Algorithmic Information Theory

A model for science?

Tom Sterkenburg

Centrum Wiskunde & Informatica

university of groningen

faculty of philosophy

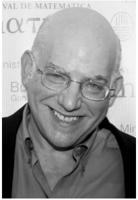
The Information Universe
Infoversum, Groningen
October 7, 2015

an objective measure of information content qua

compressibility

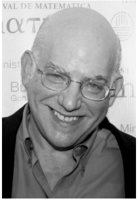
Chaitin (1966, 1969)

- “Understanding is compression”: the best theory is the one that best compresses the available data.



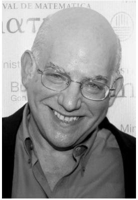
Chaitin (1966, 1969)

- "Understanding is compression": the best theory is the one that best compresses the available data.



- An instrumentalist view of scientific theorizing as *summarizing* observation data.

Chaitin (1966, 1969)

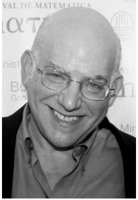


- "Understanding is compression": the best theory is the one that best compresses the available data.

- An instrumentalist view of scientific theorizing as *summarizing* observation data.
  - Summarizing/compressing is essentially a matter of finding patterns. The question is: what patterns?
  - The answer of AIT: all *effective* patterns.

Chaitin (1966, 1969)

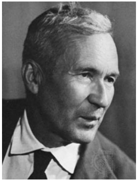
- "Understanding is compression": the best theory is the one that best compresses the available data.



- An instrumentalist view of scientific theorizing as *summarizing* observation data.
  - Summarizing/compressing is essentially a matter of finding patterns. The question is: what patterns?
  - The answer of AIT: all *effective* patterns.
- The promise of AIT is that it can *objectively quantify* compressibility.

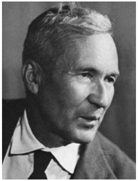
Kolmogorov (1965)

- Characterize *randomness* of data in terms of compressibility.



Kolmogorov (1965)

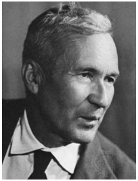
- Characterize *randomness* of data in terms of compressibility.



- Inspired by von Mises' project of giving a *frequentist* definition of probability.
  - Proposal: a sequence is random if it has no effective patterns.

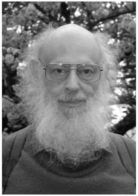
Kolmogorov (1965)

- Characterize *randomness* of data in terms of compressibility.



- Inspired by von Mises' project of giving a *frequentist* definition of probability.
  - Proposal: a sequence is random if it has no effective patterns.
- We will take a perspective on probability different from the frequentist one: namely, the *Bayesian* perspective.

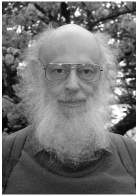
Solomonoff (1964)



- "Prediction by compression": the best method of prediction proceeds by compressing the available data.

Solomonoff (1964)

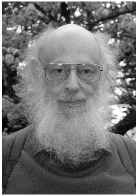
- "Prediction by compression": the best method of prediction proceeds by compressing the available data.



- Inspired by Carnap's project of delineating an "objective-logical" prediction function.
  - The promise of AIT is to give an objective prediction function that relies on an objective measure of compressibility.

Solomonoff (1964)

- "Prediction by compression": the best method of prediction proceeds by compressing the available data.



- Inspired by Carnap's project of delineating an "objective-logical" prediction function.
  - The promise of AIT is to give an objective prediction function that relies on an objective measure of compressibility.
- The promise of AIT is, in addition, that it can formally *prove* that prediction by compression is best.

The claim

- "The best method of prediction proceeds by compressing the available data."

The claim

- "The best method of prediction proceeds by compressing the available data."
  - Does predictive AIT show that prediction by compression is best?
  - Does predictive AIT give an objective definition of prediction by compression?

Sequential prediction

- Given a data sequence of bits, what bit comes next?
  - 0101010101010101010101010101010101010101..?
  - 0010010000111111011010101000100010000101..?
  - ...

