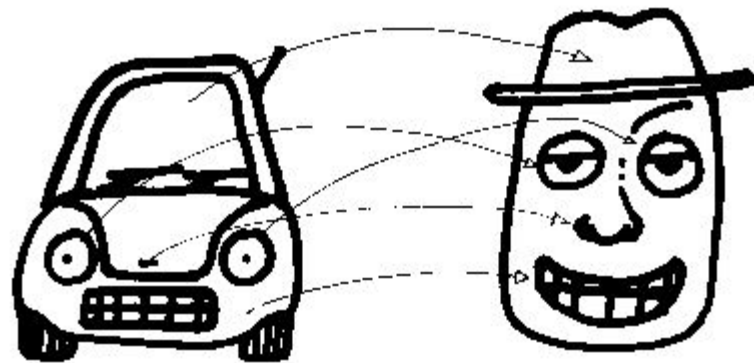


... but also metaphors!

comparing and **contrasting** is at the basis of *knowledge construction*

Analogical inference

- Analogies, in all forms, are fundamental to human cognition
- Traditional study subject in psychology and AI
- Contemporary studies in ML



visual analogy

$\text{Paris} \approx \text{France} + \text{Berlin} - \text{Germany}$

analogical inference as operations on word vectors

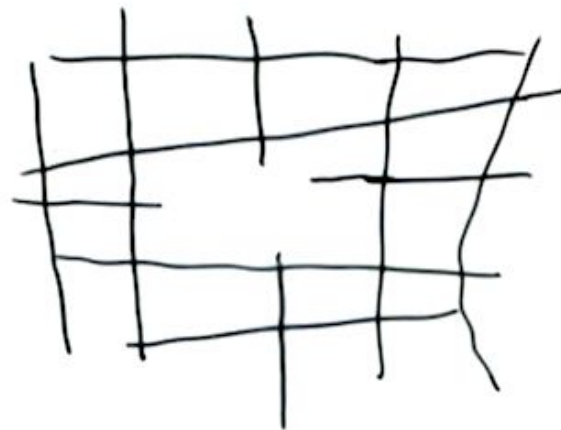
Analogical inference on symbolic analogies

- Hofstadter suggested to study phenomena about analogies on a limited domain whose primitives are meaningless symbols

ABC : ABD :: IJK : ?

string-based proportional inference

- Proposed solutions:
 - Copycat and Metacat

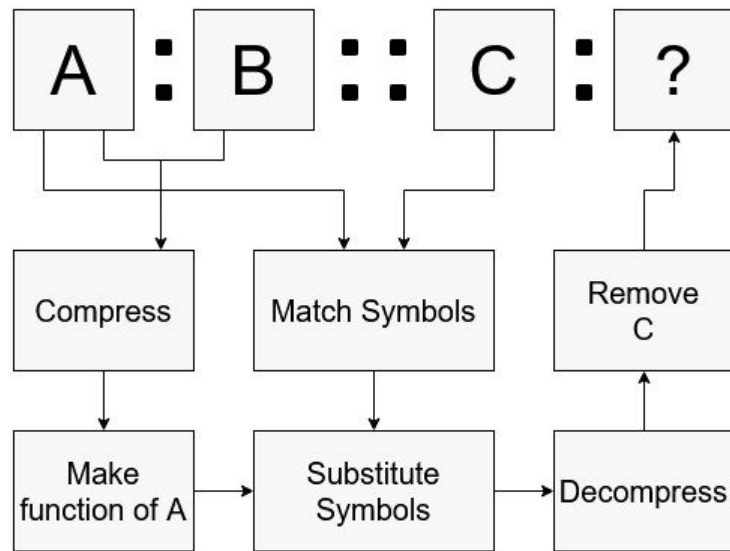


"Analogy is the Interstate Freeway System of cognition"

Integrating compression and contrast?

