Applications of Algorithmic Information Theory

Marcus Hutter
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

ANU

Abstract

Algorithmic information theory has a wide range of applications, despite the fact that its core quantity, Kolmogorov complexity, is incomputable. Most importantly, AIT allows to quantify Occam's razor, the core scientific paradigm that "among two models that describe the data equally well, the simpler one should be preferred". This led to universal theories of induction and action in the field of machine learning and artificial intelligence, and practical versions like the Minimum Encoding Length (MDL/MML) principles. The universal similarity metric probably spawned the greatest practical success of AIT. Approximated by standard compressors like Lempel-Ziv (zip) or bzip2 or PPMZ, it leads to the normalized compression distance, which has been used to fully automatically reconstruct language and phylogenetic trees, and many other clustering problems. AIT has been applied in disciplines as remote as Cognitive Sciences, Biology, Physics, and Economics.

Presented Applications of AIT

- Philosophy: problem of induction
- Machine learning: time-series forecasting
- Artificial intelligence: foundations [COMP4620/COMP8620]
- Probability theory: choice of priors
- Information theory: individual randomness/information
- Data mining: clustering, measuring similarity
- Bioinformatics: phylogeny tree reconstruction
- Linguistics: language tree reconstruction

CONTENTS

1 MINI-INTRODUCTION TO KOLMOGOROV COMPLEXITY

- Kolmogorov Complexity K(x)
- Properties of Kolmogorov complexity
- Schematic Graph of Kolmogorov Complexity
- Relation to Shannon Entropy

Kolmogorov Complexity K(x)

K. of string x is the length of the shortest (prefix) program producing x:

$$K(x|y) := \min_p\{l(p) : U(y,p) = x\}, \quad U = \text{universal TM}$$

For non-string objects o (like numbers and functions) we define $K(o) := K(\langle o \rangle)$, where $\langle o \rangle \in \mathcal{X}^*$ is some standard code for o.

+ Simple strings like $000...0$ have small K,
  irregular (e.g. random) strings have large K.
- The definition is nearly independent of the choice of U.
+ K satisfies most properties an information measure should satisfy.
+ K shares many properties with Shannon entropy but is superior.
– $K(x)$ is not computable, but only semi-computable from above.

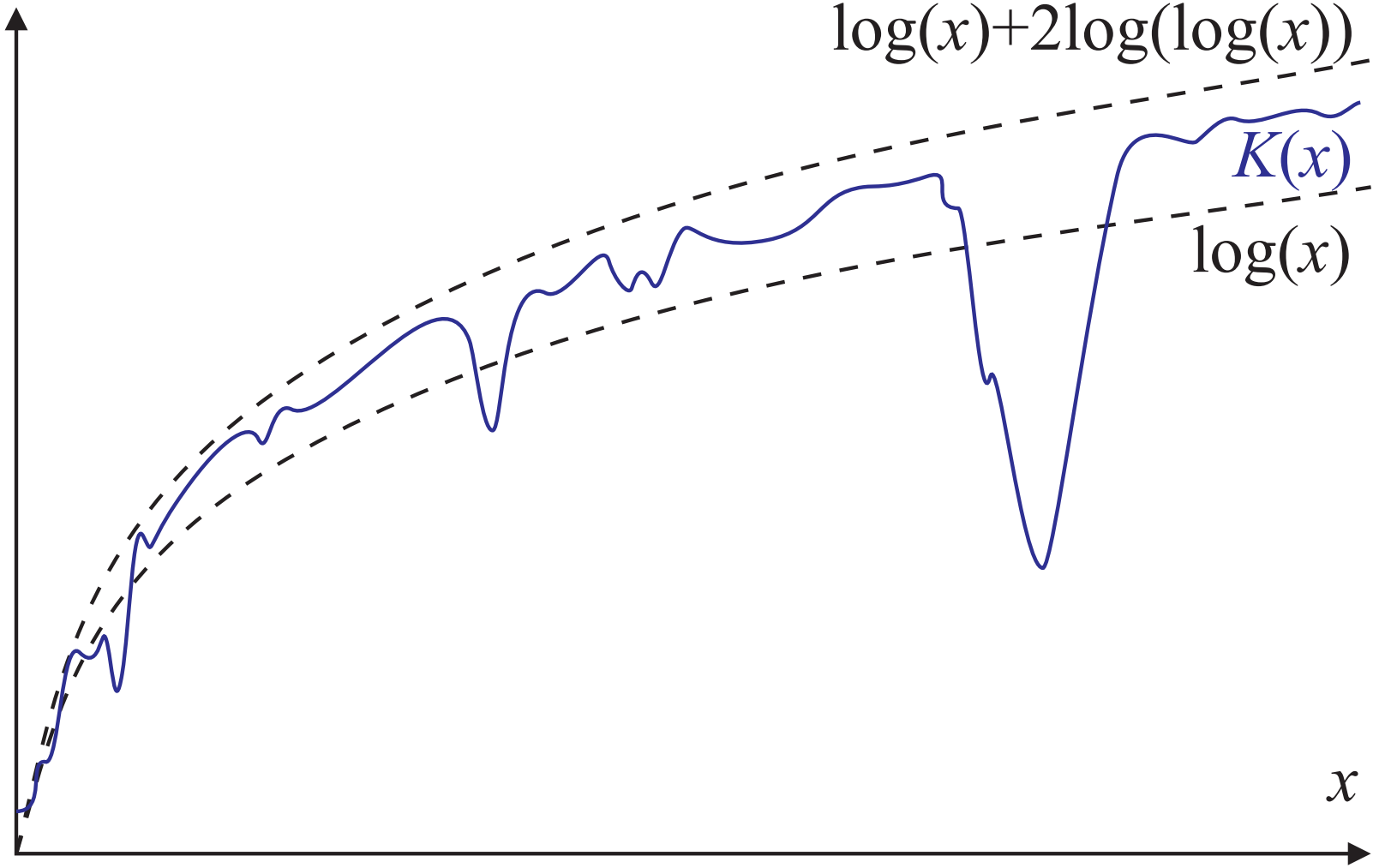
Fazit: K is an excellent universal complexity measure, suitable for quantifying Occam's razor.

