Natural Computing
Research & Applications
Group

When The Means Justifies the End: Why We Must Evaluate on More than Mere Output

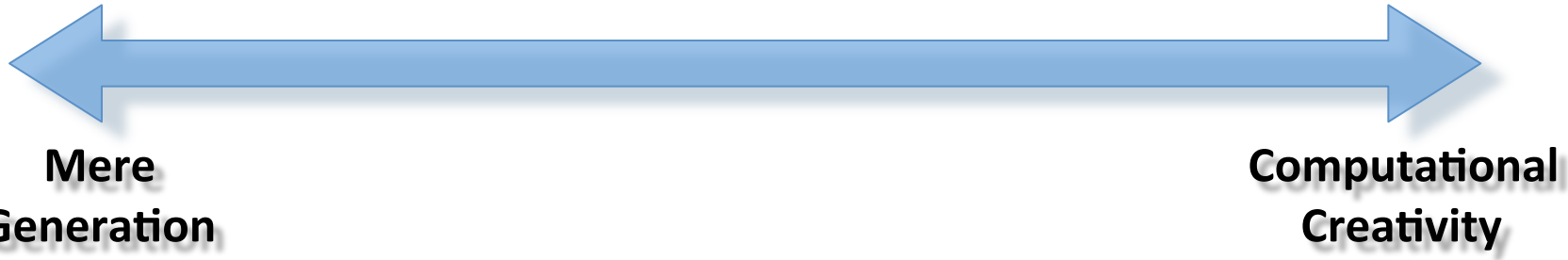
Róisín Loughran

HORSE16, Queen Mary University, London

September 19th 2016

Motivation

- Evolutionary Composition
  - Requires internal and external evaluation
- How to evaluate final system?

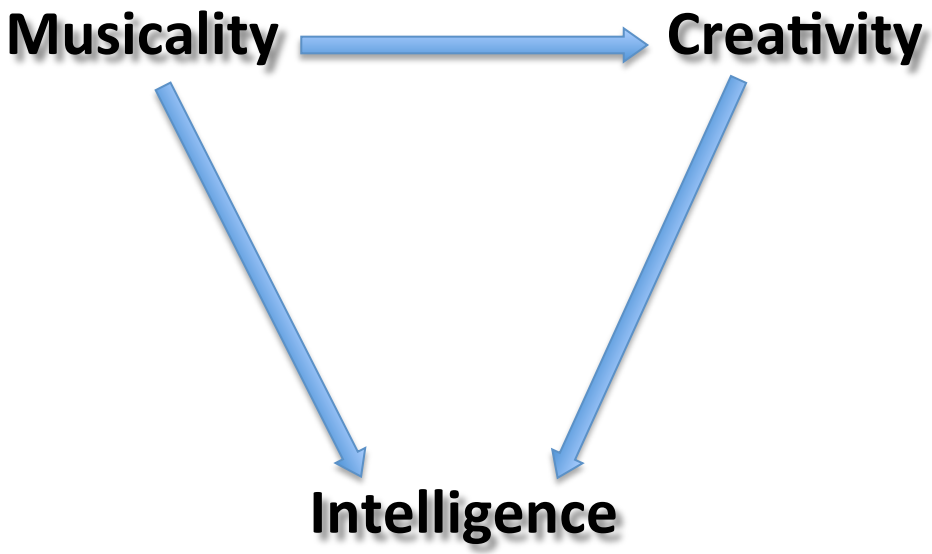


System vs. Output

Jumpy Dog Art: https://www.youtube.com/watch?v=PiF_pmLCtQM

Musicality, Creativity, Intelligence

- Does Musicality -> Creativity?
- Does Creativity -> Intelligence?



- Does this imply Musical Intelligence?
  - Can this be measured?

Musical Intelligence

- Musical Intelligence – one of the 8 identified modes of intelligence (Gardner)
- Music, Intelligence & Artificiality (Marsden, 2000)

- To be 'Musical' implies
  - Ability -> talent
  - Learning -> training
    - Knowledge
  - Intelligence?

