Assessing the Performance of Creative Systems

Graeme Ritchie
University of Aberdeen

Isn’t this just testing of software?

No!

Computational creativity (CC) has particular characteristics.

THIS TALK

- PART I : Introduction
  - What’s different about creative systems?
  - Two meanings of “creative”.
  - Two notions of “evaluation”.
  - Two types of objective in CC.
- PART II : Evaluation
  - Some ways to organise an evaluation.
- PART III : Other aspects
  - Speculations about wider issues.

PART Ⅰ : Introduction

WHAT IS DIFFERENT ABOUT CC? *(COMPARED TO SOFTWARE ENGINEERING)*

- No practical task
- No prior specification of “requirements”
- No precise definition of correctness
- No objective measure of success
- Different criteria for what counts as a “good” method of implementation

Why a creative system is different

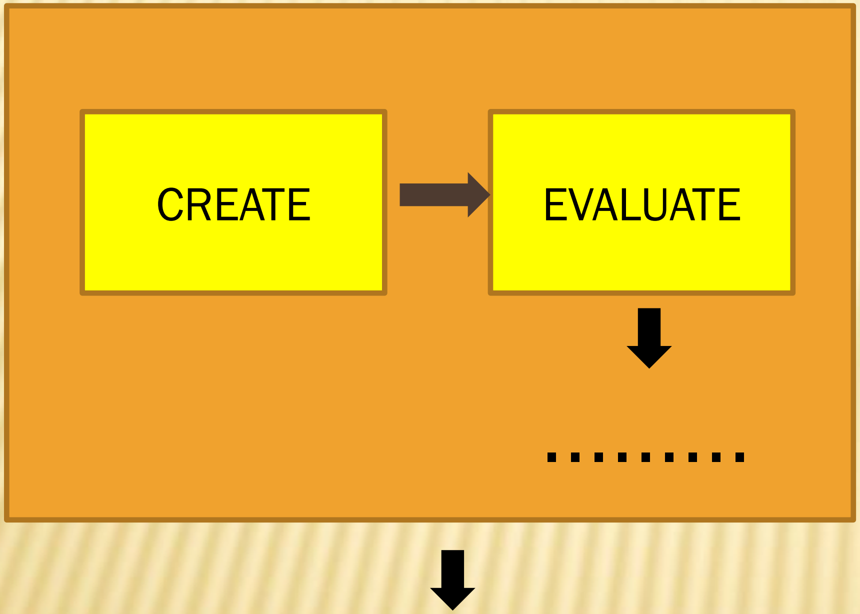
- Offers a model of some activity within society and culture.
- Embodies hypotheses about how that activity works.
- Helps us to assess these models & hypotheses.

Why bother with evaluation?

- To see if we are making any progress.
- To see which methods/models give the best results.
- To test any hypotheses (against reality).

INTERNAL (EMBEDDED) EVALUATION

Some systems contain a module which “evaluates” their creations.