Properties of Kolmogorov Complexity

- Upper bound: $K(x) \stackrel{+}{<} \ell(x) + 2\log \ell(x)$
- Kraft inequality: $\sum_x 2^{-K(x)} \leq 1, \quad K(x) \geq \ell(x)$ for 'most' x.
- Lower bound: $K(x) \geq \ell(x)$ for 'most' x, $K(n) \to \infty$ for $n \to \infty$
- Extra information: $K(x|y) \stackrel{+}{<} K(x) \stackrel{+}{<} K(x,y)$
- Symmetry: $K(x|y,K(y)) + K(y) \stackrel{+}{=} K(x,y) \stackrel{+}{=} K(y,x)$.
- Information non-increase: $K(f(x)) \stackrel{+}{<} K(x) + K(f)$ for comp. f
- MDL bound: $K(x) \stackrel{+}{<} -\log P(x) + K(P)$
  for computable $P : \{0,1\}^* \to [0,1]$ and $\sum_x P(x) \leq 1$
- K is upper semi-computable but not finitely computable.

Schematic Graph of Kolmogorov Complexity

Although $K(x)$ is incomputable, we can draw a schematic graph



Relation to Shannon Entropy

Let $X, Y \in \mathcal{X}$ be discrete random variable with distribution $P(X,Y)$.

Definition 1.1 (Definition of Shannon entropy)

$$\text{Entropy}(X) \equiv H(X) := -\sum_{x\in\mathcal{X}} P(x)\log P(x)$$
$$\text{Entropy}(X|Y) \equiv H(X|Y) := -\sum_{y\in\mathcal{Y}} P(y)\sum_{x\in\mathcal{X}} P(x|y)\log P(x|y)$$

Theorem 1.2 (Properties of Shannon entropy)

- Upper bound: $H(X) \leq \log|\mathcal{X}| = n$ for $\mathcal{X} = \{0,1\}^n$
- Extra information: $H(X|Y) \leq H(X) \leq H(X,Y)$
- Subadditivity: $H(X,Y) \leq H(X)+H(Y)$
- Symmetry: $H(X|Y)+H(Y) = H(X,Y) = H(Y,X)$
- Information non-increase: $H(f(X)) \leq H(X)$ for any f

Relations for H are essentially expected versions of relations for K.

2 UNIVERSAL A PRIORI PROBABILITY

- The Universal a Priori Probability M
- Relations between Complexities
- Fundamental Universality Property of M

Philosophy & Notation

Occam's razor: take simplest hypothesis consistent with data.

Epicurus' principle of multiple explanations: Keep all theories consistent with the data.

$\Downarrow$ $\Downarrow$

We now combine both principles:

Take all consistent explanations into account,

but weight the simpler ones higher.

Formalization with Turing machines and Kolmogorov complexity

Notation: We denote binary strings of length $\ell(x) = n$ by $x = x_{1:n} = x_1 x_2 ... x_n$ with $x_t \in \{0,1\}$ and further abbreviate $x_{<n} := x_1 ... x_{n-1}$.

The Universal a Priori Probability M

Solomonoff defined the universal probability distribution $M(x)$ as the probability that the output of a universal monotone Turing machine starts with x when provided withfair coin flips on the input tape.

Definition 2.1 (Solomonoff distribution) Formally,

$$M(x) := \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)}$$

The sum is over minimal programs p for which U outputs a string starting with x.

Since the shortest programs p dominate the sum, $M(x)$ is roughly $2^{-Km(x)}$. More precisely ...

