An introduction to Kolmogorov complexity

Liliana Salvador and Gustavo Lacerda
Complex Systems Summer School 2009

Outline

- Symbols, strings and languages
- Turing machines
- Kolmogorov complexity
- Incompressibility
- Kolmogorov Complexity and Shannon entropy
- Symmetry of Information
- Time-bounded Kolmogorov complexity
- Learning as compression
- Universal learning
- Universal measures of similarity
- Clustering by compression

Symbols, Strings and Languages

- Alphabet A = *{0,1} : finite set of symbols*
- A string over an alphabet A is a finite ordered sequence of symbols from A.
- The empty string, denoted by ε is the (unique) string of length zero.
- Given an alphabet A, we define
  - $A^0 = \{ \varepsilon \}$
  - $A^{n+1} = AA^n$
- A language L over an alphabet A is a subset of A*. That is, $L \subset A^*$.

The Roots...

- Probability theory, information theory, philosophical notions of randomness, theory of algorithms
- Regular sequence Pr(000000000000000000) = $1/2^{20}$
- Regular sequence Pr(01100110011001100110) = $1/2^{20}$
- Random sequence Pr(01000110101111100001) = $1/2^{20}$

Classical probability theory cannot express the notion of randomness of an individual sequence. It can only express expectations of properties of the total set of sequences under some distribution.

How to measure information of a single object?

- It has to be an attribute of the object itself
- Independent of the description method
- Is defined as the length of the shortest binary program from which the object can be effectively reconstructed.
- Example:

  Consider the ensemble of all binary strings of length 9999999999999999.

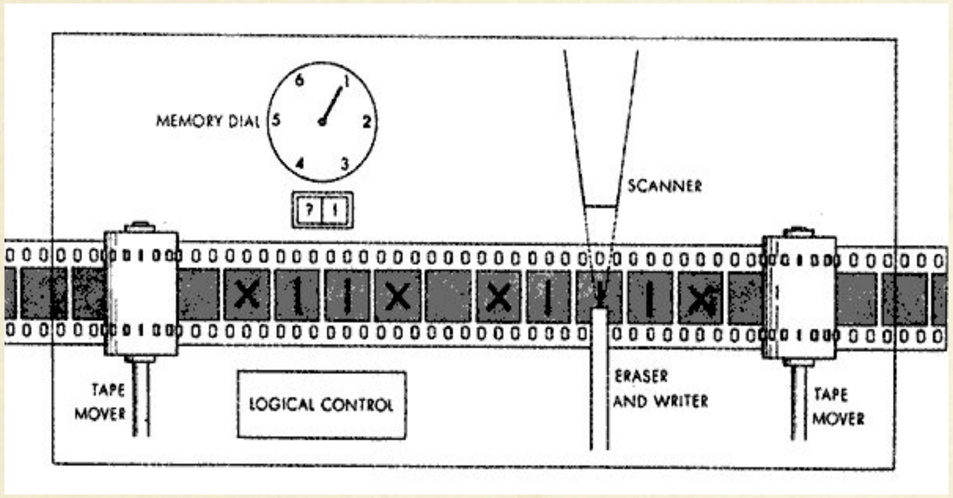
  By Shannon's measure, we require 9999999999999999 bits on average to encode a string in such an ensemble. However, this number can be encoded in about 55 bits by expressing it in binary and adding the repeated pattern 1, or even further 3^2 x 1111111111111111 (that consists of 2^4 1's).

Turing Machine

GREAT. A WAREHOUSE FILLED WITH MILES AND MILES OF REWRITABLE TAPE! WHAT ARE WE EVER GOING TO DO WITH THIS, ALAN?
...ALAN?

And thus the Turing Machine was born.

Turing Machine



A Turing machine

. $T = (S, A, \delta, s0, b, F)$

consists of:

. S = a finite set of states

. A = an alphabet

. $\delta : Sx(A \cup \{b\}) \rightarrow Sx(A \cup \{b\})x\{L, R\}$, the transition function,

. $s0 \in S$, the start state,

. b, marking unused tape cells,

. $F \subset S$, halting and/or D accepting states.

Happy Birthday Alan!

(23 June 1912 – 7 June 1954) was a British mathematician, logician, cryptanalyst and computer scientist.

Considered to be the father of modern computer science.

Provided an influential formalization of the concept of the algorithm and computation with the Turing machine.

Turing was homosexual, living in an era when homosexuality was considered a mental illness and homosexual acts were illegal. He was criminally prosecuted, which essentially ended his career.

He died not long after from what was officially declared self-induced cyanide poisoning, although his death was considered very ambiguous.