- Analogy test given in input:

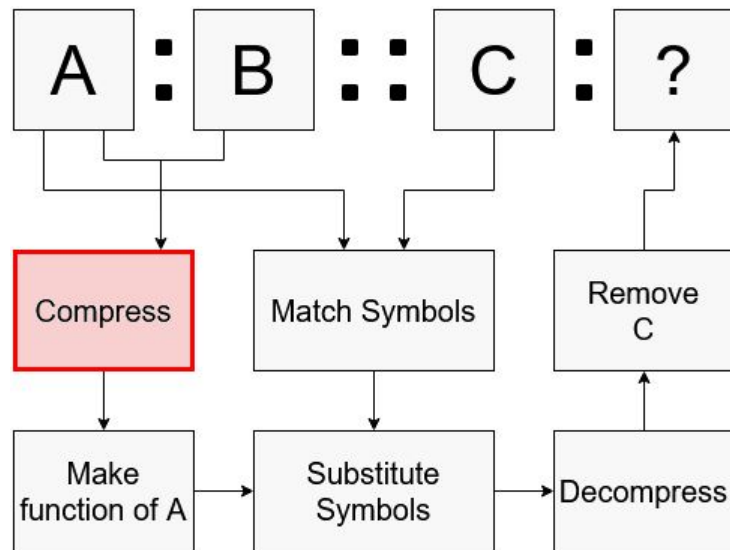
1. ABC:ABD::IJK:?



Integrating compression and contrast?

- Applying SIT, via the PISA algorithm [VanderHelm2015]

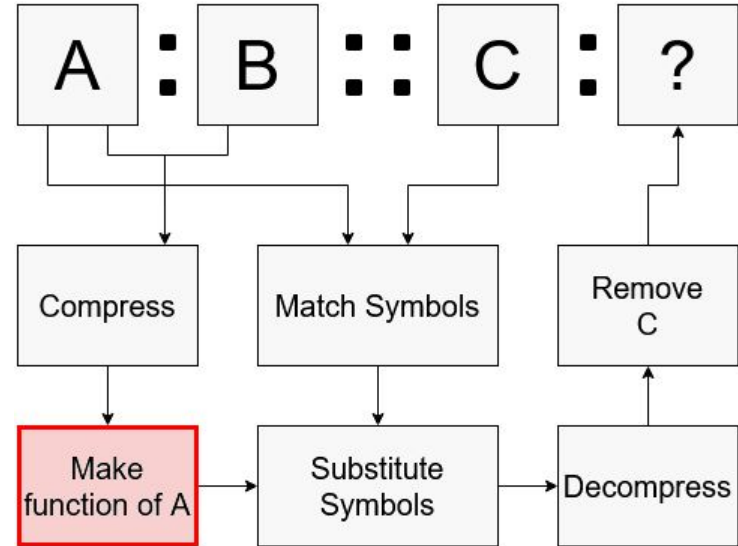
- 1. ABC:ABD::IJK:?**
- 2. <(AB)>/<(C)(D)>**



Integrating compression and contrast?

- We introduce a metric/distance between symbols.
- *Problem*: which symbol should the distance be calculated from?

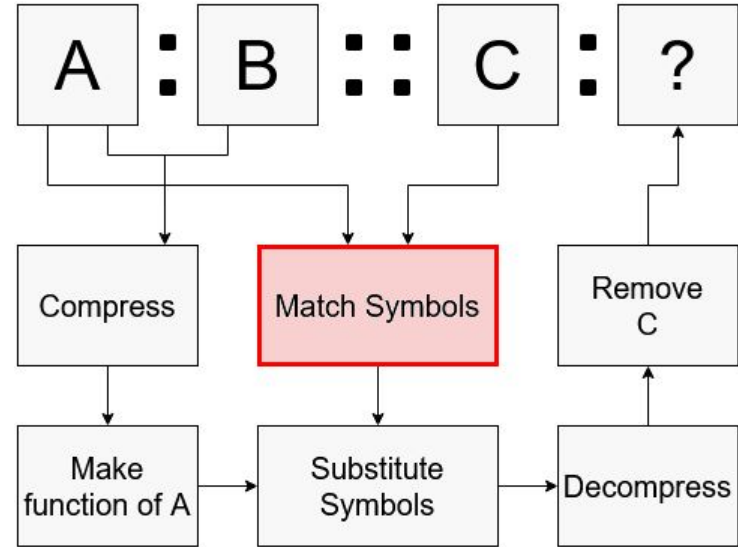
1. **ABC:ABD::IJK:?**
2. **<(AB)>/<(C)(D)>**
3. **<(AB)>/<(C)(\$+1)>**



Integrating compression and contrast?

