Evaluation of Interactive Machine Learning Systems

Nadia Boukhelifa
IAL, September 2019

visual

Evaluation of Interactive Machine Learning Systems (IVMLs)

Nadia Boukhelifa
IAL, September 2019

Who am I ? Nadia BOUKHELIFA

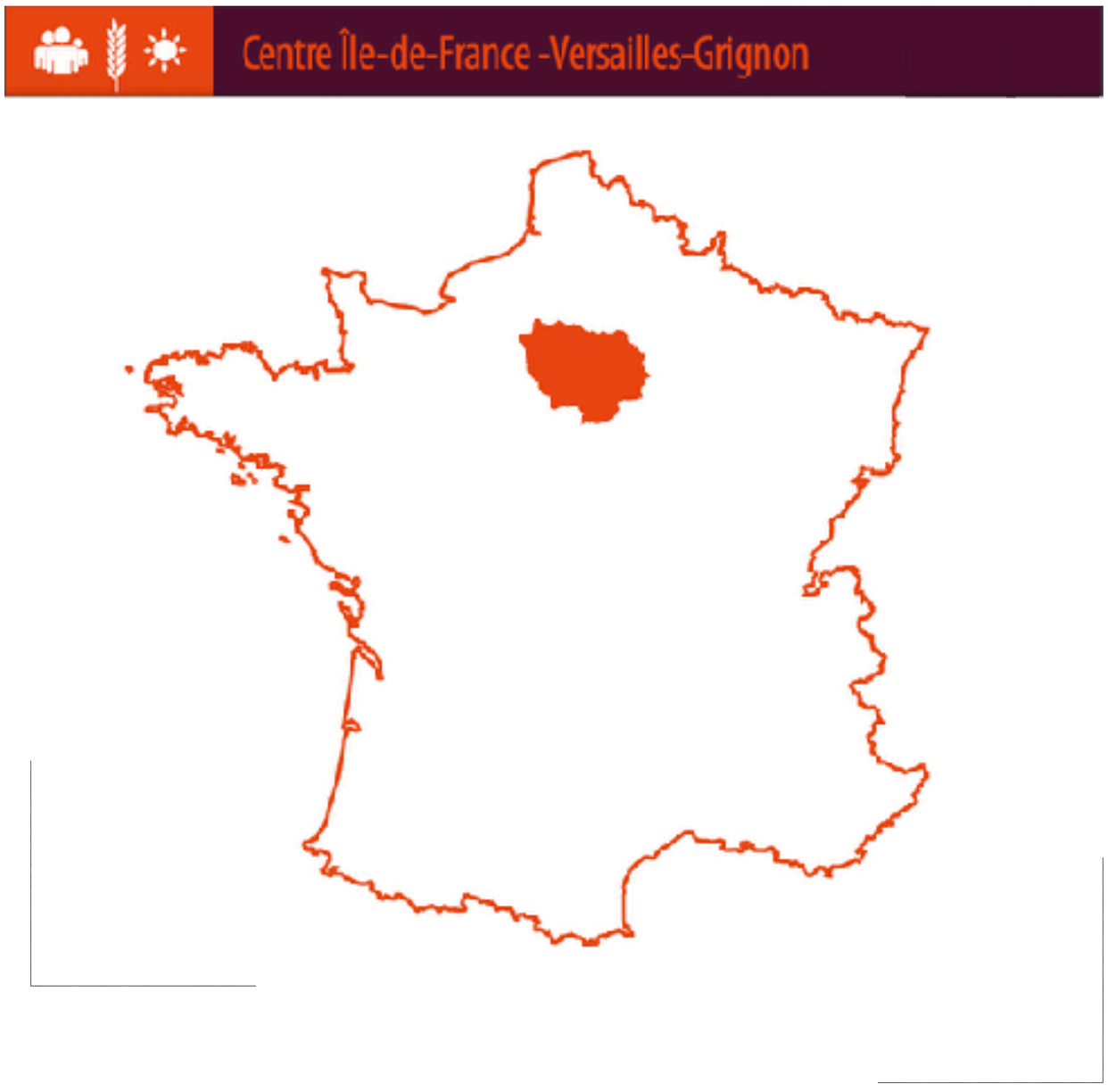
Ph.D. 2007 — University of Kent, UK
Computer Science, Information Visualization