Halting Problem

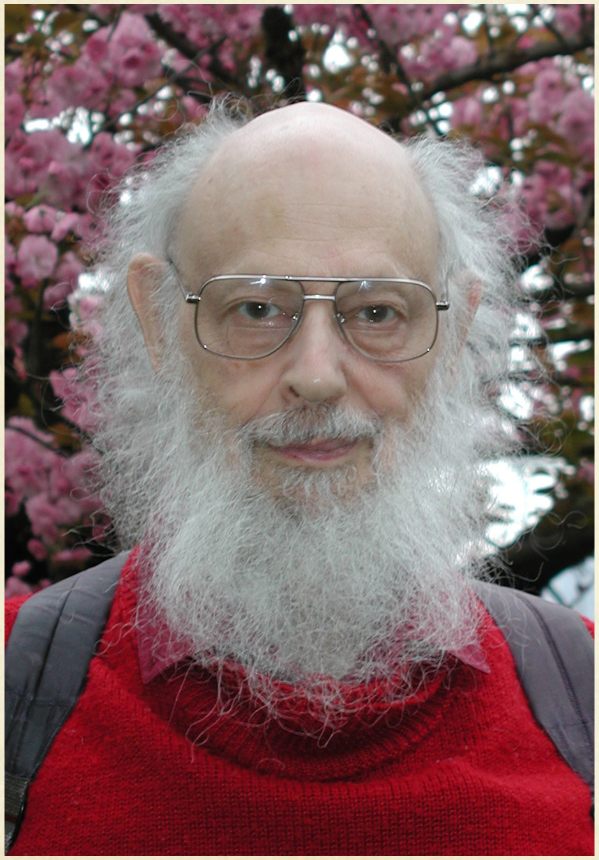
- Given a program and an input to the program, whether the program will eventually halt when run with that input.
- The halting problem is famous because it was one of the first problems proven algorithmically undecidable (not computable).
- This means there is no algorithm which can be applied to any arbitrary program and input to decide whether the program stops when run with that input.

Occam’s Razor
William of Ockham (1290--1349)

- “Entities should not be multiplied beyond necessity.”
- Commonly explained as: when have choices, choose the simplest theory.
- Bertrand Russell: ``It is vain to do with more what can be done with fewer."
- Newton (*Principia*): ``Natura enim simplex est, et rerum causis superfluis non luxuriat".

Ray Solomonoff

(born 1926, Cleveland, Ohio)



Algorithmic probability

Theory of inductive inference

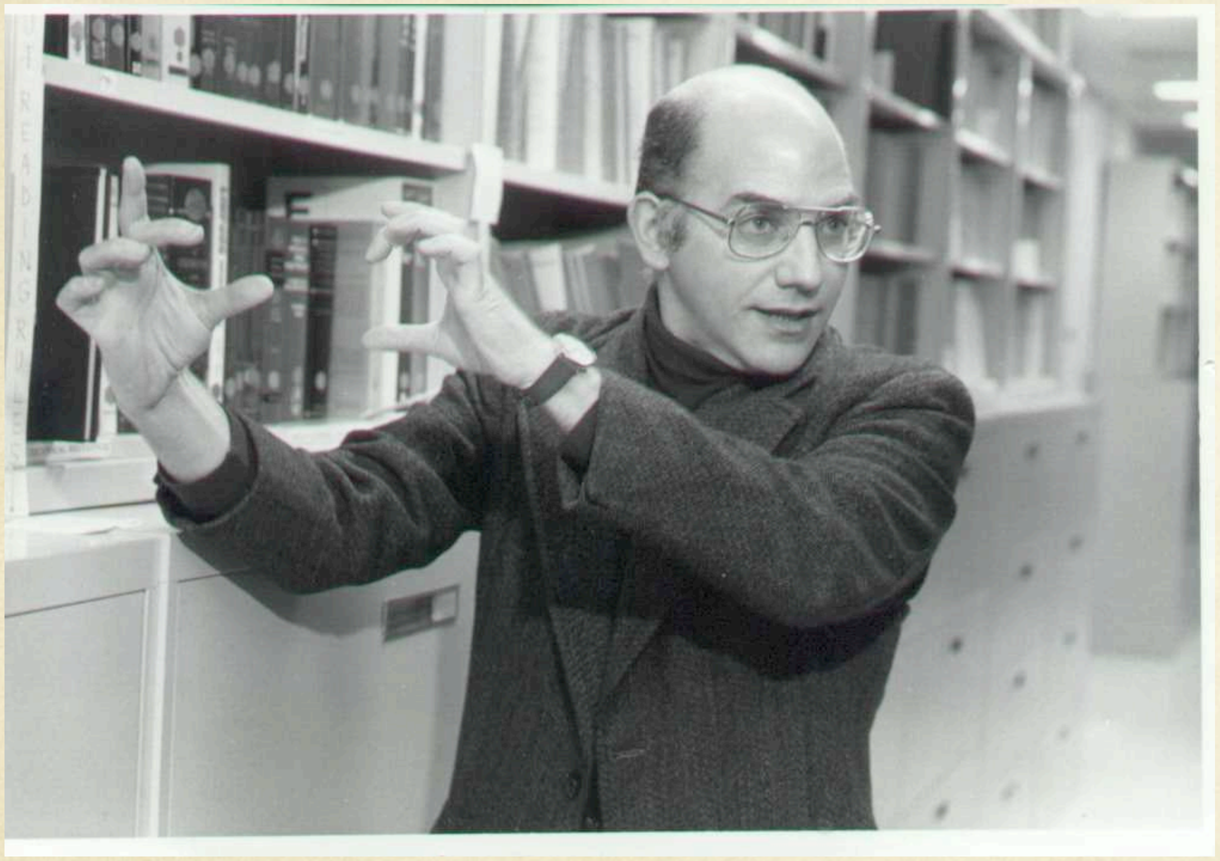
Attended the first meeting
Where AI became a field
.....