- We formulate a mapping between symbols in A and C
- *Problem:* Difficult when A and C are of different lengths

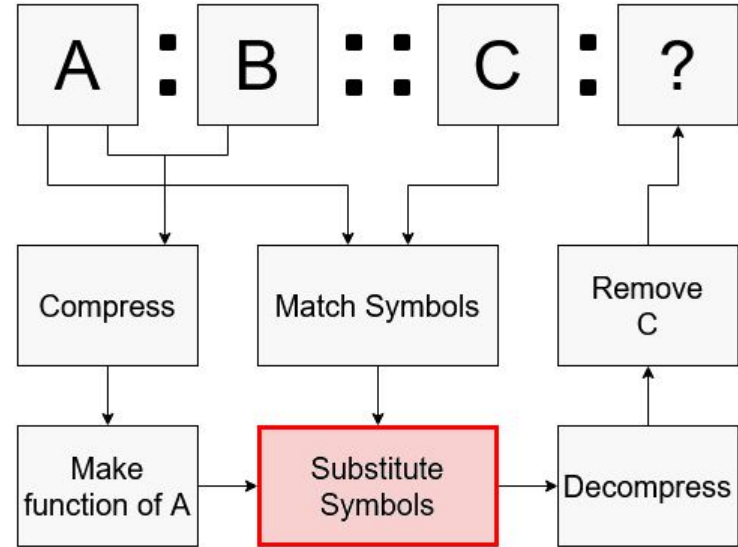
1. **ABC:ABD::IJK:?**
2. **<(AB)>/<(C)(D)>**
3. **<(AB)>/<(C)(\$+1)>**
4. **A:I, B:J, C:K**



Integrating compression and contrast?

- We create a new SIT code with the mapping

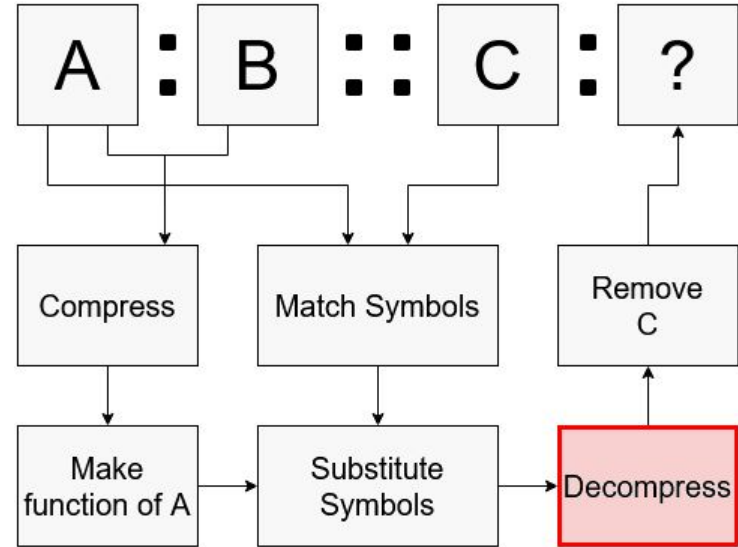
1. **ABC:ABD::IJK:?**
2. **<(AB)>/<(C)(D)>**
3. **<(AB)>/<(C)(\$+1)>**
4. **A:I, B:J, C:K**
5. **<(IJ)>/<(K)(\$+1)>**
6. **<(IJ)>/<(K)(L)>**



Integrating compression and contrast?

- We decompress the SIT encoding

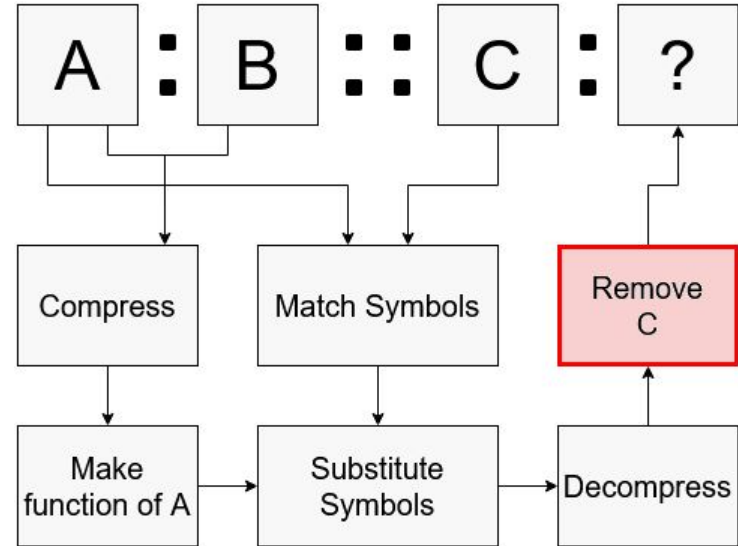
1. **ABC:ABD::IJK:?**
2. **<(AB)>/<(C)(D)>**
3. **<(AB)>/<(C)(\$+1)>**
4. **A:I, B:J, C:K**
5. **<(IJ)>/<(K)(\$+1)>**
6. **<(IJ)>/<(K)(L)>**
7. **IJKIJL**



Integrating compression and contrast?