Sequential prediction

- Given a data sequence of bits, what bit comes next?
  - 0101010101010101010101010101010101010101..?
  - 0010010000111111011010101000100010000101..?
  - ...
- Predictive AIT: assign higher probability to more compressible continuations.

Kolmogorov complexity

- Let C be a computer.
- Let σ be a data sequence, and σ^* be the shortest C-instruction for σ.
- Data sequence σ is more *compressible* as instruction σ^* is *shorter*.

Kolmogorov complexity

- Let C be a computer.
- Let σ be a data sequence, and σ^* be the shortest C-instruction for σ.
- Data sequence σ is more *compressible* as instruction σ^* is *shorter*.

Definition (Solomonoff, Kolmogorov, Chaitin)

The Kolmogorov complexity via C of data sequence σ is given by

$$K_C(\sigma) = \ell(\sigma^*).$$

Algorithmic probability

- ▷ Let C be a computer.
- ▷ Let σ be a data sequence, and σ^* be the shortest C-instruction for σ.
- ▷ Data sequence σ is more *compressible* as instruction σ^* is *shorter*.

Definition (Solomonoff, Kolmogorov, Chaitin)

The Kolmogorov complexity via C of data sequence σ is given by

$$K_C(\sigma) = \ell(\sigma^*).$$

Definition (Solomonoff, 1964)

The algorithmic probability via C of data sequence σ is given by

$$Q_C(\sigma) = \text{[imposing definition]} \approx 2^{-K(\sigma)}.$$

Algorithmic probability

- Let C be a computer.
- Let σ be a data sequence, and σ^* be the shortest C-instruction for σ.
- Data sequence σ is more *compressible* as instruction σ^* is *shorter*.

Definition (Solomonoff, Kolmogorov, Chaitin)

The Kolmogorov complexity via C of data sequence σ is given by

$$K_C(\sigma) = \ell(\sigma^*).$$

Definition (Solomonoff, 1964)

The algorithmic probability via C of data sequence σ is given by

$$Q_C(\sigma) = [\text{imposing definition}] \approx 2^{-K(\sigma)}.$$

- Assign higher probability to more compressible data.

Predictive AIT

- Assign higher probability to more compressible continuations.

Predictive AIT

- Assign higher probability to more compressible continuations.

- An algorithmic probability distribution gives a prediction method.
    - Let's refer to the algorithmic probability distributions as the **predictors of type** $\mathcal{Q}$.

Predictive AIT

- Assign higher probability to more compressible continuations.

- An algorithmic probability distribution gives a prediction method.
  - Let's refer to the algorithmic probability distributions as the **predictors of type** $\mathcal{Q}$.
- Moreover, we can prove that these predictors are good.

Theorem (Solomonoff, 1978)

If effective P^* is the true data-generating source, then the predictions of Q_C converge with P^*-probability 1 to the true probability values.

The claim, again

- "The best method of prediction proceeds by compressing the available data."
  - Does predictive AIT show that prediction by compression is best?
  - Does predictive AIT give an objective definition of prediction by compression?

A different pair of glasses

- Let's reconsider things from the perspective of the **Bayesian** model of scientific inference.

Bayesian prediction (1)

- Select (countable) *hypothesis class* $\mathcal{H}$.
  - The patterns in the data we hold possible.

Bayesian prediction (1)

- Select (countable) *hypothesis class* $\mathcal{H}$.
  - The patterns in the data we hold possible.
- Define *prior distribution* W over $\mathcal{H}$.
  - The weights we assign to the possible patterns.

Bayesian prediction (1)

- Select (countable) *hypothesis class* $\mathcal{H}$.
  - The patterns in the data we hold possible.
- Define *prior distribution* W over $\mathcal{H}$.
  - The weights we assign to the possible patterns.
- Our choice of hypothesis class plus prior embodies our *inductive assumptions*.

Bayesian prediction (2)

- A Bayesian distribution is a prior-weighted mean over all hypotheses:

Definition

The Bayesian distribution $P_W^{\mathcal{H}}$ via prior W on hypothesis class $\mathcal{H}$ is given by

$$P_W^{\mathcal{H}}(\sigma) := \sum_{P \in \mathcal{H}} W(P)P(\sigma).$$

- This gives a Bayesian *prediction method*.

