Five years of Facticity research

Pieter Adriaans, Peter Bloem, Steven de Rooij

Ivl, SNE group Universiteit van Amsterdam

COMMIT/ A public-private research community

Info-Metrics Institute
College of Arts and Sciences, American University

eScience center
Research cycle

- System
- Observations
- Processing
- Model

[Images of a planet, a portrait, a historical figure, and a pie chart]
Facticity

• Objective: Studies the balance between structural and ad hoc data in (large) data sets.
• Methodology: Kolmogorov complexity.
• Model class: partial recursive functions (all possible computer programs).
• Most general theory of model induction we currently have.
Turing two-part code optimal compression

Data $x$

Index

Program

$K(x) = \min\{|ip| : U(ip) = x\}$

Kolmogorov complexity

Facticity

Residual Entropy

Randomness Deficiency
Turing two-part code optimal compression

\[ |x| \]

\[ K(x) \]

\[ (x) \]
Two problems (and their solutions)

1) Kolmogorov complexity is not computable?
   
   *A Safe Approximation for Kolmogorov Complexity, P Bloem, F Mota, S de Rooij, L Antunes, P Adriaans - Algorithmic Learning Theory, 2014*

2) Are indexes of TM’s indexes a reliable measure for model complexity (Nickname Problem)?
   
   *Invariance theorem for indexes (Bloem, de Rooij 2014): there is a compact acceptable enumeration of the partial recursive functions*
Facticity of random strings

$K(x) \mid x \mid$
Facticity for compressible datasets with small $K(x)$
Facticity for large compressible sets: Collapse onto smallest Universal Machine
Collapse point Conjecture

- If we have a data set D and a Universal Turing machine $U_u$ such that:

$$\log(K(D)) > |u|$$

then $U_u$ will be the preferred model under facticity.

Realistic example: If $|u| = 20$ then all models above $10^6$ bits will collapse
Conclusions Further research

• Compression based automated model selection for unique strings? No!
• Is Big Data just more Simple data? No!
• What is a really complex system? Collapse Conjecture: model collapse phenomena are relevant for the study of complex systems in the real world. (Brain, Cell, Stock Exchange etc.)
• Wat are viable restrictions on model classes?