- If **C** is not in the decompressed code, the answer is dismissed

1. **ABC:ABD::IJK:?**
2. **<(AB)>/<(C)(D)>**
3. **<(AB)>/<(C)(\$+1)>**
4. **A:I, B:J, C:K**
5. **<(IJ)>/<(K)(\$+1)>**
6. **<(IJ)>/<(K)(L)>**
7. **IJKIJL**
8. **IJL**



Results!

Given problem	Solutions	Selected by	P_s	P_M
ABA:ACA::ADA:?	AEA	97.1%	1	1
	AFA	2.9%	-	-
ABAC:ADAE::BACA:?	DAEA	60%	2	-
	BCCC	28.6%	21	-
AE:BD::CC:?	DB	68.5%	3	1
	CC	17.1%	-	2
ABBB:AAAB::IIJJ:?	IIJJ	57.1%	1	-
	JJII	14,3%	-	-
ABC:CBA::MLKJI:?	IJKLM	88.6%	1	1
	-	-	-	-
ABCB:ABCB::Q:?	Q	100.0%	1	-
	-	-	-	-
ABC:BAC::IJKL:?	JKLM	54.3%	-	-
	KIJL	14.3%	2	-
ABACA:BC::BACAD:?	AA	57.1%	1	-
	BCD	31.4%	-	-
AB:ABC::IJKL:?	IJKLM	85.7%	1	1
	IJKLMNOP	11.4%	-	-
ABC:ABBACCC::FED:?	FEEFDDD	91.4%	2	1
	-	-	-	-

ABC:BBC::IKM:?	JKM	57.1%	7	-
	KKM	37.1%	2	-
ABAC:ACAB::DEFG:?	DGFE	68.6%	2	-
	FGDE	14.3%	1	-
ABC:ABD::CBA:?	DBA	51.4%	1	2
	CBB	45.7%	2	1
ABAC:ADAE::FBFC:?	FDFE	94.3%	1	-
	FDFA	2.9%	-	-
ABCD:CDAB::IJKLMNOP:?	LMNIJK	80.0%	-	-
	-	-	-	-
ABC:AAABBBCCC::ABCD:?	AAABBBCCCDDD	74.3%	1	1
	AAAABBBBCCCCDDDD	17.1%	-	-
ABC:ABBCCC::ABCD:?	ABBCCCDDDD	85.7%	-	-
	ABBCCCDDD	8.6%	1	-
ABBCCC:DDDEEF::AAABBC:?	DEEEFF	77.1%	1	-
	DCCDDF	8.6%	-	-
A:AA::AAA:?	AAAAAA	62.8%	1	-
	AAAA	25.7%	2	1
ABBA:BAAB::IJKL:?	JILK	71.4%	-	-
	IJJM	11.4%	5	-

generating most common human answer: our solver 16/20 (80%), Metacat 7/20 (35%)

best answer in the top 2 human answers: 13/20 (65%) vs 8/20 (40%)

best answer matching most common human answer: 10/20 (50%) vs 6/20 (30%)

What about LLMs?

“[...] This work provides evidence that, despite previously reported successes of LLMs on zero-shot analogical reasoning, these models often lack the robustness of zero-shot human analogy-making, exhibiting brittleness on most of the variations we teste [...]”

