Unidade Curricular

“Teoria Algorítmica da Informação”

António José Ribeiro Neves

an@ua.pt

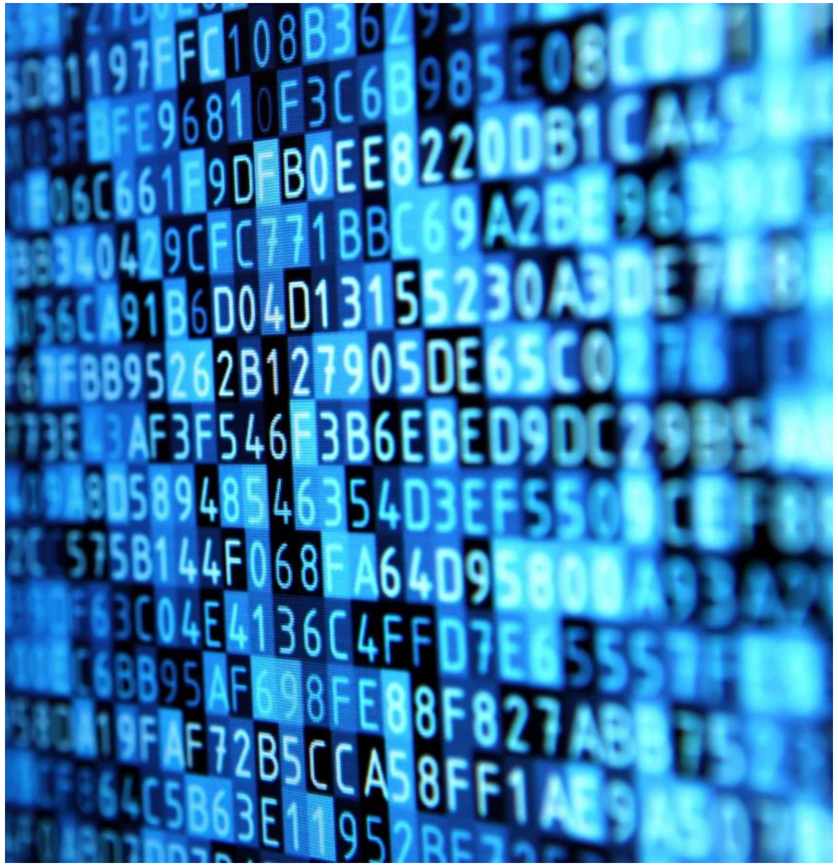
https://www.ua.pt/pt/uc/13643

Outline

- Guiding questions
- Data
- Information
- Challenges
- Compression
- Computation