Relations between Complexities

Theorem 2.2 (Relations between Complexities)

$KM := -\log M$, Km, and K are ordered in the following way:

$$0 \;\leq\; K(x|\ell(x)) \;\stackrel{+}{<}\; KM(x) \;\leq\; Km(x) \;\leq\; K(x) \;\stackrel{+}{<}\; \ell(x) + 2\log\ell(x)$$

Proof sketch:

The second inequality follows from the fact that,
given n and Kraft's inequality $\sum_{x\in\mathcal{X}^n} M(x) \leq 1$,
there exists for $x \in \mathcal{X}^n$ a Shannon-Fano code of length $-\log M(x)$,
which is effective since M is enumerable.

Now use the MDL bound conditioned to n.

The other inequalities are obvious from the definitions. ■

3 UNIVERSAL SEQUENCE PREDICTION

- Solomonoff, Occam, Epicurus
- Prediction
- Simple Deterministic Bound
- Solomonoff's Major Result
- Implications of Solomonoff's Result
- Universal Inductive Inference
- More Stuff / Critique / Problems

Solomonoff, Occam, Epicurus

- In which sense does M incorporate Occam's razor and Epicurus' principle of multiple explanations?
- From $M(x) \approx 2^{-K(x)}$ we see that M assigns high probability to simple strings (Occam).
- More useful is to think of x as being the observed history.
- We see from Definition 2.1 that every program p consistent with history x is allowed to contribute to M (Epicurus).
- On the other hand, shorter programs give significantly larger contribution (Occam).

Prediction

How does all this affect prediction?

If $M(x)$ correctly describes our (subjective) prior belief in x, then

$$M(y|x) := M(xy)/M(x)$$

must be our posterior belief in y.

From the symmetry of algorithmic information $K(x,y) \stackrel{+}{=} K(y|x,K(x)) + K(x)$, and assuming $K(x,y) \approx K(xy)$, and approximating $K(y|x,K(x)) \approx K(y|x)$, $M(x) \approx 2^{-K(x)}$, and $M(xy) \approx 2^{-K(xy)}$ we get:

$$M(y|x) \approx 2^{-K(y|x)}$$

This tells us that M predicts y with high probability iff y has an easy explanation, given x (Occam & Epicurus).

Simple Deterministic Bound

Sequence prediction algorithms try to predict the continuation $x_t \in \{0,1\}$ of a given sequence $x_1...x_{t-1}$. Simple deterministic bound:

$$\sum_{t=1}^{\infty} |1-M(x_t|x_{<t})| \overset{a}{\leq} -\sum_{t=1}^{\infty} \ln M(x_t|x_{<t}) \overset{b}{=} -\ln M(x_{1:\infty}) \overset{c}{\leq} \ln 2 \cdot Km(x_{1:\infty})$$

(a) use $|1-a| \leq -\ln a$ for $0 \leq a \leq 1$.

(b) exchange sum with logarithm and eliminate product by chain rule.

(c) used Theorem 2.2.

If $x_{1:\infty}$ is a computable sequence, then $Km(x_{1:\infty})$ is finite,
which implies $M(x_t|x_{<t}) \to 1 \quad (\sum_{t=1}^{\infty} |1-a_t| < \infty \Rightarrow a_t \to 1)$.

$\Rightarrow$ if environment is a computable sequence (digits of π or Expert or ...), after having seen the first few digits, M correctly predicts the next digit with high probability, i.e. it recognizes the structure of the sequence.

More Stuff / Critique / Problems

- Other results: M convergence rapidly also on stochastic sequences; solves the zero-prior & old evidence & new theories problems; can confirm universal hypotheses; is reparametrization invariant; predicts better than all other predictors.
- Prior knowledge y can be incorporated by using "subjective" prior $w^U_{\nu|y} = 2^{-K(\nu|y)}$ or by prefixing observation x by y.
- Additive/multiplicative constant fudges and U-dependence is often (but not always) harmless.
- Incomputability: K and M can serve as "gold standards" which practitioners should aim at, but have to be (crudely) approximated in practice (MDL [Ris89], MML [Wal05], LZW [LZ76], CTW [WSTT95], NCD [CV05]).

4 MARTIN-LÖF RANDOMNESS

- When is a Sequence Random? If it is incompressible!
- Motivation: For a fair coin 00000000 is as likely as 01100101, but we "feel" that 00000000 is less random than 01100101.
- Martin-Löf randomness captures the important concept of randomness of **individual** sequences.
- Martin-Löf random sequences pass all effective randomness tests.

When is a Sequence Random?

a) Is 01100101001011011010011110 11 generated by a fair coin flip?
b) Is 11111111111111111111111111 11 generated by a fair coin flip?
c) Is 11001001000011111110110101010 generated by a fair coin flip?
d) Is 01010101010101010101010101 01 generated by a fair coin flip?

- Intuitively: (a) and (c) look random, but (b) and (d) look unlikely.
- Problem: Formally (a-d) have equal probability $(\frac{1}{2})^{length}$.
- Classical solution: Consider hypothesis class $H := \{\text{Bernoulli}(p) : p \in \Theta \subseteq [0,1]\}$ and determine p for which sequence has maximum likelihood $\Longrightarrow$ (a,c,d) are fair Bernoulli($\frac{1}{2}$) coins, (b) not.
- Problem: (d) is non-random, also (c) is binary expansion of π.
- Solution: Choose H larger, but how large? Overfitting? MDL?
- AIT Solution: A sequence is **random** *iff* it is **incompressible**.