Bayesian prediction (2)

- A Bayesian distribution is a prior-weighted mean over all hypotheses:

Definition

The Bayesian distribution $P_W^{\mathcal{H}}$ via prior W on hypothesis class $\mathcal{H}$ is given by

$$P_W^{\mathcal{H}}(\sigma) := \sum_{P \in \mathcal{H}} W(P)P(\sigma).$$

  - This gives a Bayesian *prediction method*.

- Moreover, we can prove that Bayesian predictors are *consistent*:

Theorem

If $P^* \in \mathcal{H}$ is the true data-generating source, then the predictions of $P_W^{\mathcal{H}}$ converge with P^*-probability 1 to the true probability values.

A very general inductive assumption

- Consider the hypothesis class $\mathcal{H}^{\text{eff}}$ of computably approximable or simply *effective* hypotheses (Levin, 1970).

A very general inductive assumption

- Consider the hypothesis class $\mathcal{H}^{\text{eff}}$ of computably approximable or simply *effective* hypotheses (Levin, 1970).
- The Bayesian predictors with an effective prior over this hypothesis class are the predictors *operating under the inductive assumption of effectiveness*.
  - We'll call these the **predictors of type** $\mathcal{R}$.

It's the same thing!

- The predictors of type $\mathcal{Q}$ and the predictors of type $\mathcal{R}$ **are the same**.

It's the same thing!

- The predictors of type $\mathcal{Q}$ and the predictors of type $\mathcal{R}$ **are the same**.

Theorem (Wood, Sunehag and Hutter, 2013)

$$\mathcal{Q} = \mathcal{R}.$$

- ▷ The choice of computer is the choice of effective Bayesian prior over $\mathcal{H}^{\text{eff}}$.

It's the same thing!

- The predictors of type $\mathcal{Q}$ and the predictors of type $\mathcal{R}$ **are the same**.

Theorem (Wood, Sunehag and Hutter, 2013)

$$\mathcal{Q} = \mathcal{R}.$$

  - The choice of computer is the choice of effective Bayesian prior over $\mathcal{H}^{\text{eff}}$.

- **Predictive AIT is prediction under the inductive assumption of effectiveness.**

Back to the beginning

- "The best method of prediction proceeds by compressing the available data."

Back to the beginning

- "The best method of prediction proceeds by compressing the available data."
  - Does predictive AIT show that prediction by compression is best?
  - Does predictive AIT give an objective definition of prediction by compression?

Relativity and subjectivity

- **Does predictive AIT give an objective definition of prediction by compression?**

Relativity and subjectivity

- **Does predictive AIT give an objective definition of prediction by compression?**
- The choice of computer = prior: how to weigh each pattern?
  - An inevitable element of subjectivity.

Relativity and subjectivity

- **Does predictive AIT give an objective definition of prediction by compression?**
- The choice of computer = prior: how to weigh each pattern?
  - ▷ An inevitable element of subjectivity.
- The choice of hypothesis class: what class of patterns?
  - ▷ The effective patterns – but what level of effectiveness?

$$
\begin{tikzcd}
 & \boldsymbol{\Sigma}^0_1 \arrow[dr] & & \boldsymbol{\Sigma}^0_2 & \cdots \\
\boldsymbol{\Delta}^0_1 \arrow[ur] \arrow[dr] & & \boldsymbol{\Delta}^0_2 \arrow[ur] \arrow[dr] & & \cdots \\
 & \boldsymbol{\Pi}^0_1 \arrow[ur] & & \boldsymbol{\Pi}^0_2 & \cdots
\end{tikzcd}
$$

What does the convergence theorem show? (1)

- **Does predictive AIT show that prediction by compression is best?**

What does the convergence theorem show? (1)

- **Does predictive AIT show that prediction by compression is best?**
- Does it show that prediction by compression converges to the truth?