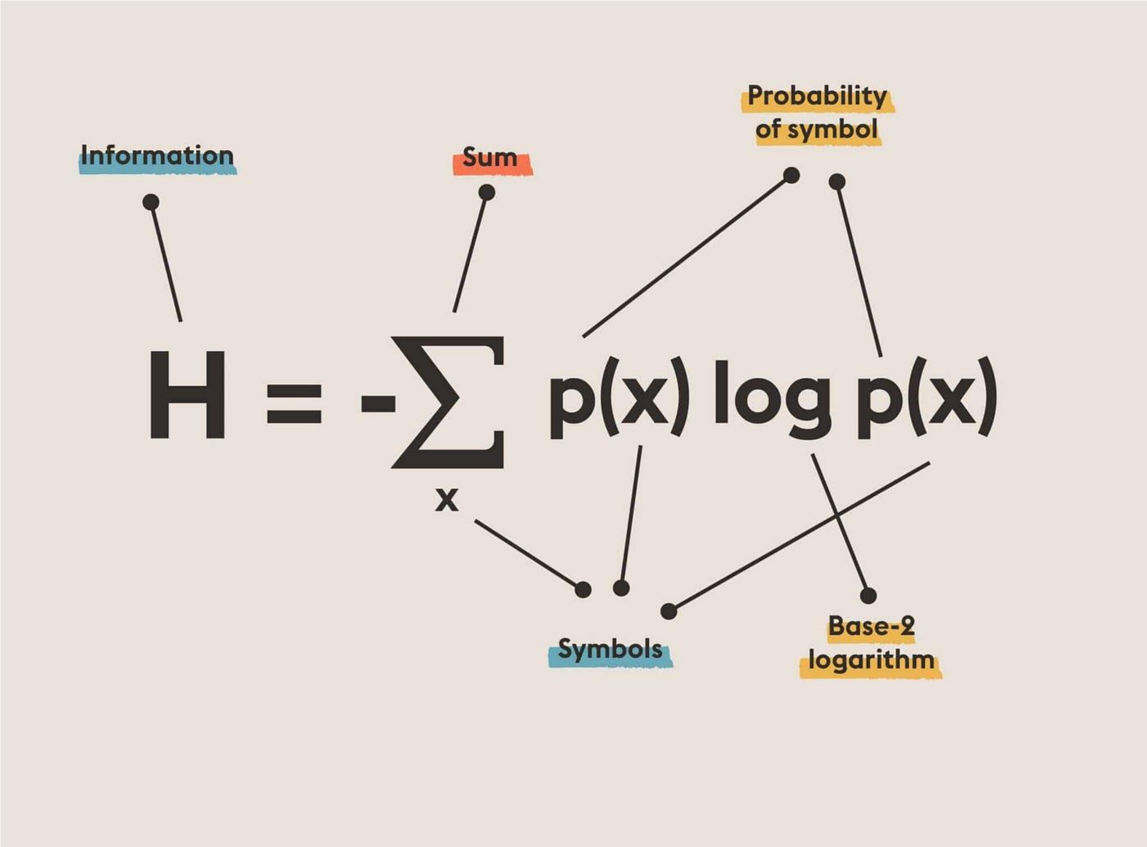
Data

- Data is any representation of facts, figures, or instructions that can be stored, communicated, or processed.
- • **Forms of Data:**
  - Text
  - Images
  - Audio
  - Video
- Data is the raw material we process to extract **information**.
- **How to Measure Data:**
  - **Bits and Bytes:** Basic units of data storage (1 byte = 8 bits).
  - **Symbols:** In a message, data can also be quantified by the number of symbols or characters.
- **Example:**
  - A 1000-character text document in ASCII requires 1000 bytes (1 byte per character).



Information

- Information is a measure of the reduction of uncertainty when an event or message is revealed.
- Information is the meaningful content extracted from data.
- **Example:**
  - A coin flip reveals 1 bit of information (if it’s a fair coin).
- Entropy
  - More uncertain outcomes provide more information when revealed.
  - A measure of the uncertainty or average amount of information in a data source.
- Example:
  - Fair coin toss: H = 1 bit.
  - Biased coin (90% heads): H < 1 bit.
- Higher entropy = more information (less predictable).



Measuring Information: Combinatorial Approach

- Measures information based on the number of possible combinations or arrangements of symbols, objects, or events.
- Information is related to the number of distinct possibilities in a given system.
- The more possible combinations, the more information is needed to describe a specific arrangement.
- Example:
  - If we want to know the arrangement of 5 cards drawn from a deck of 52, the number of possible combinations is large, which implies a significant amount of information is needed to specify the exact combination.
- For a set of size N , the amount of information is proportional to log2(N) .
- Applications in cryptography, data transmission, and permutational problems where information is linked to the number of possible configurations or sequences.

Measuring Information: Probabilistic Approach

- Measures information based on uncertainty and the probability of outcomes, as developed by Claude Shannon. This is the foundation of information theory.
- The less likely an event is, the more information it provides when it occurs.
- Information is quantified in terms of entropy, which represents the average uncertainty of a random variable.
- Entropy!
- Widely used in data compression, telecommunications, and error-correction systems where quantifying uncertainty and information is key to efficient data handling.