Martin-Löf Random Sequences

Characterization equivalent to Martin-Löf's original definition:

Theorem 4.1 (Martin-Löf random sequences)
A sequence $x_{1:\infty}$ is μ-random (in the sense of Martin-Löf)
$\iff$ there is a constant c such that $M(x_{1:n}) \leq c \cdot \mu(x_{1:n})$ for all n.

Equivalent formulation for computable μ:

$$x_{1:\infty} \text{ is } \mu\text{.M.L.-random} \quad \iff \quad Km(x_{1:n}) \stackrel{+}{=} -\log\mu(x_{1:n}) \; \forall n, \quad (4.2)$$

Theorem 4.1 follows from (4.2) by exponentiation, "using $2^{-Km} \approx M$" and noting that $M \stackrel{\times}{>} \mu$ follows from universality of M.

Properties of ML-Random Sequences

- Special case of μ being a fair coin, i.e. $\mu(x_{1:n}) = 2^{-n}$, then $x_{1:\infty}$ is random $\iff$ $Km(x_{1:n}) \stackrel{+}{=} n$, i.e. iff $x_{1:n}$ is incompressible.
- For general μ, $-\log\mu(x_{1:n})$ is the length of the Shannon-Fano code of $x_{1:n}$, hence $x_{1:\infty}$ is μ-random $\iff$ the Shannon-Fano code is optimal.
- One can show that a μ-random sequence $x_{1:\infty}$ passes **all** thinkable effective randomness tests, e.g. the law of large numbers, the law of the iterated logarithm, etc.
- In particular, the set of all μ-random sequences has μ-measure 1.

Summary

- Solomonoff's universal a priori probability $M(x)$

  $=$ Occam $+$ Epicurus $+$ Turing $+$ Bayes $+$ Kolmogorov

  $=$ output probability of a universal TM with random input

  $=$ enum. semimeasure that dominates all enum. semimeasures

  $\approx 2^{-\text{Kolmogorov complexity}}$

- $M(x_t|x_{<t}) \to \mu(x_t|x_{<t})$ rapid w.p.1 $\forall$ computable μ.

- M solves/avoids/meliorates many if not all philosophical and statistical problems around induction.

- Fazit: M is universal predictor.

- Matin-Löf /Kolmogorov define randomness of individual sequences: A sequence is random *iff* it is incompressible.

5 THE MINIMUM DESCRIPTION LENGTH PRINCIPLE

- MDL as Approximation of Solomonoff's M
- The Minimum Description Length Principle

MDL as Approximation of Solomonoff's M

- Approximation of Solomonoff, since M incomputable:
- $M(x) \approx 2^{-Km(x)}$ (excellent approximation)
- $Km(x) \equiv Km_U(x) \approx Km_T(x)$
  (approximation quality depends on T and x)
- Predict y of highest $M(y|x)$ is approximately same as
- MDL: Predict y of smallest complexity $Km_T(xy)$.
- Examples for x: Daily weather or stock market data.
- Example for T: Lempel-Ziv decompressor.

The Minimum Description Length Principle

Identification of probabilistic model "best" describing data:

Probabilistic model(=hypothesis) H_ν with $\nu \in \mathcal{M}$ and data D.

Most probable model is $\nu^{\text{MDL}} = \arg\max_{\nu\in\mathcal{M}} p(H_\nu|D)$.

Bayes' rule: $p(H_\nu|D) = p(D|H_\nu)\cdot p(H_\nu)/p(D)$.

Occam's razor: $p(H_\nu) = 2^{-Kw(\nu)}$.

By definition: $p(D|H_\nu) = \nu(x)$, $D = x$ =data-seq., $p(D)$ =const.

Take logarithm:

Definition 5.1 (MDL) $\nu^{\text{MDL}} = \arg\min_{\nu\in\mathcal{M}}\{K\nu(x) + Kw(\nu)\}$