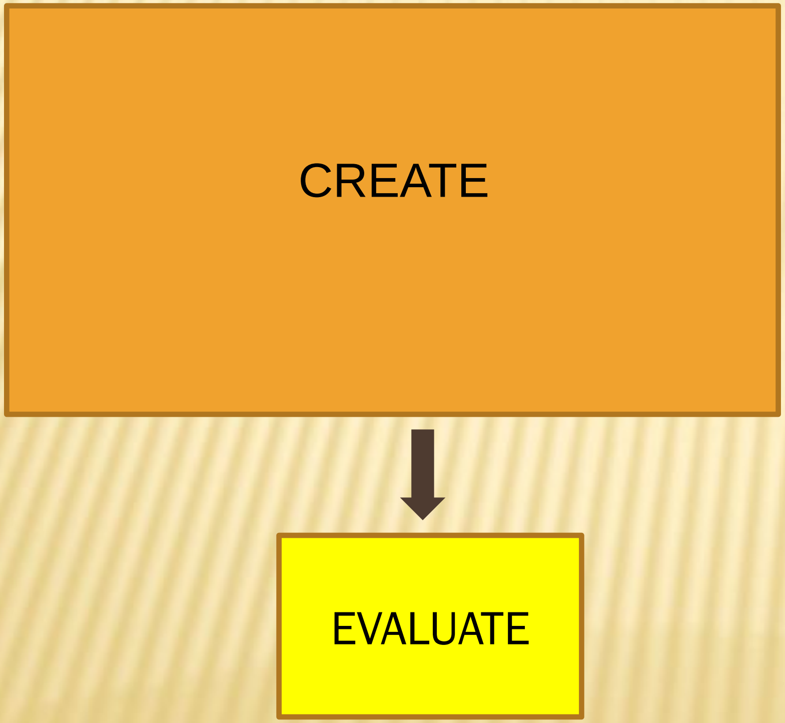
*THIS IS **NOT** WHAT WE ARE TALKING ABOUT TODAY!*

✖

EXTERNAL (OVERALL) EVALUATION

We need to assess how well the creative system has performed.

THIS ***IS*** *WHAT WE ARE TALKING ABOUT TODAY!*



CREATE

EVALUATE

What is meant by the term “creative system”?

Software which exhibits “behaviours that would be deemed creative if exhibited by a person”.

Two common usages of “creative”

“loose” version

(a) an activity (e.g. painting, musical composition, writing poetry) which is regarded as inherently “creative”

(b) displaying particular skill or artistry or ingenuity

“strict” version

creative$_{L}$ vs. **creative**$_{S}$

A DIFFICULT QUESTION SOMETIMES ASKED:

- “Is this creative$_L$ system being creative$_s$?”

POSSIBLE AIMS FOR CC RESEARCH

- “**Weak**” objectives: Modelling some (creative$_L$) domain (e.g. visual art)
  - *Engineering*: create artefacts for some purpose

OR

  - *Science*: better understand that domain
- “**Strong**” objectives: Modelling creativity$_S$ – getting the computer to be “genuinely creative”.

SUCCESS WITH “WEAK” OBJECTIVES

- Model is about **the domain**.
- Main interest: how **accurate** is this?
- Hence: is the output (a) **well-formed** (b) of **good quality**?

SUCCESS WITH “STRONG” OBJECTIVES

- Model/hypothesis is about **how to be creative$_S$**.
- Main interest : are outputs being produced **in a way which is creative$_S$** ?

THIS TALK

- PART I : Introduction
- PART II : Evaluation
  - Two aims in evaluation.
  - Generating things to evaluate : 5 approaches.
  - Measuring.
- PART III : Other aspects

PART II - Evaluation

DIFFERENT NOTIONS OF “TESTING”

- **Debugging:** informal, to find bugs, etc.
- **Formative evaluation:** organised and systematic tests to check progress during development.
- **Summative evaluation:** controlled assessment of the success of the final version of the system.

