Information Transformation

An Underpinning Theory for Software Engineering

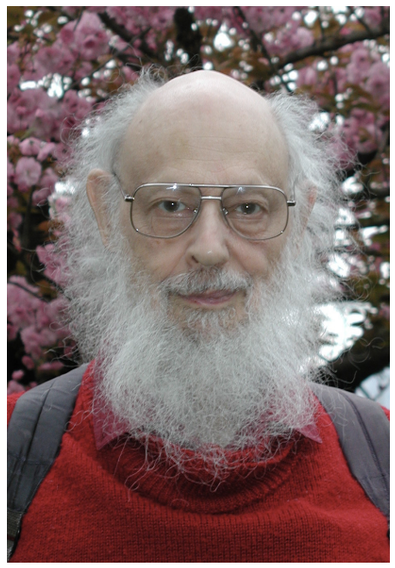
David Clark, Robert Feldt, Simon Poulson, Shin Yoo

Shannon Entropy

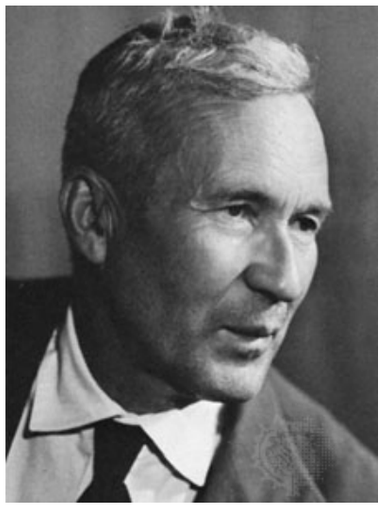
$$\mathcal{H}(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

randomness of a
random variable

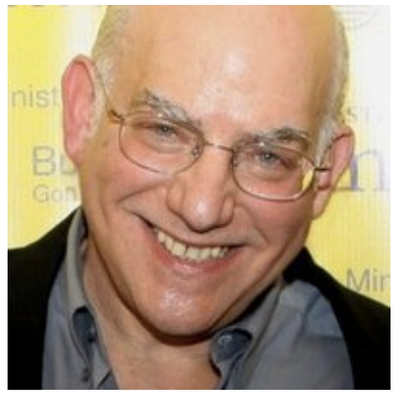
Kolmogorov Complexity



Solomonoff



Kolmogorov



Chaitin

The length of the shortest program that can produce a given string from no inputs

randomness of a string

source  transmitted

expressions
statements
paths
programs

Channel Capacity of the program's I/O channel

$$\max_{\sigma \in \Sigma_{\mathcal{I}}} \mathcal{M}(\mathcal{I}; \mathcal{O})$$

$$\max_{\sigma \in \Sigma_{\mathcal{I}}} \mathcal{H}(\mathcal{O}) = \log_2(|\mathcal{O}|)$$

$\mathcal{H}(f^{-1}o)$

$f^{-1}o$

f

o

$p(o)$

Test Suite Selection

Output Diversity

1. Generate random inputs
2. Discard inputs if they lead to discovered outputs
3. Repeat until too hard to find new test inputs

High Pearson correlation with statement, branch and path coverage but 47% improvement in bug finding over these.

Alshawan and Harman, ICSE 2012, ISSTA 2014

Conditional Entropy of the program's I/O channel

$$\mathcal{H}(\mathcal{I}|\mathcal{O}) = \mathcal{H}(\mathcal{I},\mathcal{O}) - \mathcal{H}(\mathcal{O}) = \mathcal{H}(\mathcal{I}) - \mathcal{H}(\mathcal{O})$$

- Information lost through program execution

$\mathcal{H}(f^{-1}o)$

$f^{-1}o$

f

o

$p(o)$

Intended

t

P

A

C

pp

Q

B

o

Unintended

t

P''

A'

$[Q](\llbracket Q \rrbracket \Sigma_{pp'})$

$\Sigma_{pp'}$

C'

pp'

Q

B'

$\llbracket Q \rrbracket \Sigma_{pp'}$

o

Coincidental Correctness in test suites

High correlation between failed error propagation to oracles and the conditional entropy of the program region between the bug and the program point.

Androutsopoulos, Clark, Dan, Hierons and Harman. ICSE 2014

Laws?

- A test suite is I/O channel adequate if it achieves the channel capacity of the program

- The conditional entropy of the I/O channel of the program is a measure of how difficult it is to test

- e.g. ID and ML

Ideas

- using conditional entropy to optimise oracle placement
- using Kolmogorov complexity to measure test suite diversity
- using Kolmogorov complexity to measure program cohesion
- using channel capacity to measure coupling between program components
- using channel capacity to estimate path feasibility
- using Kolmogorov complexity to measure software evolution
- using entropy to measure stability of test cases *Gao et alia, ICSE 2015*