Evaluating the Robustness of Analogical Reasoning in Large Language Models

Martha Lewis^{1,2} and Melanie Mitchell²

¹ILLC, University of Amsterdam, Amsterdam, The Netherlands

²Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

m.a.f.lewis@uva.nl, mm@santafe.edu

Abstract

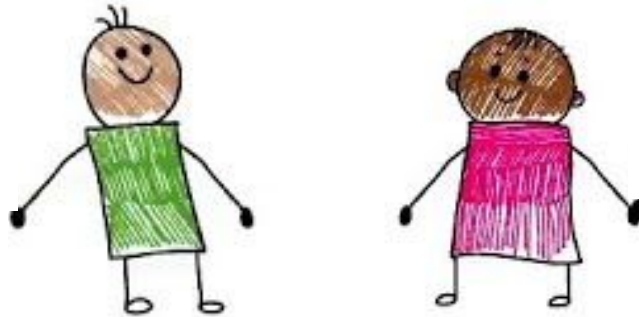
Large language models (LLMs) have performed well on several reasoning benchmarks, including ones that test analogical reasoning abilities. However, there is debate on the extent to which they are performing general abstract reasoning versus employing shortcuts or other non-robust processes, such as ones that overly rely on similarity to what has been seen in their training data. Here we investigate the robustness of analogy-making abilities previously claimed for LLMs on three of four domains studied by [35]: letter-string analogies, digit matrices, and story analogies. For each of these domains we test humans and GPT models on robustness to variants of the original analogy problems—versions that test the same abstract reasoning abilities but that are likely dissimilar from tasks in the pre-training data. The performance of a system that uses robust abstract reasoning should not decline substantially on these variants.

On simple letter-string analogies, we find that while the performance of humans remains high for two types of variants we tested, the GPT models’ performance declines sharply. This pattern is less pronounced as the complexity of these analogy problems is increased, as both humans and GPT models perform poorly on both the original and variant problems requiring more complex analogies. On digit-matrix problems, we find a similar pattern but only on one out of the two types of variants we tested. Lastly, we assess the robustness of humans and GPT models on story-based analogy problems, finding that, unlike humans, the performance of GPT models are susceptible to answer-order effects, and that GPT models also may be more sensitive than humans to paraphrasing.

Approaches to Fairness

There exists two main families of approaches to fairness:

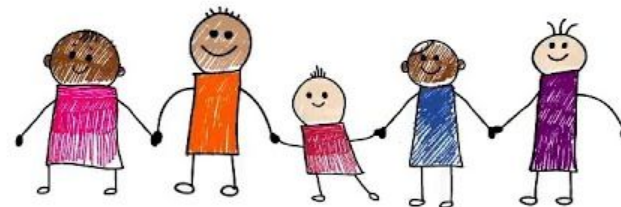
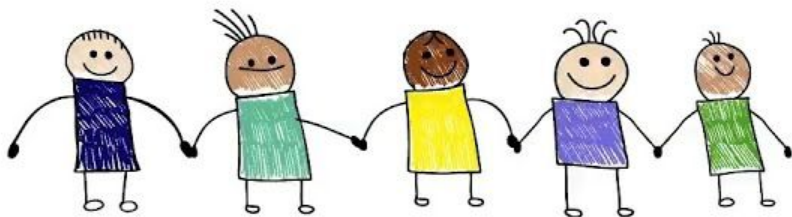
- **Individual fairness** compares treatment amongst individuals, based upon the principle that *“similar should be treated similarly”*



Approaches to Fairness

There exists two main families of approaches to fairness:

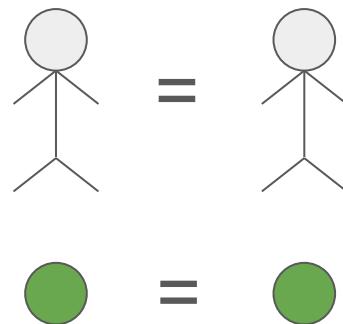
- **Individual fairness** compares treatment amongst individuals, based upon the principle that “*similar should be treated similarly*”
- **Group fairness** compares treatment amongst groups (rather sub-groups of a given population) typically identified by sensitive attributes (e.g. *gender*, *ethnicity*, *age*, etc.), e.g. “women should be treated as men”.



Individual fairness

- In the traditional formulation:

We are similar $\rightarrow$ we are treated similarly



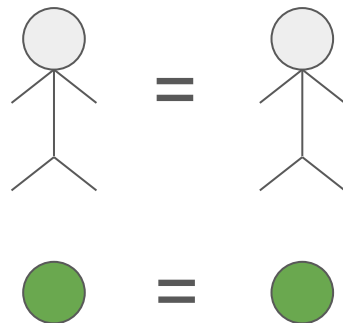
Individual fairness

- In the traditional formulation:

We are similar $\rightarrow$ we are treated similarly

- We can write it in its contrapositive form:

We are not treated similarly $\rightarrow$ we are not similar



Individual fairness

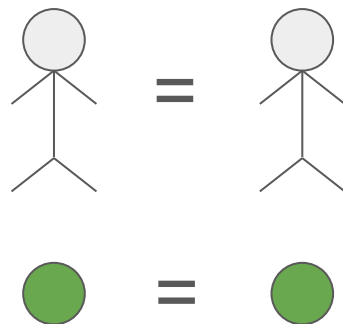
- In the traditional formulation:

We are similar $\rightarrow$ we are treated similarly

- We can write it in its contrapositive form:

We are not treated similarly $\rightarrow$ we are not similar

➡ The contrapositive makes clear that the decision has consequences on determining the similarity metric. In other words, a classifier presupposes classes of equivalence applicable to the dataset.