Andrey Nikolaevich Kolmogorov
(1903, Tambov, Russia–1987 Moscow)

- Measure Theory
- Probability
- Analysis
- Intuitionistic Logic
- Cohomology
- Dynamical Systems
- Hydrodynamics
- Kolmogorov complexity

Gregory Chaitin

(born 1947, is an Argentine-American)



Algorithmic information theory
Chaitin's constant Ω
Biology
Neuroscience
Philosophy
……

Algorithmic information

(Kolmogorov complexity)

- Solomonoff (1960)-Kolmogorov (1965)-Chaitin (1969): The amount of information in a string is the size of the smallest program of an optimal Universal TM U generating that string.

$$K_U(x) = \min_p \{|p| : U(p) = x\}$$

- *Invariance Theorem*: It does not matter which optimal universal Turing machine U we choose. I.e. all "universal encoding methods" are ok.

[Paul Vitanyi slide]

Proof of the Invariance theorem

- Fix an effective enumeration of all
  Turing
  m
  achines (TM's): T_1, T_2, ... Define $K = \min \{|p| : T(p) = x\}$
- U is an optimal universal TM such that (p produces x)

$$\mathbf{U(1^n0p) = T_n(p)}$$

- Then for all x:

$K_U(x) \leq K_{T_n}(x) + n+1$, and $|K_U(x) - K_{U'}(x)| \leq c$

Fixing U, we write K(x) instead of $K_U(x)$.

[Paul Vitanyi slide]

Properties and examples

- Intuitively: *K(x)*= length of shortest description of *x*

$$K(x) \leq |x| + O(1)$$

- K(*x*|*y*)=length of shortest description of *x* given *y*.

$$K(x|y) \leq K(x) + O(1)$$

  - For all x,
  - $K(x|x) = O(1)$
  - $K(x|\varepsilon) = k(x)$; $K(\varepsilon|x) = O(1)$

  - $K(x,x) = K(x) + O(1)$
  - $K(x,y) \leq K(x) + K(y) + O(\log(\min\{K(x),K(y)\})$

Randomness

- Randomness of strings mean that they do not contain regularities.
- If the regularities are not effective, then we cannot use them.
- Hence, we consider randomness of strings as the lack of effective regularities (that can be exploited).
- For example: a random string cannot be compressed by any known or unknown real-world compressor.

[Paul Vitanyi slide]

Intuition: Randomness = Incompressibility

- For constant $c>0$, a string $x \in \{0,1\}^*$ is c-incompressible if

$$K(x) \geq |x|-c.$$

Usually, we often simply say that x is incompressible
(We will call incompressible strings random strings.)

Shannon Entropy and Kolmogorov complexity

- Shannon entropy of random variable X over sample size S

  $H(X) = \sum P(X=x) \log 1/P(X=x)$

- H(X) bits are necessary on P-average to describe the outcome x.
- Example. For P uniform over finite S, we have

  $H(X)=\sum (1/|S|) \log |S| = \log |S|$.

- Kolmogorov complexity, is the minimum description (smallest program) for one fixed x. And the expected H(X) and the P-expectation of K(x) converge to the same thing.

  $H(P) = - \sum P(x) \log P(x)$ is assimptotically equal to the expected complexity

  $\sum_x P(x)K(x) \leq - \log P(x) + O(1)$

[Paul Vitanyi slide]

Symmetry of information

$$I(x;y) = K(y) - K(y|x)$$

$$= K(x) - K(x|y)$$

$$= I(x;y)$$ (up to an additive log term)

(the first term is read as “the information x knows about y)

Entropy or CK?

"It has been shown that although in practice we can't be guaranteed to get the right answer to either the entropy or computational complexity values, we can be sure that they are (essentially) equal to each other, so both methods can be useful, depending on what we know about the system, and what our local goals are. "

- Tom Carter's notes

Resource bounded KC

- K complexity depends on unlimited computational resources. Kolmogorov himself first observed that we can put resource bounds on such computations. This was subsequently studied by Barzdins, Loveland, Daley, Levin, Adleman.

