Stochastic Process Model-Log Quality Dimensions

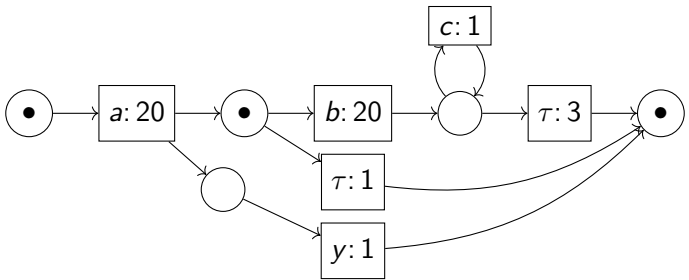
an experimental study

Adam Burke, *Sander Leemans*, Moe Wynn,
Wil van der Aalst and Arthur ter Hofstede

Stochastic Models

- Event logs are stochastic
  e.g. the log
  $[\langle a, b\rangle^{20}, \langle a, b, c\rangle^{2}, \langle a, b, c, c\rangle^{1}, \langle e, f\rangle^{1}]$
  ... already has frequency information
- Control-flow models discard stochastic information
- Stochastic process models retain stochastic information
- Simulation, analysis and recommendation need stochastic information

A Stochastic Model



How to compare?

What dimensions describe the quality of stochastic process models?

Stochastic Conformance Checking Measures

Exploration measures (13 new)

- Earth Movers' trace-wise (1)
- Probability mass (2)
- Fitness (6)
- Precision (2)
- Simplicity (3)
- Generalisation (4)

Discover dimensions

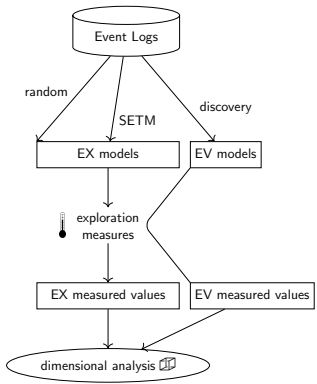
Evaluation measures

- Earth-Movers' Stochastic Conformance
- Entropy recall
- Entropy precision

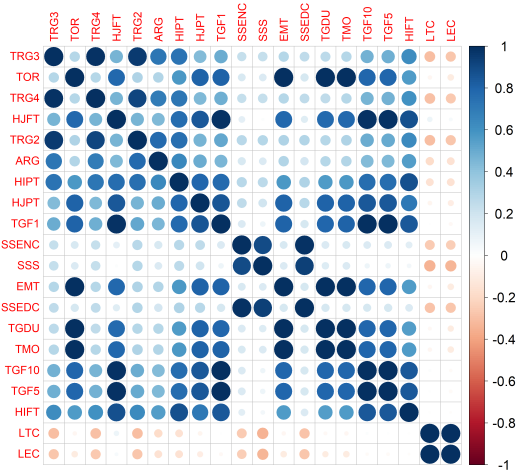
Identify dimensions

Discovering the Dimensions

1. Use 6 public logs
2. 9301 stochastic process models
   random, new genetic algorithm & discovered
3. 18 exploration measures
4. Dimensional analysis



Dimensional Analysis 1: Correlations



- Baselines: 2 log-only measures
- Remove: 3 too-correlated measures:
  Trace Overlap Ratio,
  Trace Generalization Floor-1,
  Trace Generalization Floor-10

Dimensional Analysis 2: Principal Component Analysis

- Find linear relation that best describes the data
- Find linear relation that best describes the data, orthogonal to first relation
- . . . (15 times)

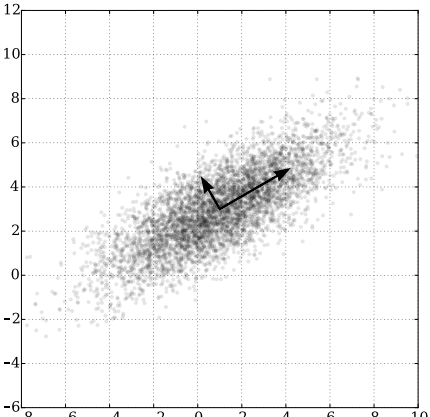
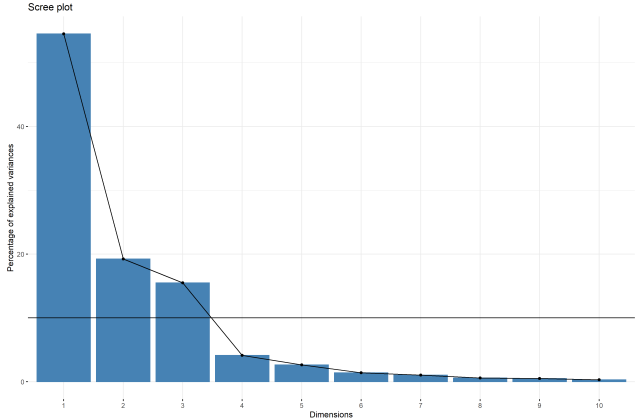