- How could this be measured?

- Should a generative music system display Intelligence?
  - Does this *MI* require AI?

Generative Music

- What does a generative system actually create
  - Notation
  - Audio
  - Theory
  - Representation
  - Music?

'Music, in its own right, does not exist' (Wiggins, 2010)

- How can we evaluate it?
  - Are we left with subjective evaluation?

Musical Computational Creativity

'The philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative ' (Colton & Wiggins, 2012)

- Creativity still difficult to define/distinguish
- H-creativity vs. P-creativity
  - Dependent on field of study
- 3 ways computers can create new ideas (Boden):
  - Combining novel ideas
  - Exploring the limits of conceptual space
  - Transforming established ideas that enable the emergence of unknown ideas

Why do we Limit to 'Human'?

- Generative music often evaluated:
  - Similarity to style
  - Turing-esque tests
  - Musical measures
  - Crowd sourcing

- Two Distinct issues:
  1. Creativity (musical, computational or otherwise) is hard to define and hence hard to evaluate
  2. There is a persistent, limiting tendency to evaluate creative artifacts, systems or results purely using human opinion

"The ultimate vindication of AI-creativity would be a program that generated novel ideas which initially perplexed or even repelled us, but which was able to persuade us that they were indeed valuable" (Boden)

Alternative Evaluation

Creativity

- Lovelace test
  - Agent ***A***, output ***o*** human architect ***H***
  - If ***H*** cannot explain how ***A*** produced ***o*** -> test passed.
- Evaluation Frameworks:
  - Empirical Criteria (Ritchie 2001, 2007)
  - Creative Tripod (Colton, 2007)
  - CCT - FACE/IDEA descriptive models (Pease & Colton, 2011)
  - SPECS (Jordanous, 2012)

Evolutionary Composition

- Evaluation vs. fitness
- Alternative criteria?
  - Defensibility: Choose one item over another from a logical system of comparing between items and determining a decisive preference
  - Sociability: Similarity between agents repertoires
  - Combination of metrics?

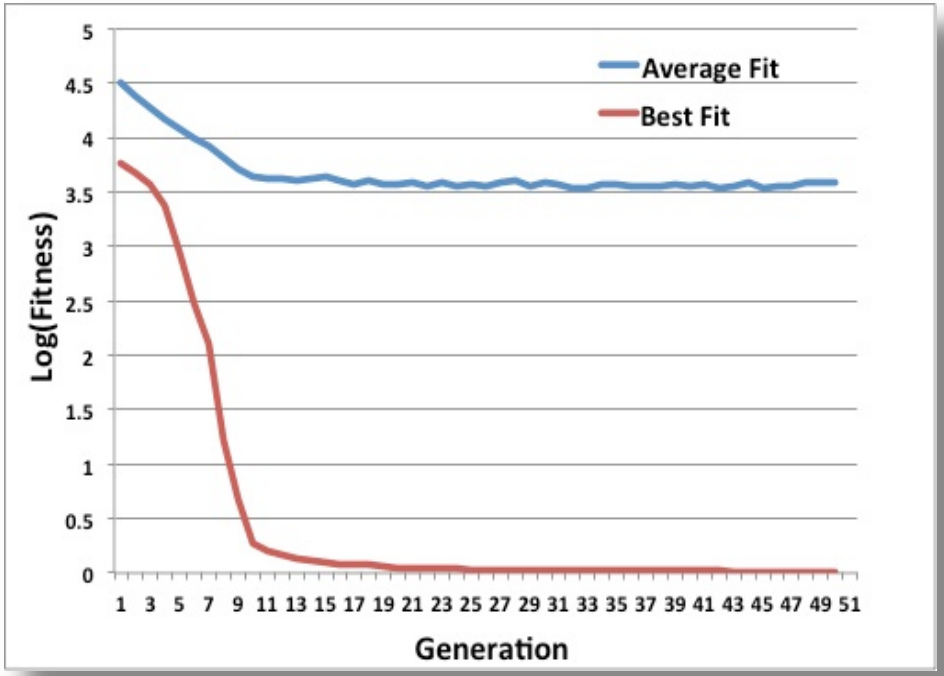
Musical Horses

- Generative Music Systems: How can we detect Horses?
  - Figure of Merit
    - Evaluation
  - Irrelevant Transformations
    - If we only evaluate on musical output -> cannot establish a 'ground truth'
      - Cannot detect and therefore cannot avoid a Horse