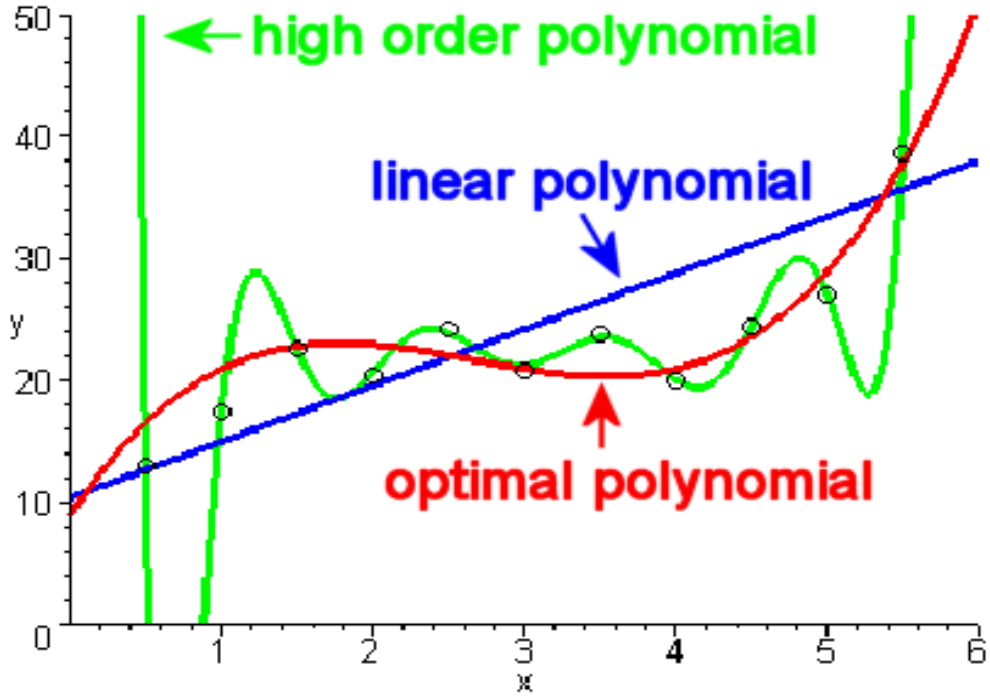
$K\nu(x) := -\log\nu(x)$ = length of Shannon-Fano code of x given H_ν.

$Kw(\nu)$ = length of model H_ν.

Names: Two-part MDL or MAP or MML ($\exists$ slight/major differences)

Application: Regression / Polynomial Fitting

- Data $D = \{(x_1, y_1), ..., (x_n, y_n)\}$
- Fit polynomial $f_d(x) := a_0 + a_1 x + a_2 x^2 + ... + a_d x^d$ of degree d through points D
- Measure of error: $SQ(a_0...a_d) = \sum_{i=1}^{n} (y_i - f_d(x_i))^2$
- Given d, minimize $SQ(a_{0:d})$ w.r.t. parameters $a_0...a_d$.
- This classical approach does not tell us how to choose d? ($d \geq n-1$ gives perfect fit)



6 THE UNIVERSAL SIMILARITY METRIC

- Conditional Kolmogorov Complexity
- The Universal Similarity Metric
- Tree-Based Clustering
- Genomics & Phylogeny: Mammals, SARS Virus & Others
- Classification of Different File Types
- Language Tree (Re)construction
- Classify Music w.r.t. Composer
- Further Applications
- Summary

Based on [Cilibrasi&Vitanyi'05]

Conditional Kolmogorov Complexity

Question: When is object=string x similar to object=string y?

Universal solution: x similar y $\Leftrightarrow$ x can be easily (re)constructed from y $\Leftrightarrow$ Kolmogorov complexity $K(x|y) := \min\{\ell(p) : U(p,y) = x\}$ is small

Examples:

1) x is very similar to itself ($K(x|x) \stackrel{+}{=} 0$)
2) A processed x is similar to x ($K(f(x)|x) \stackrel{+}{=} 0$ if $K(f) = O(1)$).
   e.g. doubling, reverting, inverting, encrypting, partially deleting x.
3) A random string is with high probability not similar to any other string ($K(\text{random}|y) = \text{length}(\text{random})$).

The problem with $K(x|y)$ as similarity=distance measure is that it is neither symmetric nor normalized nor computable.

The Universal Similarity Metric

- Symmetrization and normalization leads to a/the universal metric d:

$$0 \;\leq\; d(x,y) := \frac{\max\{K(x|y),K(y|x)\}}{\max\{K(x),K(y)\}} \;\leq\; 1$$

- Every effective similarity between x and y is detected by d
- Use $K(x|y)\approx K(xy)-K(y)$ (coding T) and $K(x)\equiv K_U(x)\approx K_T(x)$ $\Longrightarrow$ computable approximation: Normalized compression distance:

$$d(x,y) \;\approx\; \frac{K_T(xy)-\min\{K_T(x),K_T(y)\}}{\max\{K_T(x),K_T(y)\}} \;\lesssim\; 1$$