Post-doc — INRIA, Télécom ParisTech, France
Visualization, Human-Computer Interaction

Researcher, Tenured, 2016 — INRA, France
Multi-dimensional Data Visualization, Interactive Modelling

INRA — National Institute of Agricultural Research

Nadia BOUKHELIFA



INRA — National Institute of Agricultural Research

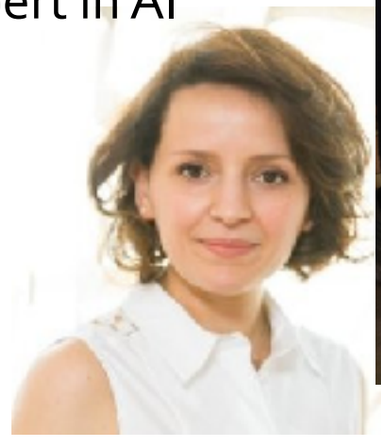
Nadia BOUKHELIFA

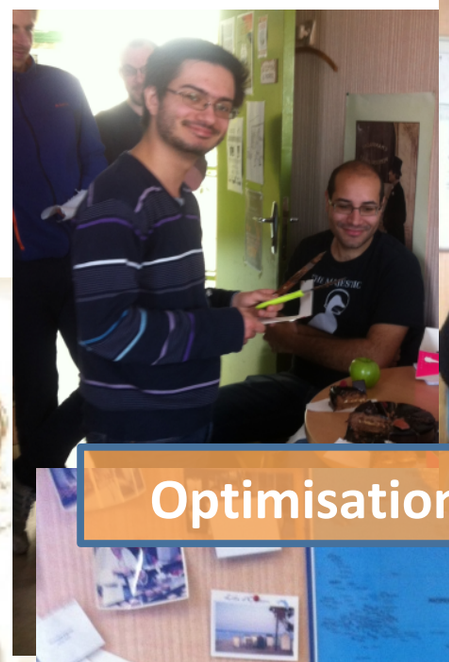
http://www.cfsg.fr/site-de-grignon

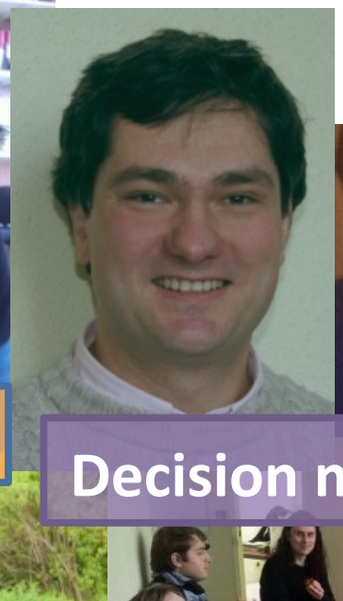
MALICES Team: Modelling and Knowledge Integration

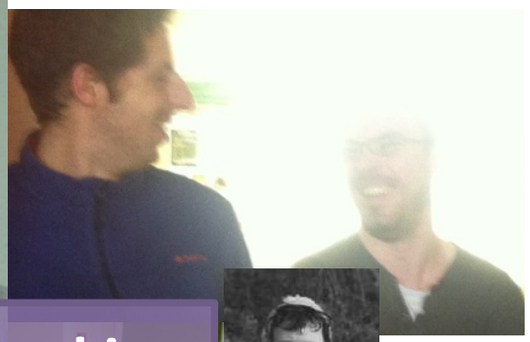
MALICES Team: Modelling and Knowledge Integration