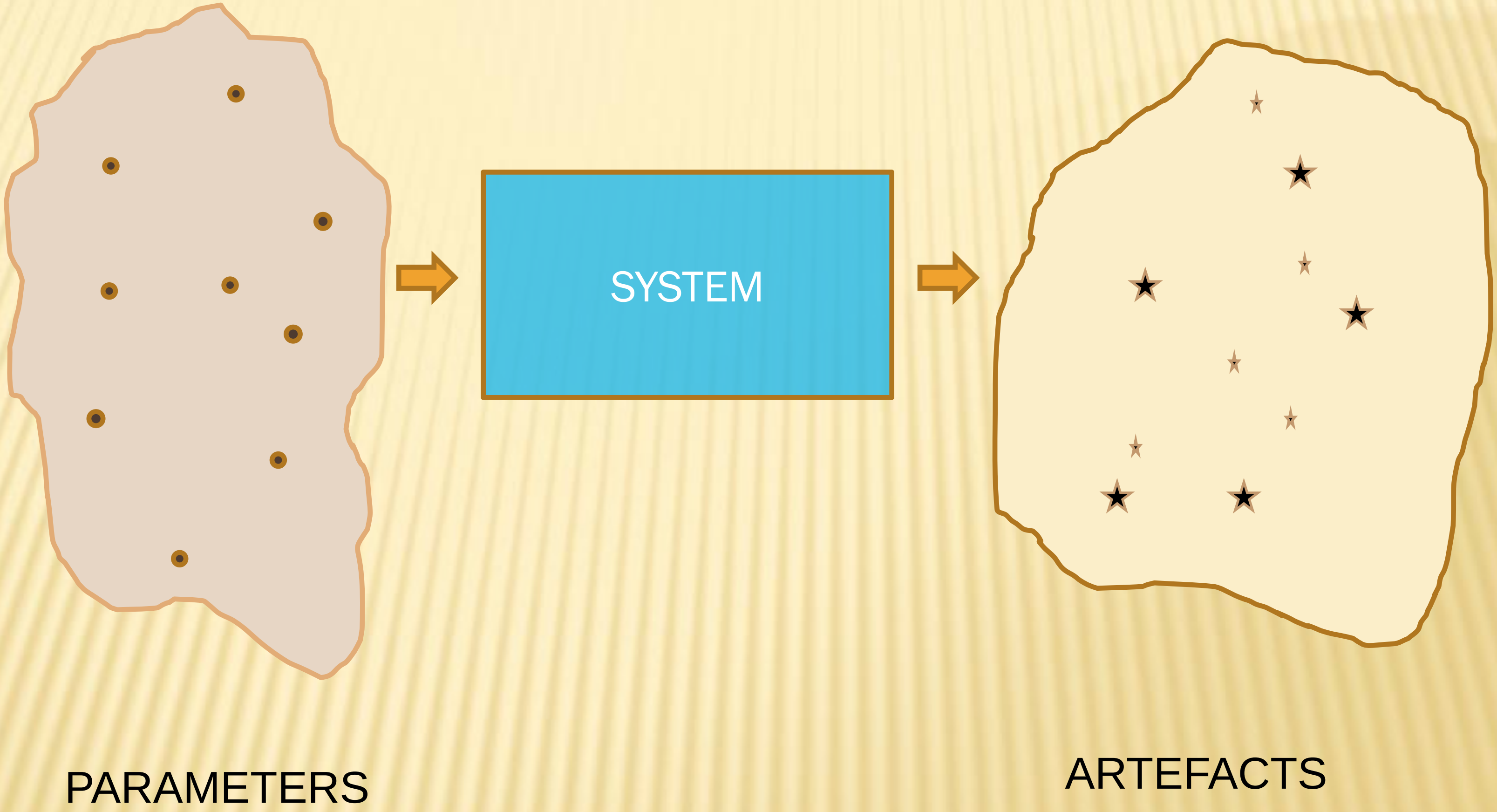
METHODS DEPEND ON WHAT YOU ARE DOING

STAGE → ------------------------------------- OBJECTIVES ↓	FORMATIVE (progress checking)	SUMMATIVE (final assessment)
WEAK (modelling domain)		
STRONG (modelling creativity)		

FACETS OF EVALUATION

- Collecting output items:
  - generating
  - selecting
- What to measure
- How to measure

HOW SHOULD ITEMS BE GENERATED?



SYSTEM

PARAMETERS

ARTEFACTS

1. RE-CREATING KNOWN EXEMPLARS

Parameter values that reproduce existing artefacts?

This tests the accuracy of the model.

Success is not a sign of “strict” creativity.

Not appropriate for all situations.