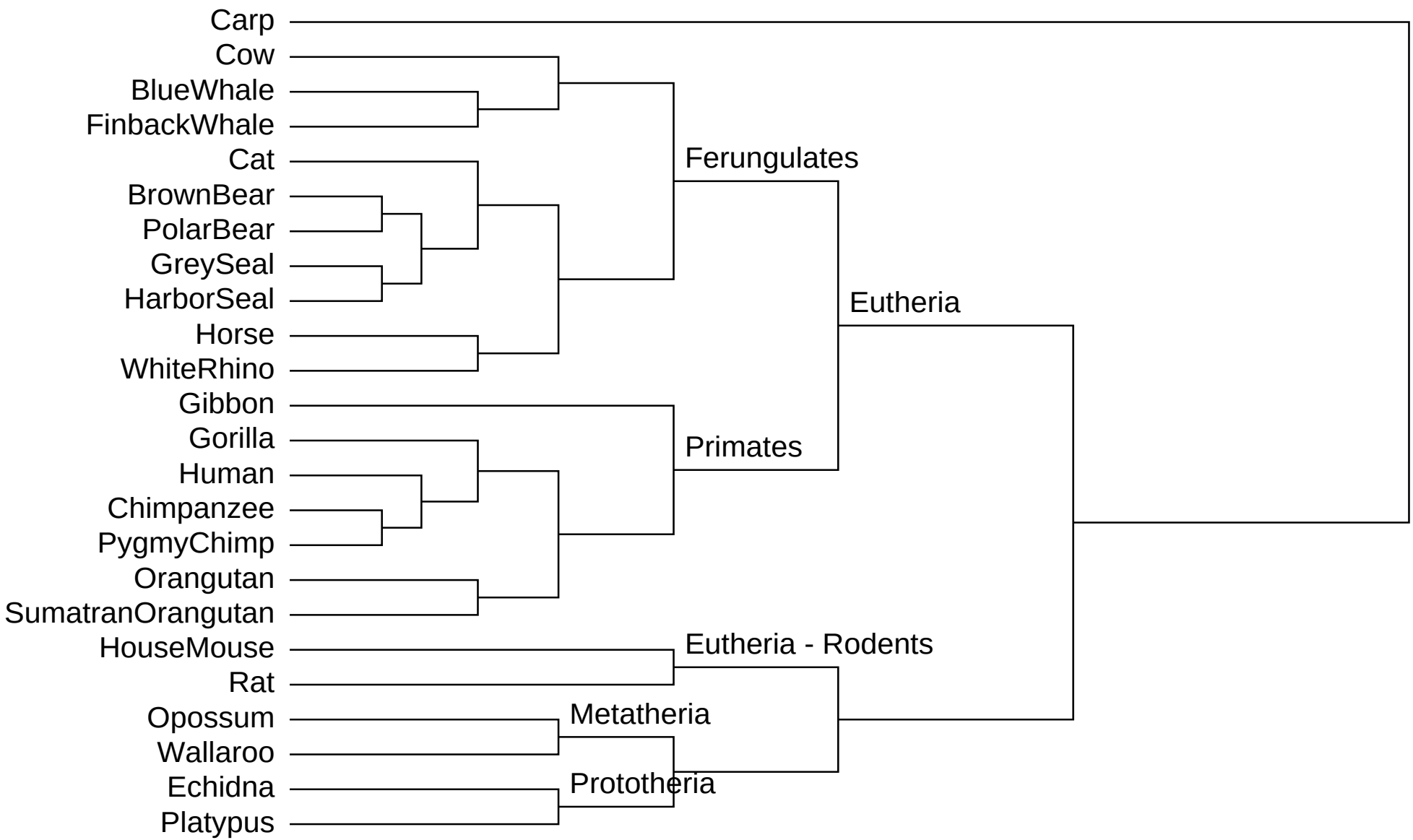
- For T choose Lempel-Ziv or gzip or bzip(2) (de)compressor in the applications below.
- Theory: Lempel-Ziv compresses asymptotically better than any probabilistic finite state automaton predictor/compressor.

Tree-Based Clustering [CV'05]

- If many objects $x_1, ..., x_n$ need to be compared, determine the Similarity matrix: $M_{ij} = d(x_i, x_j)$ for $1 \leq i, j \leq n$
- Now cluster similar objects.
- There are various clustering techniques.
- Tree-based clustering: Create a tree connecting similar objects,
- e.g. quartet method (for clustering)
- Applications: Phylogeny of 24 Mammal mtDNA,
  50 Language Tree (based on declaration of human rights),
  composers of music, authors of novels, SARS virus, fungi,
  optical characters, galaxies, ... [Cilibrasi&Vitanyi'05]