Individual fairness

- In the traditional formulation:

We are similar $\rightarrow$ we are treated similarly

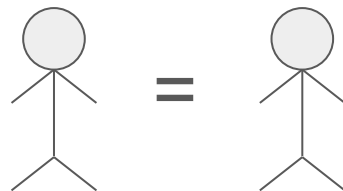
- We can write it in its contrapositive form:

We are not treated similarly $\rightarrow$ we are not similar

 The contrapositive makes clear that the decision has consequences on determining the similarity metric. In other words, a classifier presupposes classes of equivalence applicable to the dataset.

- **Unfairness** is defined as the negation of fairness:

We are similar, yet we are not treated similarly.

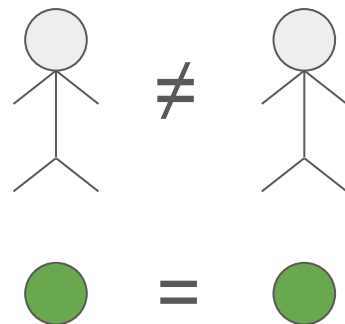


Group fairness

- Group fairness has a different formulation than individual fairness. Here, we have sensitive attributes (gender, etc.) that should not be used to reach a conclusion.

Group fairness

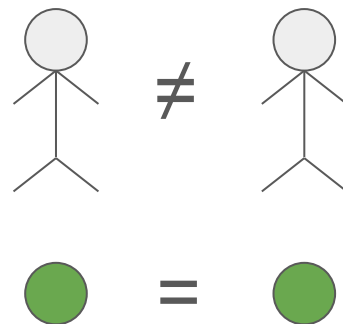
- Group fairness has a different formulation than individual fairness. Here, we have sensitive attributes (gender, etc.) that should not be used to reach a conclusion.
- Therefore, even if we are different with respect to a certain sensitive feature, we should be treated the same, because this feature has to be treated as ***irrelevant*** in the decision process:



Even if we are different for a sensitive feature, we are treated the same.

Group fairness

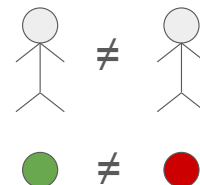
- Group fairness has a different formulation than individual fairness. Here, we have sensitive attributes (gender, etc.) that should not be used to reach a conclusion.



- Therefore, even if we are different with respect to a certain sensitive feature, we should be treated the same, because this feature has to be treated as **irrelevant** in the decision process:

Even if we are different for a sensitive feature, we are treated the same.

- Unfairness** is defined as the negation of the above rule:



We are different for a feature, AND we are treated differently.

Individual vs Group fairness

- If individual fairness makes reference to variation on input having effects on output, group fairness appeals to absence of effects on output
- The keyword unifying the two views is “relevant”:
 - Individual: *We are RELEVANTLY similar, yet we are not treated similarly.*
 - Group: *We are IRRELEVANTLY different, yet we are not treated similarly.*

Individual vs Group fairness

- If individual fairness makes reference to variation on input having effects on output, group fairness appeals to absence of effects on output.
- The keyword unifying the two views is “relevant”:
 - Individual: *We are RELEVANTLY similar, yet we are not treated similarly.*
 - Group: *We are IRRELEVANTLY different, yet we are not treated similarly.*



Group-fairness seems easier to apply when we prescribe explicitly irrelevant features!

Individual vs Group fairness

- If individual fairness makes reference to variation on input having effects on output, group fairness appeals to absence of effects on output.
- The keyword unifying the two views is “relevant”:
 - Individual: *We are RELEVANTLY similar, yet we are not treated similarly.*
 - Group: *We are IRRELEVANTLY different, yet we are not treated similarly.*



Group-fairness seems easier to apply when we prescribe explicitly irrelevant features!