Suitable for:

- ➢Weak objectives (formative or summative)
- ➢Strong objectives (formative)

2. EXPLORING THE NEIGHBOURHOOD OF KNOWN EXEMPLARS

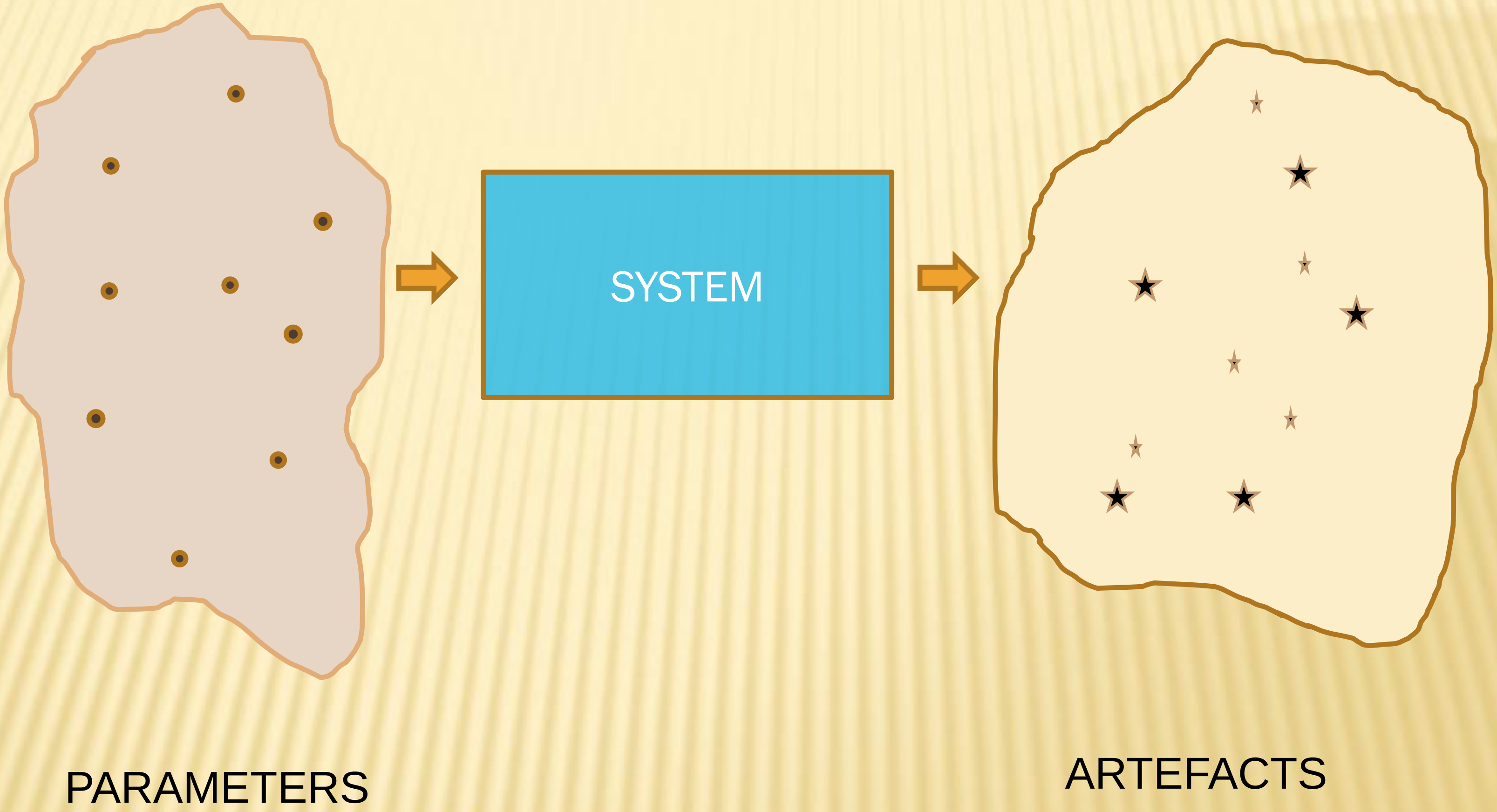
Parameter values that result in output *similar* to existing artefacts?