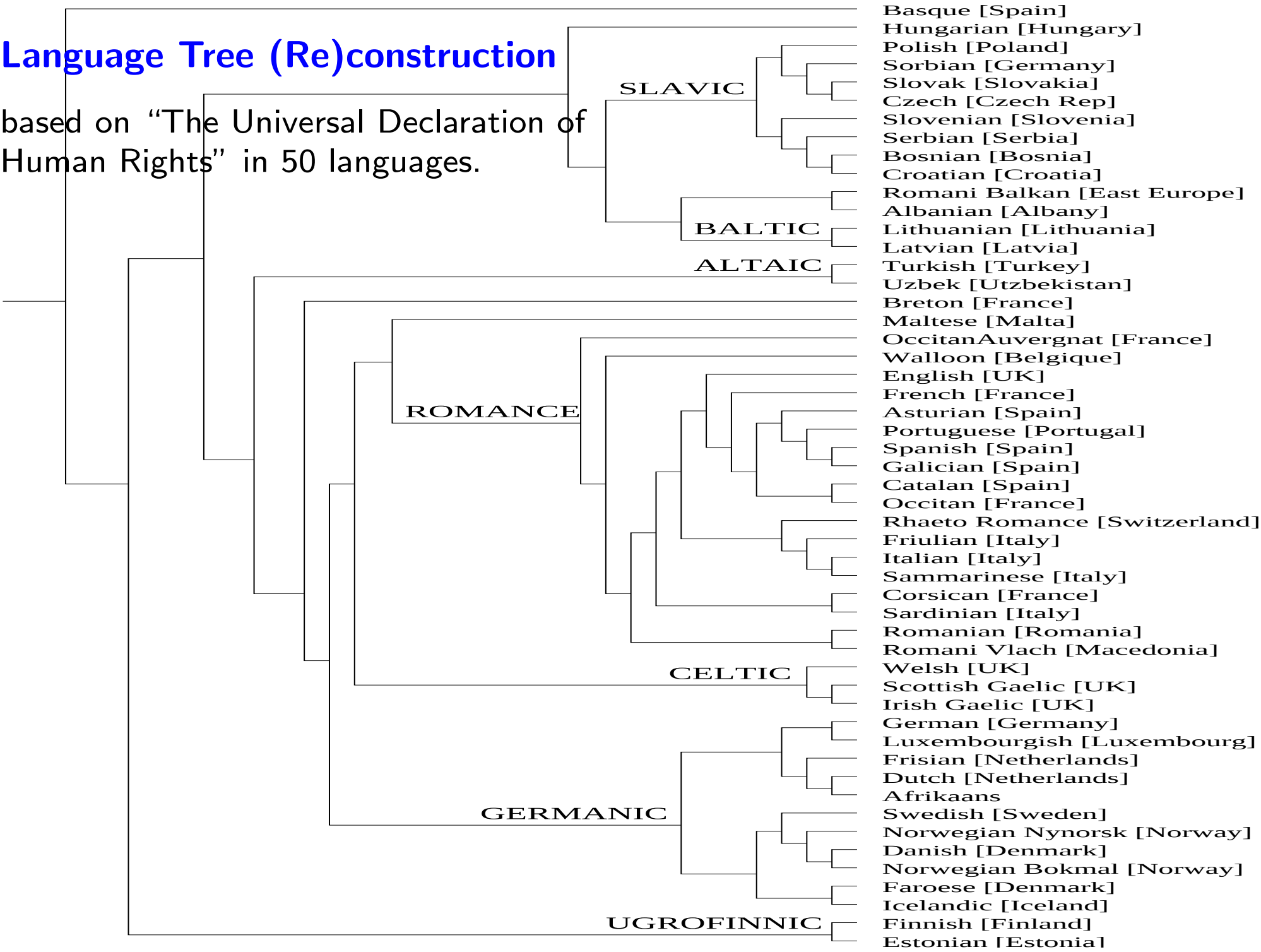
Genomics & Phylogeny: Mammals

Evolutionary tree built from complete mammalian mtDNA of 24 species:



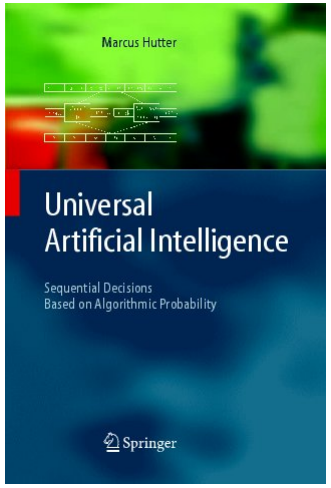
Language Tree (Re)construction

based on "The Universal Declaration of Human Rights" in 50 languages.