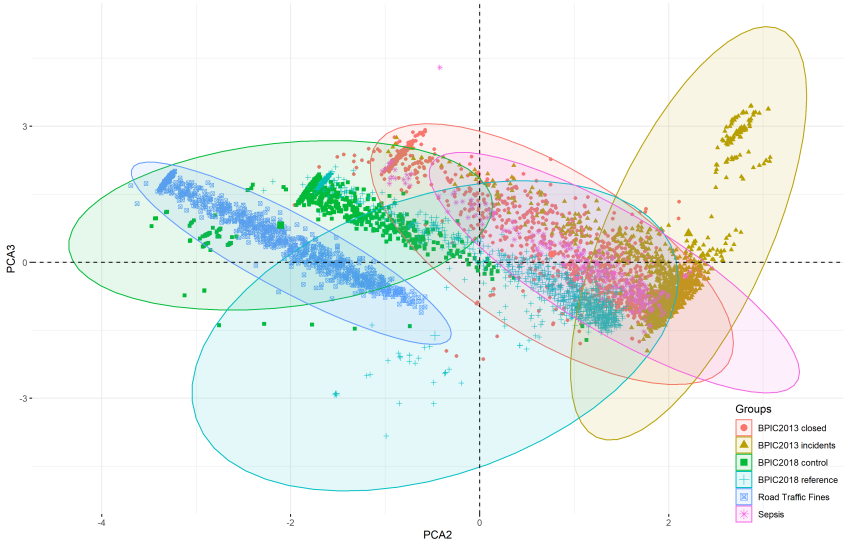
image source: wikipedia

Dimensional Analysis 2: Principal Component Analysis

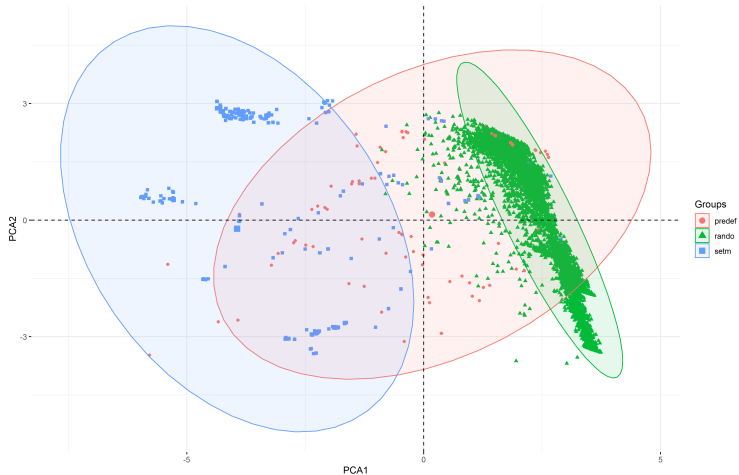
- 15 linear combinations of measures
- Scree plot: we choose 3



Principal Components - Variation By Log

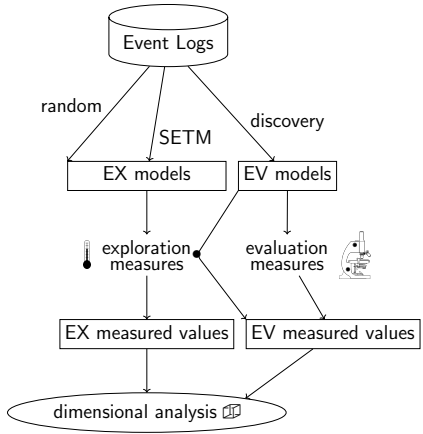


Principal Components - By Model Generator



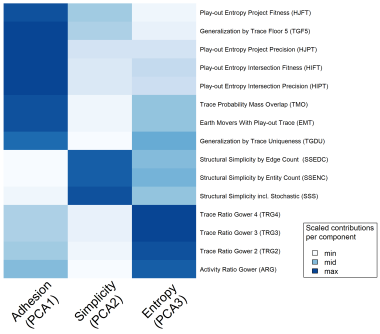
Identifying the Dimensions

- Remove random & genetic models
- Add the 3 evaluation measures *on EV models only*
- Redo principal component analysis