Measuring Information: Algorithmic Approach (Kolmogorov Complexity)

- Measures information based on the complexity of a sequence by finding the shortest algorithm (or computer program) that can generate the sequence.
- Information is related to the complexity of the description of data.
- A random sequence has high algorithmic complexity because it cannot be compressed, while a repetitive sequence has low complexity because it can be described with a simple rule (e.g., repeating patterns).
- Kolmogorov Complexity (K): $K(x)$ is the length of the shortest program (in a fixed universal language) that outputs x .
- Example:
  - The sequence “1010101010” can be generated by a simple algorithm (e.g., “repeat ‘10’ five times”), so its algorithmic complexity is low.
  - A truly random sequence like “0110100110101101” cannot be compressed and has high algorithmic complexity.
- Relevant in machine learning, data compression, and the study of randomness and complexity in computational systems.

Discussion topic 1

Alice wants to play a game with Bob, where a coin has to be tossed. To show that she is using a fair coin, she tosses it twenty times. However, in all trials, heads comes up. Can Bob trust this coin? Why?

Would it be different if the sequence of outcomes had been “00101001110101010110”? Why?

Discussion topic 2

You work at a company that sells large sequences of random numbers. Your boss thinks he needs to have some way of testing the “randomness” of the sequences before they are sent to the clients (let us consider this as a kind of quality control of the production. . .). Because you are the best informatics engineer of the company, he asks you to develop a program for testing the randomness of arbitrary sequences.

What should you answer him? Have you any ideas of how to do it? Can you define “randomness”?

Some guiding questions (1)

How does compression relate to information?

What are the limits of compression?

Why are some forms of data more compressible than others?

Data Compression and Information

- **Compression:**
  - Reducing the number of bits required to represent data.
- **Types of Compression:**
  - **Lossless:** No information is lost (e.g., Huffman coding).
  - **Lossy:** Some information is lost for better compression (e.g., JPEG, MP3).
- **The Role of Entropy:**
  - Entropy gives the theoretical limit for lossless compression.
- The entropy $H(X)$ of a source sets the limit on how much you can compress without losing information.
- No lossless compression algorithm can reduce data size below its entropy.

Discussion topic

Alice told Bob that she found an image compression software that is able to reduce the size of every image in, at least, 50% of its original size - for example, if originally the image occupies 1 000 000 bytes, after compression it will require at most 500 000 bytes.

Bob was not very impressed by Alice's statement. However, when Alice added that she would also be able to always recover the original image from its compressed version, Bob immediately replied:

"THAT IS IMPOSSIBLE! Every lossless (i.e., reversible) compression method is limited, i.e., it cannot compress all messages!"

Is Bob correct? Why?