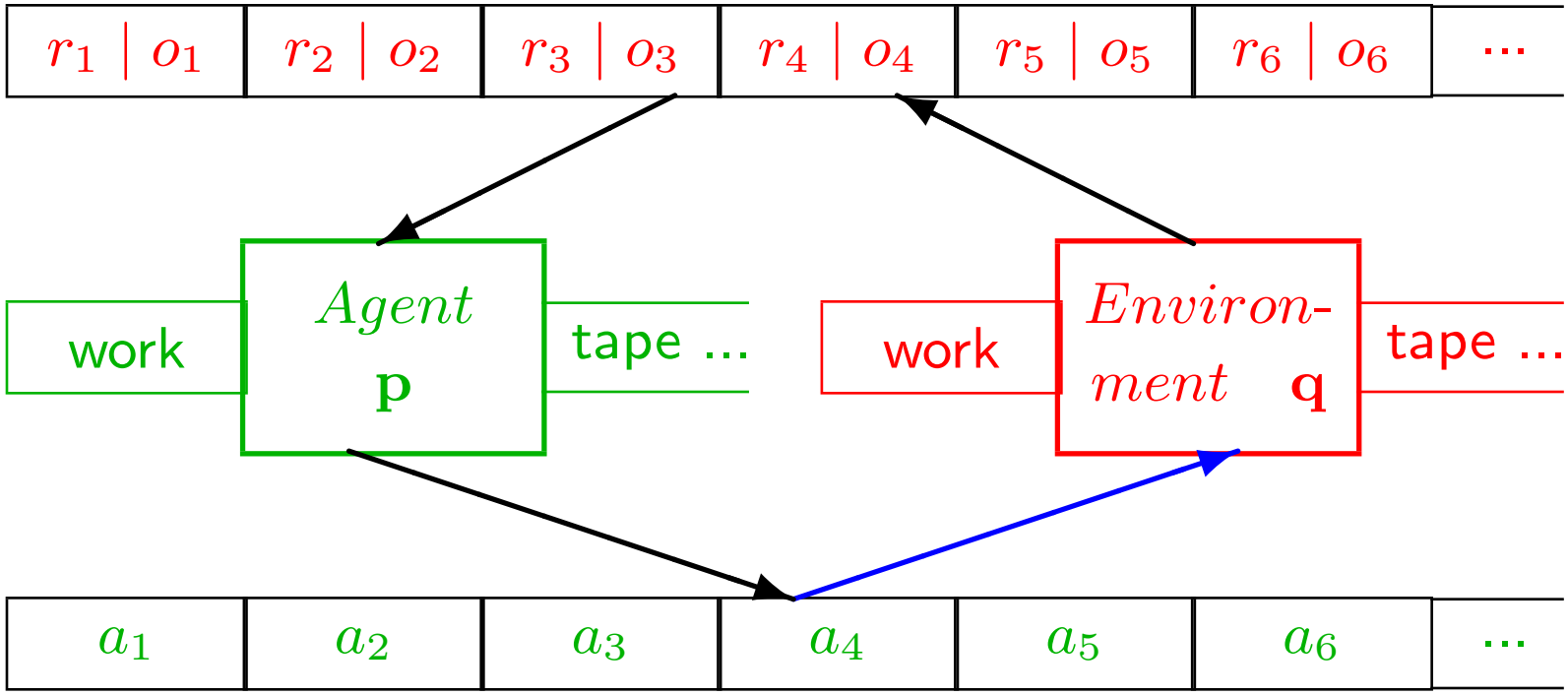
7 ARTIFICIAL INTELLIGENCE

- The Agent Model
- Universal Artificial Intelligence



The Agent Model

Most if not all AI problems can be formulated within the agent framework



Formal Definition of Intelligence

- Agent follows policy $\pi : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^* \rightsquigarrow \mathcal{A}$
- Environment reacts with $\mu : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^*\times\mathcal{A} \rightsquigarrow \mathcal{O}\times\mathcal{R}$
- Performance of agent π in environment μ
  $=$ expected cumulative reward $= V_\mu^\pi := \mathbb{E}_\mu[\sum_{t=1}^\infty r_t^{\pi\mu}]$
- True environment μ unknown
  $\Rightarrow$ average over wide range of environments
- Ockham+Epicurus: Weigh each environment with its Kolmogorov complexity $K(\mu) := \min_p\{length(p) : U(p) = \mu\}$
- Universal intelligence of agent π is $\Upsilon(\pi) := \sum_\mu 2^{-K(\mu)} V_\mu^\pi$.
- Compare to our informal definition: Intelligence measures an agent's ability to perform well in a wide range of environments.
- AIXI $= \arg\max_\pi \Upsilon(\pi) =$ most intelligent agent.

Computational Issues: Universal Search

- Levin search: Fastest algorithm for inversion and optimization problems.
- Theoretical application:
  Assume somebody found a non-constructive proof of P=NP, then Levin-search is a polynomial time algorithm for every NP (complete) problem.
- Practical applications (J. Schmidhuber)
  Maze, towers of hanoi, robotics, ...
- FastPrg: The asymptotically fastest and shortest algorithm for all well-defined problems.

- AIXItl and ΦMDP: Computable variants of AIXI.
- Human Knowledge Compression Prize: (50'000€)

8 MORE APPLICATIONS OF AIT/KC

- Computer science: string matching, complexity/formal-language/automata theory
- Math: ∞ primes, quantitative Goedel incompleteness
- Physics: Boltzmann entropy, Maxwell daemon, reversible computing
- Operations research: universal search
- Others: Music, cognitive psychology, OCR

Literature

[LV07] M. Li and P. M. B. Vitányi.
Applications of Algorithmic Information Theory.
Scholarpedia, 2:5 (2007) 2658 [an extended encyclopedic entry]

[LV08] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, Berlin, 3rd edition, 2008.

[CV05] R. Cilibrasi and P. M. B. Vitányi. *Clustering by compression.* IEEE Trans. Information Theory, 51(4):1523–1545, 2005.

[RH11] S. Rathmanner and M. Hutter. A philosophical treatise of universal induction. *Entropy*, 16(6):1076–1136, 2011.

[Hut04] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

See also Advanced AI course COMP4620/COMP8620 @ ANU