More realistic than aiming for exact replication of exemplars.

Success here may indicate *extremely low* level of "strict" creativity.

Suitable for:

- Weak objectives (formative or summative)
- Strong objectives (formative)



SYSTEM

PARAMETERS

ARTEFACTS

3. EXPLORING THE PARAMETER SPACE

What happens if you make very small changes to the parameters?

More interesting than aiming for replication of exemplars.

Hard to predict what degree of “strict” creativity might result.

Suitable for:

- Weak objectives (formative or summative)
- Strong objectives (formative)

4. RANDOM SAMPLING OF PARAMETER SPACE

Choose parameter values randomly.

What is meant by “random” in this context?

Need good definition of parameter ranges.

Could show evidence of “strict” creativity.

Suitable for:

- ➢Weak objectives (formative or summative)
- ➢Strong objectives (formative or summative)

5. STRUCTURED SAMPLING OF PARAMETER SPACE

Choose parameter values using a theoretical hypotheses.

You need to have some theoretical ideas.

More oriented towards summative evaluation.

Could be relevant to “strict” creativity.

Suitable for:

- ➢Weak objectives (formative or summative)
- ➢Strong objectives (formative or summative)

SELECTING OUTPUT TO EVALUATE

Possible ways to decide which creations to evaluate:

? Assess all output from all evaluation runs of the system.

? Randomly select output items.

? Select output items in a structured way.

? Select output items in an ad hoc or subjective way.

FACETS OF EVALUATION

- Collecting output items:
  - generating
  - selecting
- What to measure
- How to measure

SUCCESS WITH “WEAK” OBJECTIVES

- Model is about **the domain**.
- Main interest is: how **accurate** is this?
- Hence: is the output (a) well-formed (b) of good quality?
- Two central concepts here: the **“acceptability”/“typicality”**, and the **“value”/“quality”**, of output items.

SUCCESS WITH “STRONG” OBJECTIVES

- Model/hypothesis is about **how to be creative**$_S$.
- Main interest is: are outputs being produced **in a way which is creative**$_S$?
- It is tricky to make this question precise – much debate about this.
- Two central concepts often proposed: the **“value”**/**“quality”** and the **“novelty”**/ **“originality”** of output items.

HOW TO EVALUATE THE QUALITY OF THE OUTPUT

"Quality" is subjective.
Two main approaches:

- Rating of output items by sample audiences.
- Rating of output items by expert annotators.

RATING BY SAMPLE AUDIENCES

- A large number of judges (for statistical reasons).
- Judges **not** knowledgeable about your research.
- No technical/theoretical questions.
- Just a few questions.
- Preparing the data – follow guidelines from experimental psychology (control items, fillers, warm-up/practice, varied sequencing, randomising...)

WHAT TO ASK THE JUDGES?

Various possibilities; e.g.:

“Is this a poem/story/joke/tune?”

“How good is this, on a scale of .....?”

“How original do you think this is, on a scale of .....?”

- Think carefully about the wording.
- Only a small number of questions.
- Keep them simple.

WHAT STATISTICS TO USE?

- Plan this before the testing.
- Have a hypothesis or two.
- What are you interested in:
  - Average ratings?
  - Best rating?
  - Spread of ratings?
  - Difference from control items?
- Lots of established practice in experimental psychology.

HOW TO EVALUATE THE QUALITY OF THE OUTPUT

Since “quality” is subjective, two main approaches:

- Rating of output items by sample audiences.
- Rating of output items by expert annotators.

RATING BY EXPERT ANNOTATATORS

- Just a few judges.
- Judges should be knowledgeable.
- Questions/tasks can be more technical.
- Questions/tasks may need to be structured.