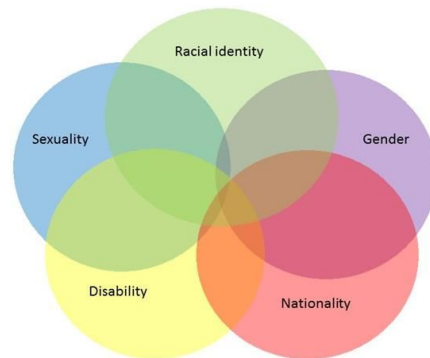


Yet, irrelevance depends on context. For instance gender is not necessarily problematic when suggesting dresses on a web-shop.

Algorithmic fairness: Parity between whom?

- statistical measures relies on identifying groups (→ group fairness)
- but which groups? generally protected groups groups defined by sensitive dimensions (eg. sex or gender, race or ethnicity)

➔ problem of *intersectionality*



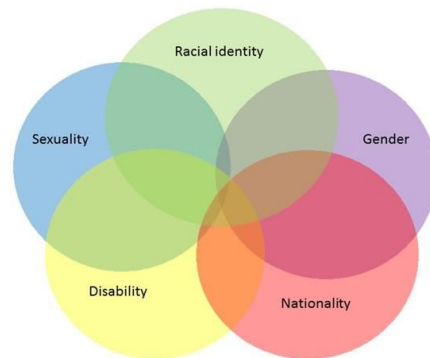
Algorithmic fairness: Parity between whom?

- statistical measures relies on identifying groups (→ group fairness)
- but which groups? generally protected groups groups defined by sensitive dimensions (eg. sex or gender, race or ethnicity)

 problem of *intersectionality*

Example:

- women vs men
- black women vs black men
- low-class black women vs low-class black men

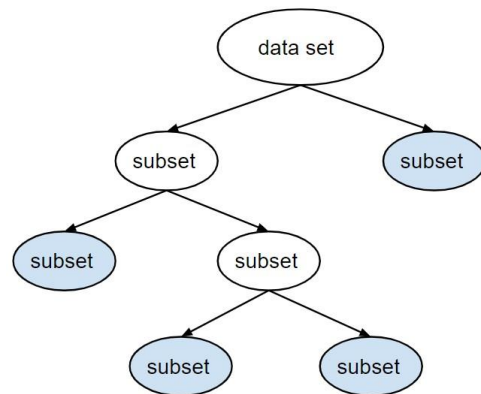


Which is the right granularity? (problem with affinities to Simpson's paradox!)

Algorithmic fairness: Parity between whom?

With a Master IS student we worked on inverting the problem:

- identifying **descriptively relevant subgroups** (*profiles*) in the data space, based on an iterative application of clustering algorithms
- check how the algorithm works on these subgroups in general, and w.r.t. protected features.



Some results on German credit dataset

– Note: here ML is used to unveil
historical (human) bias!

demographic disparity
against women
with low resources
when asking credit
for cars

(slight)
demographic disparity
against women
with low resources
when asking credit
for furniture/equipment

Profile	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Sample
0	26	male	2	rent	moderate	unknown	3577	9	car	859
1	37	male	2	own	unknown	unknown	7409	36	business	868
2	39	male	3	own	little	unknown	6458	18	car	106
3	26	male	2	own	little	little	4370	42	radio/TV	639
4	31	male	2	own	quite rich	unknown	3430	24	radio/TV	19
5	38	female	2	own	unknown	unknown	1240	12	radio/TV	135
6	43	male	1	own	little	little	1344	12	car	929
7	36	male	2	rent	little	little	2799	9	car	586
8	39	male	2	own	little	little	2922	30	radio/TV	239
9	31	male	2	own	little	moderate	1935	24	business	169
10	33	female	2	own	little	little	1131	18	furniture/equipment	166
11	26	male	1	own	little	moderate	625	12	radio/TV	220
12	23	male	2	own	unknown	moderate	1444	15	radio/TV	632
13	42	male	2	own	little	little	4153	18	furniture/equipment	899
14	29	male	2	own	unknown	unknown	3556	15	car	962
15	37	female	2	own	little	moderate	3612	18	furniture/equipment	537
16	27	female	2	own	little	little	2389	18	radio/TV	866
17	26	female	2	rent	little	unknown	1388	9	furniture/equipment	582
18	29	male	2	own	little	unknown	2743	28	radio/TV	426
19	53	male	2	free	little	little	4870	24	car	4
20	36	male	2	own	little	little	1721	15	car	461
21	38	male	2	own	little	unknown	804	12	radio/TV	997
22	29	male	2	own	little	moderate	1103	12	radio/TV	696
23	43	male	2	own	unknown	unknown	2197	24	car	406
24	27	male	2	own	little	little	3552	24	furniture/equipment	558
25	30	male	2	own	little	moderate	1056	18	car	580
26	24	female	2	own	little	moderate	2150	30	car	252
27	34	male	2	own	little	unknown	2750	12	furniture/equipment	452
28	24	female	2	rent	little	little	2124	18	furniture/equipment	761
29	34	male	2	own	little	moderate	5800	30	car	893
30	34	female	2	own	little	unknown	1493	12	radio/TV	638
31	30	female	2	own	little	unknown	1055	18	car	161
32	35	male	2	own	little	unknown	2346	24	car	654
33	35	male	2	own	unknown	unknown	1979	15	radio/TV	625