- In the 1960's, two parallel theories were developed:
  - Computational complexity – Hartmanis and Stearns, Rabin, Blum, measuring time/space complexity
  - Kolmogorov complexity, measuring information.
- Resource bounded KC links the two theories.

[Paul Vitanyi slide]

Theory

- $C^{t,s}(x|y)$ is the t-time s-space limited Kolmogorov complexity of x condition on y. I.e. the length of shortest program that with input y, *produces* x in time t(n) and space s(n).

- In standard K complexity, it does not matter if we say "produces x" or "accepts x", they are the same. But in resource bounded case, they are likely to be different. So Sipser defined $CD^{t,s}(x|y)$ to be the length of the shortest program that with input y *accepts* x in time t(n) and space s(n).
- When we use just one parameter such as time, we will simply write C^t or CD^t.

[Paul Vitanyi slide]

Learning as (lossless) compression:

- Q: how do you compress well?
- A: by finding patterns in the data, i.e. by learning to predict it
- compression is learning: in the long run, it's impossible to compress a source without learning the patterns in it
- learning is compression: the output of learning from data is a more compact representation of the data

Universal learning

- KC suggests a universal learning algorithm: given data from a source, search for a TM that outputs the same distribution.
- Of course, there are infinitely many such programs.
- One way to encode Occam's razor is to select the smallest such program. This is roughly the idea behind MDL learning (in reality, MDL uses a restricted class of programs)

Universal Measures of Similarity

Normalized Compression Distance:

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}.$$

Since KC is uncomputable, we estimate it using gzip (this is an overestimate)

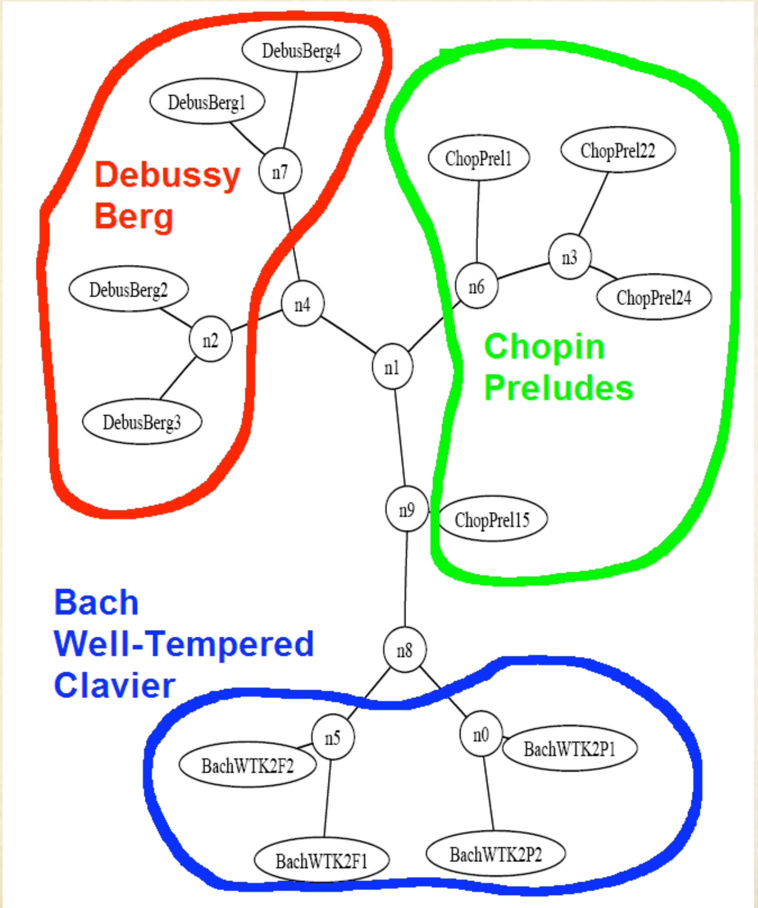
Grabbing random articles Wikipedia in 4 languages: (Portuguese, Spanish, Dutch, German), we compute NCD, and find the following distances:

NL1-NL2 0.9062

PT-ES .9774, .9698
NL-DE .9801, .9812
PT-NL .9872, .9871
PT-DE .9965, .9957
ES-NL .9917, .9961
ES-DE .9975, 1.000

Clustering by Compression

Clustering music



from Rudi Cilibrasi and Paul M.B. Vitányi - "Clustering by Compression"
IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 51, NO 4, APRIL 2005, 1523–1545

Warnings!

- Warning: gzip uses superficial features! It won't capture capture deeper similarities, since that would require lots of data and computing time.
- This is universal learning: an ideal compression algorithm will find/exploit any pattern. Of course, this means that this method really sucks in practice!