AGREEMENT BETWEEN JUDGES

Borrow methodology from existing areas; e.g.:

- “annotation” - marking up text as a “gold standard”
- “coding” – labelling data in the social sciences

“**reliability**” – to what extent do judges agree?

There are various statistical measures of “agreement” (Cohen’s Kappa, Krippendorf’s Alpha, etc.)

PLAN THE EVALUATION!

*(Very important for **summative** evaluation.)*

- Write up a detailed design first.
- Get comments on the design and revise it.
- Carry out a trial study ("pilot") to check your design.
- Fix any bugs found in the design
- Then do the main evaluation.

THIS TALK

- PART I : Introduction
- PART II : Evaluation
- PART III : Other aspects
  - The “Turing Test”
  - Can evaluation be automated?
  - More “natural” testing?
  - Novelty
  - Effects
  - What about basic software engineering?

PART III : Other aspects

WHAT ABOUT THE “TURING TEST”?

- The original TT is not about creativity
- The TT uses an interactive dialogue
- The TT acts only as a “classifier” (pass/fail)
- The TT encourages trickery that helps to pass the test without addressing the real issues

COMPARING TO HUMAN ARTEFACTS?

Is it helpful to have judges rate both computer-generated artefacts and human-created artefacts?

YES!

✓ **Not** a Turing Test.
✓ Establishes a standard of the “normal” range of ratings for this kind of artefact.
✓ Allows statistical comparison of ratings (of computer items) with this standard.
✓ These comparisons can take account of variations between judges (variable tastes).

AUTOMATED TESTING?

In other fields, there are test suites, benchmark data, etc.

Similarity to known exemplars (see earlier) can be tested automatically (for “weak” objectives, or formative evaluation).

But “quality” and “creativity” are open-ended.

Similarity to benchmarks might show **lack** of “strict” creativity.

If there were mechanisable criteria – why not build them into the creative model?

COULD THE SETTING BE MORE NATURAL?

E.g.

visual art in a gallery
music at a concert

How would we measure success?

How could we distinguish creativity from ‘mere’ quality?

Novelty

Distinguish (Boden 1990):

- **H-creativity**: novel within history
- **P-creativity**: novel for the creator

Correspondingly, two notions of “novelty”:

- H-novelty
- P-novelty

We are interested in P-creativity....
and hence **P-novelty.**

WHAT ABOUT MEASURING NOVELTY?

Lack of similarity to known artefacts? (H-novelty).

Variety amongst the system's output? (P-novelty)?

Asking judges? (What questions should be asked?)

How can judges detect P-novelty as opposed to H-novelty?

CAN WE MEASURE EFFECTS?

Outside CC, **task-based** evaluation is sometimes used.

For creative systems, what is the “task”?

Perhaps “artistic” artefacts should evoke an **emotional response**.

For example – jokes: amusement.

What about forms that can evoke a wide range of emotions?
E.g. music, poetry, visual art, stories.

Could we measure this?

WHICH METHODOLOGIES?

Is software engineering methodology irrelevant ?

We still neec
- design proc
- implement

No!

Still need abstraction, modularity, clear coding, etc.

The differences arise regarding *requirement specification* and *testing*.

SUMMARY

- Evaluating a creative system is not like testing conventional software.
- Usually “creativity” is measured indirectly, as “quality”, “novelty”, and other properties.
- Much of the methodology can be borrowed from the social sciences.
- There are still open questions about measurement of “creativity”.

RECOMMENDED READING

- Two chapters in “Computational Creativity”, eds. A. Cardoso & T. Veale [Springer, 2017]:
    - *The Evaluation of Creative Systems*. Graeme Ritchie.
    - *Evaluating Evaluation: Assessing Progress and Practices in Computational Creativity Research*. Anna Jordanous.
- *Empirical Methods for Artificial Intelligence*. Paul R. Cohen. MIT Press, Cambridge, Mass. 1995.
- Any good psychology textbooks on experimental design & statistics.