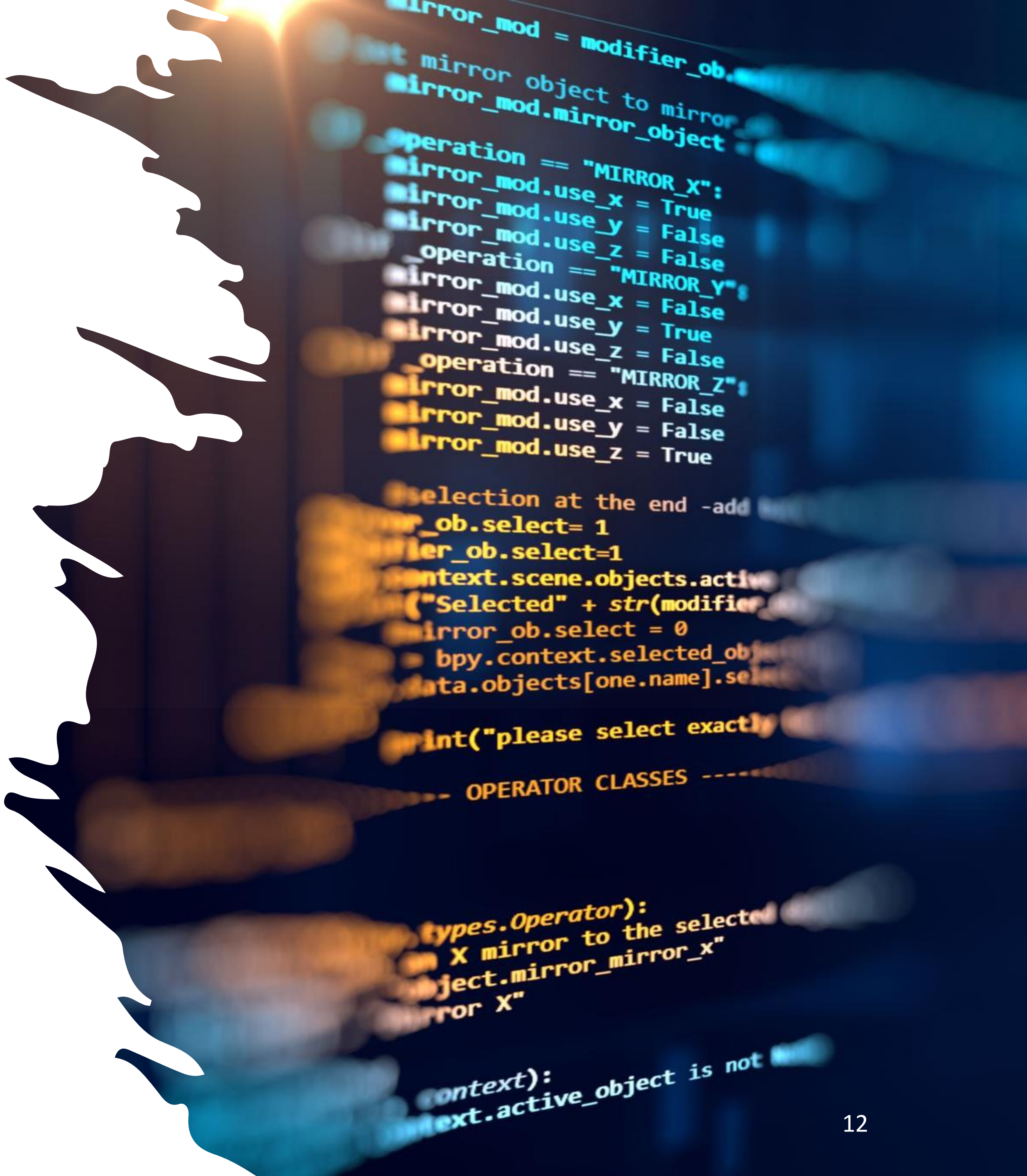
Some guiding questions (2)

What is a computation?

Are there problems that computers cannot solve?