What does the convergence theorem show? (1)

- **Does predictive AIT show that prediction by compression is best?**
- Does it show that prediction by compression converges to the truth?
  - The convergence proof of predictive AIT is an instance of Bayesian consistency.
  - It only shows convergence to the truth insofar as the hypothesis class $\mathcal{H}$ contains the true data-generating source.

What does the convergence theorem show? (1)

- **Does predictive AIT show that prediction by compression is best?**
- Does it show that prediction by compression converges to the truth?
  - The convergence proof of predictive AIT is an instance of Bayesian consistency.
  - It only shows convergence to the truth insofar as the hypothesis class $\mathcal{H}$ contains the true data-generating source.
  - In our case, insofar as the true data-generating source is effective.
  - This commits us to an assumption of effectiveness on the world.

What does the convergence theorem show? (2)

- **Does predictive AIT show that prediction by compression is best?**
- Could it show that prediction by compression is best in a different sense?

What does the convergence theorem show? (2)

- **Does predictive AIT show that prediction by compression is best?**
- Could it show that prediction by compression is best in a different sense?
  - The best *we* (with our limited epistemic means) could ever do.

What does the convergence theorem show? (2)

- **Does predictive AIT show that prediction by compression is best?**
- Could it show that prediction by compression is best in a different sense?
  - The best *we* (with our limited epistemic means) could ever do.
  - Effectiveness seems a reasonable outer constraint on our own predictive capabilities.
  - So, prediction by compression might thus be shown to be *optimal*: at least as good as any possible prediction method.

What does the convergence theorem show? (2)

- **Does predictive AIT show that prediction by compression is best?**
- Could it show that prediction by compression is best in a different sense?
  - The best *we* (with our limited epistemic means) could ever do.
  - Effectiveness seems a reasonable outer constraint on our own predictive capabilities.
  - So, prediction by compression might thus be shown to be *optimal*: at least as good as any possible prediction method.
  - Unfortunately, this also doesn't quite work, no matter the exact level of effectiveness.

$$
\begin{tikzcd}
 & \mathbf{\Sigma}^0_1 \arrow[dr] & & \mathbf{\Sigma}^0_2 & \cdots \\
\mathbf{\Delta}^0_1 \arrow[ur] \arrow[dr] & & \mathbf{\Delta}^0_2 \arrow[ur] \arrow[dr] & & \cdots \\
 & \mathbf{\Pi}^0_1 \arrow[ur] & & \mathbf{\Pi}^0_2 & \cdots
\end{tikzcd}
$$

To conclude

Algorithmic information theory gives an idealized model of reasoning under a particular assumption of *effectiveness*.

To conclude

Algorithmic information theory gives an idealized model of reasoning under a particular assumption of *effectiveness*.

Indeed, predictive AIT is an instance of Bayesian reasoning with this particular assumption of effectiveness. This clearly reveals the subjectivity involved in the choice of prior and the relativity involved in the choice of hypothesis class – the inductive assumption.

To conclude

> Algorithmic information theory gives an idealized model of reasoning under a particular assumption of *effectiveness*.

Indeed, predictive AIT is an instance of Bayesian reasoning with this particular assumption of effectiveness. This clearly reveals the subjectivity involved in the choice of prior and the relativity involved in the choice of hypothesis class – the inductive assumption.

The interest of AIT thus lies mainly in the interest of effectiveness as an inductive assumption. While effectiveness looks interesting as a highly general and unrestrictive assumption, the prospects for showing that this is somehow the best assumption look bleak.

To conclude

Algorithmic information theory gives an idealized model of reasoning under a particular assumption of *effectiveness*.

Indeed, predictive AIT is an instance of Bayesian reasoning with this particular assumption of effectiveness. This clearly reveals the subjectivity involved in the choice of prior and the relativity involved in the choice of hypothesis class – the inductive assumption.

The interest of AIT thus lies mainly in the interest of effectiveness as an inductive assumption. While effectiveness looks interesting as a highly general and unrestrictive assumption, the prospects for showing that this is somehow the best assumption look bleak.

www.cwi.nl/~tom