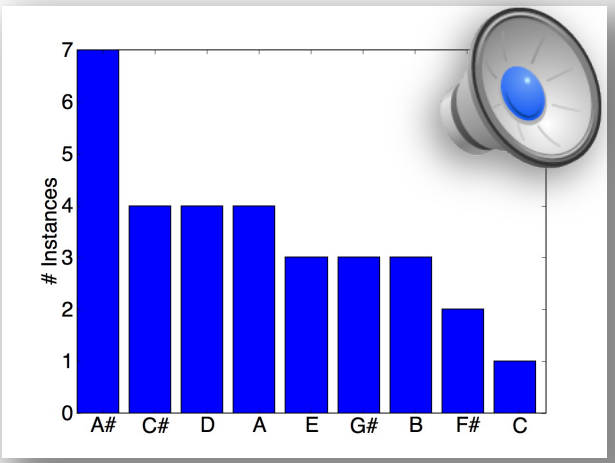
- Can we detect a Horse with system evaluation?

Musical Horses: The Composing Pony

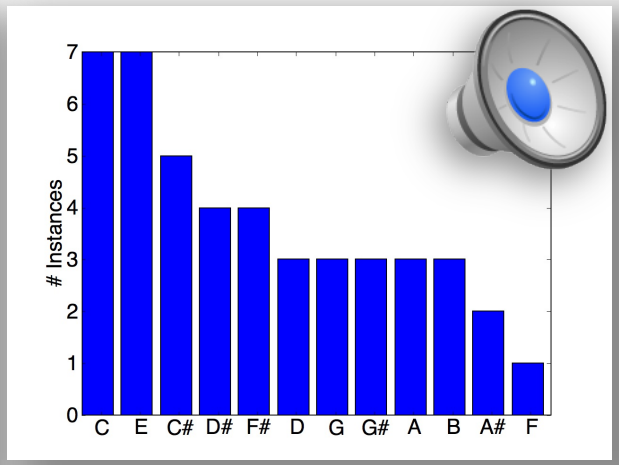
- Average vs. best fitness for 30 runs on a population of 100 over 50 generations:



Musical Horses: The Composing Pony



Phenotype length = 200
Clear primary pitch
Only 9 pitches used
Top7 = 81%, Top9 = 100%
Fitness = 1 (perfect)



Phenotype length = 325
2 equally strong pitches
All 12 pitches played
Top7 = 67%, Top9 = 80%
Fitness = 34,330

(Loughran et al, CEC 2015)

Musical Horse: Bet on it!

*"a 'horse' is just a system that is not actually addressing the problem **it appears to be solving**" (Sturm, 2015)*

- Problem: Generative Music System
  - Subjective output
  - Evaluation methods in development

- Unless the system has a very precise goal and criteria, it is likely that it could be classified as a 'Horse'
  - Precise experimentation design is vital

Conclusions

- Can we detect a Horse in a generative music system?
  - Possibly!
- Evaluation of any creatively tasked ML system is dependent a clearly focused intent and experimental design
  - 'ask the right question' (Wiggins)
- A 'Horse' is not necessarily a failure
  - Painting dog may have more practical application
  - Learn from the process developed and not merely the output produced

'Science is not about building a body of known 'facts'. It is a method for asking awkward questions and subjecting them to a reality-check, thus avoiding the human tendency to believe whatever makes us feel good'

(Pratchett, SoD)

Acknowledgements

This work is part of the App'Ed (Applications of Evolutionary Design) project funded by Science Foundation Ireland under grant 13/IA/1850

Melodies and technical papers available at:
https://loughranroisin.wordpress.com/about/