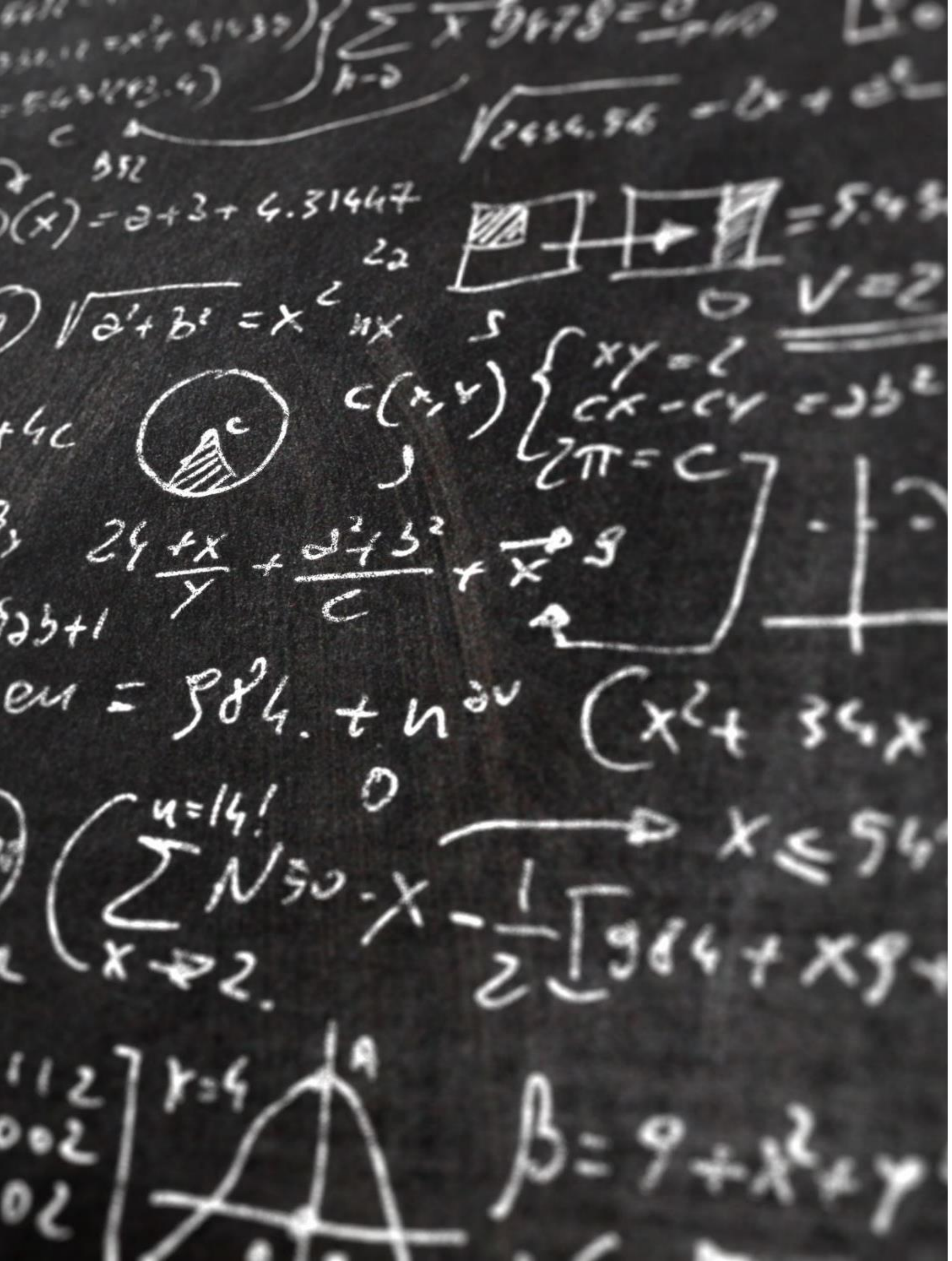
What is Computation?

- Computation refers to the process of transforming input data into output data using a series of operations.
- **Algorithms** perform computations.
- **Example:**
  - Sorting numbers, finding the shortest path in a graph, or compressing a file.
- Computation allows us to extract useful information from raw data.
- **Information Theory:**
  - Helps us quantify and compress data.
- **Computation:**
  - Provides algorithms to process, compress, and transmit data.



Are There Problems Computers Cannot Solve?

- **Undecidable Problems:**
  - Problems that have no algorithm that can solve them for all inputs.
  - Example: **The Halting Problem**, which asks whether a program will eventually halt or run forever.
- **Intractable Problems:**
  - Problems that may have a solution, but it’s computationally impractical to solve them (e.g., NP-hard problems like the Traveling Salesman Problem).
- There are theoretical and practical limits to what computers can compute.



Discussion topic

10 min

Consider the set of all functions $f : N \to \{0, 1\}$. *Alice told Bob that there are functions in that set that cannot be calculated by any finite program, regardless of the programming language used to implement it. Bob thinks Alice is wrong. Is she wrong? Why?*