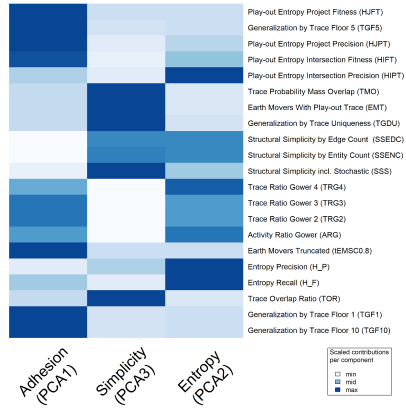


Identifying the Dimensions: comparison

Discovered dimensions



Identified dimensions

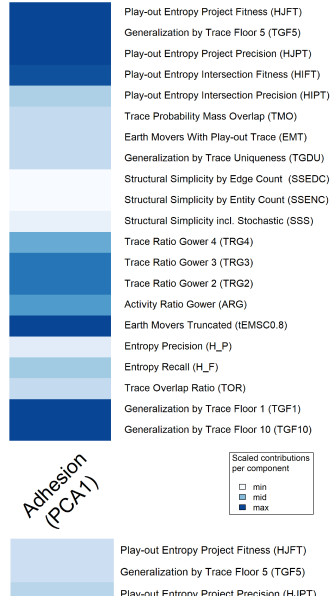


Play-out Entropy Project Fitness (HJFT)
Generalization by Trace Floor 5 (TGF5)
Play-out Entropy Project Precision (HJPT)
Play-out Entropy Intersection Fitness (HIFT)
Play-out Entropy Intersection Precision (HIPT)
Trace Probability Mass Overlap (TMO)
Earth Movers With Play-out Trace (EMT)
Generalization by Trace Uniqueness (TGDU)
Structural Simplicity by Edge Count (SSEDC)

Three Empirical Dimensions

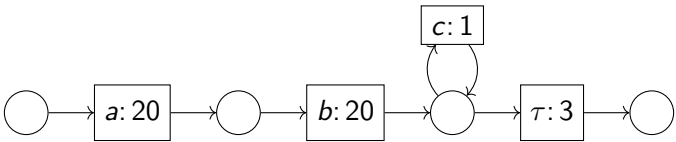
- Adhesion
  How little effort is required to transform one stochastic language into another
- Entropy
  The amount of information in a system

  In this case, the combination of log and model
- Simplicity
  Structural simplicity of the model



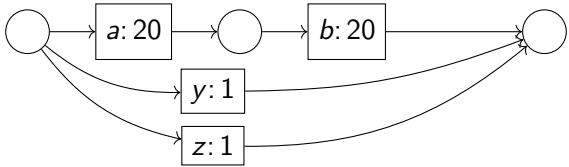
Example 1

Adhesion + entropy + simplicity +



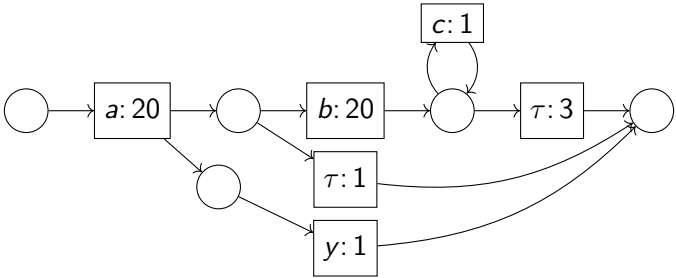
Example 2

Adhesion + entropy - simplicity +



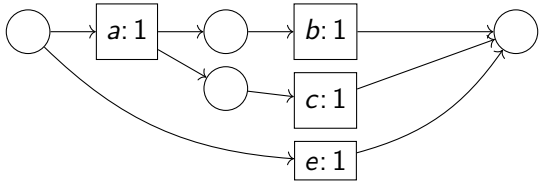
Example 3

Adhesion $\sim$ entropy $\sim$ simplicity -



Example 4

Adhesion - entropy - simplicity $\sim$



Limitations

- First models are process trees $\rightarrow$ representational bias
- SETM evolutionary fitness function may tend to correlate measures
  - Robustness tests excluding SETM still show the effect, though
- Largest log 200 000 traces

Conclusion / Future Work

- Three empirically derived dimensions
- Focus on empirical and orthogonality
- Other measures and principles may be *non-orthogonal* but still *useful*, eg recall and precision entropy measures
- Future work
  - Theoretical grounded measures for these dimensions
  - Further tests