Some results on German credit dataset

– Note: here ML is used to unveil
historical (human) bias!

demographic disparity
against women
with low resources
when asking credit
for cars

(slight)
demographic disparity
against women
with low resources
when asking credit

Profile	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Sample
0	26	male	2	rent	moderate	unknown	3577	9	car	859
1	37	male	2	own	unknown	unknown	7409	36	business	868
2	39	male	3	own	little	unknown	6458	18	car	106
3	26	male	2	own	little	little	4370	42	radio/TV	639
4	31	male	2	own	quite rich	unknown	3430	24	radio/TV	19
5	38	female	2	own	unknown	unknown	1240	12	radio/TV	135
6	43	male	1	own	little	little	1344	12	car	929
7	36	male	2	rent	little	little	2799	9	car	586
8	39	male	2	own	little	little	2922	30	radio/TV	239
9	31	male	2	own	little	moderate	1935	24	business	169
10	33	female	2	own	little	little	1131	18	furniture/equipment	166
11	26	male	1	own	little	moderate	625	12	radio/TV	220
12	23	male	2	own	unknown	moderate	1444	15	radio/TV	632
13	42	male	2	own	little	little	4153	18	furniture/equipment	899
14	29	male	2	own	unknown	unknown	3556	15	car	962
15	37	female	2	own	little	moderate	3612	18	furniture/equipment	537
16	27	female	2	own	little	little	2389	18	radio/TV	866
17	26	female	2	rent	little	unknown	1388	9	furniture/equipment	582
18	29	male	2	own	little	unknown	2743	28	radio/TV	426
19	53	male	2	free	little	little	4870	24	car	4
20	36	male	2	own	little	little	1721	15	car	461
21	38	male	2	own	little	unknown	804	12	radio/TV	997
22	29	male	2	own	little	moderate	1103	12	radio/TV	696
23	43	male	2	own	unknown	unknown	2197	24	car	406
24	27	male	2	own	little	little	3552	24	furniture/equipment	558
25	30	male	2	own	little	moderate	1056	18	car	580
26	24	female	2	own	little	moderate	2150	30	car	252
27	34	male	2	own	little	unknown	2750	12	furniture/equipment	452
28	24	female	2	rent	little	little	2124	18	furniture/equipment	761
29	34	male	2	own	little	moderate	3800	30	car	893
30	34	female	2	own	little	unknown	1493	12	radio/TV	638
31	29	female	2	own	little	unknown	1055	18	car	161

In general, parity at aggregate level does not imply parity at subgroup level!

Conclusions!

- We still miss several important pieces of the puzzle of how (human, and not only) intelligence works.

Conclusions!

- We still miss several important pieces of the puzzle of how (human, and not only) intelligence works.
- Compression is clearly a core component to cognition...
...but it is not the only one.

Conclusions!

- We still miss several important pieces of the puzzle of how (human, and not only) intelligence works.
- Compression is clearly a core component to cognition...
...but it is not the only one.
- If we aim to have *intelligent* and *intelligible* systems, we need to keep a structural alignment with human cognitive mechanisms, such as *simplicity* *biases*, as well as *contrast* and other primitive mental operations (*merge*, *fusion*, ...)!

Conclusions!

- We still miss several important pieces of the puzzle of how (human, and not only) intelligence works.
- Compression is clearly a core component to cognition...
...but it is not the only one.
- If we aim to have *intelligent* and *intelligible* systems, we need to keep a structural alignment with human cognitive mechanisms, such as *simplicity* *biases*, as well as *contrast* and other primitive mental operations (*merge*, *fusion*, ...)!

→ *this path is relevant also because it will plausibly bring to more efficient AI!*

Comprehension as compression?

Università di Salerno

24 July 2026



Giovanni Sileno

g.sileno@uva.nl

University of Amsterdam