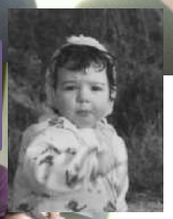
Not expert in AI









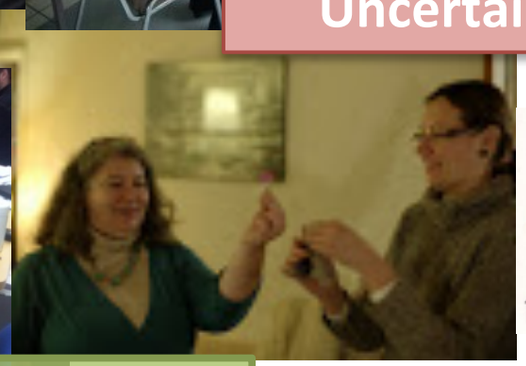


Optimisation, visualisation

Decision making

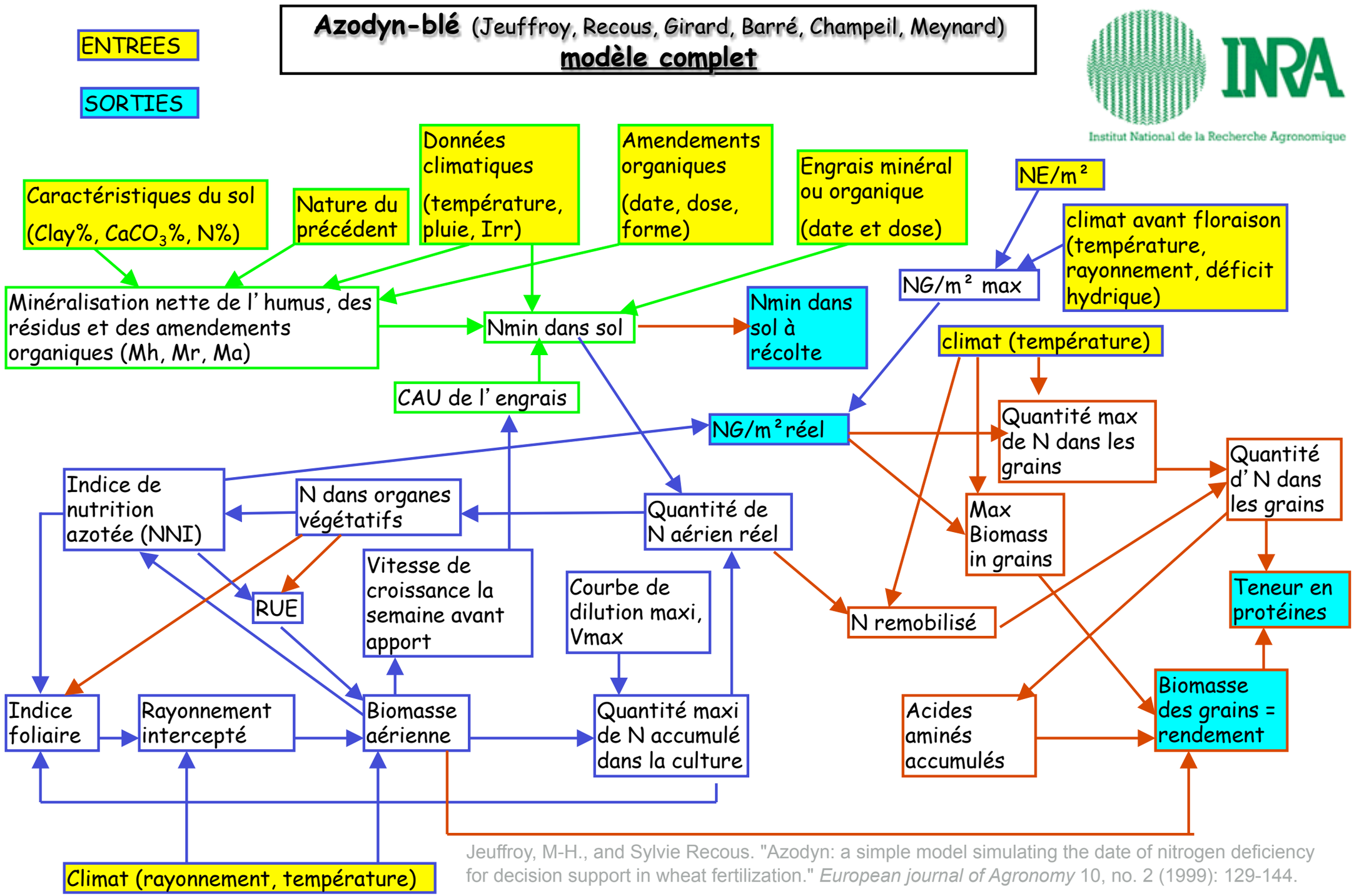
Uncertainty

Expertise formalisation



Deterministic and stochastic modelling

Context: complex systems, complex datasets, complex models …



Jeuffroy, M-H., and Sylvie Recous. "Azodyn: a simple model simulating the date of nitrogen deficiency for decision support in wheat fertilization." *European journal of Agronomy* 10, no. 2 (1999): 129-144.

Context : Interactive Model Exploration

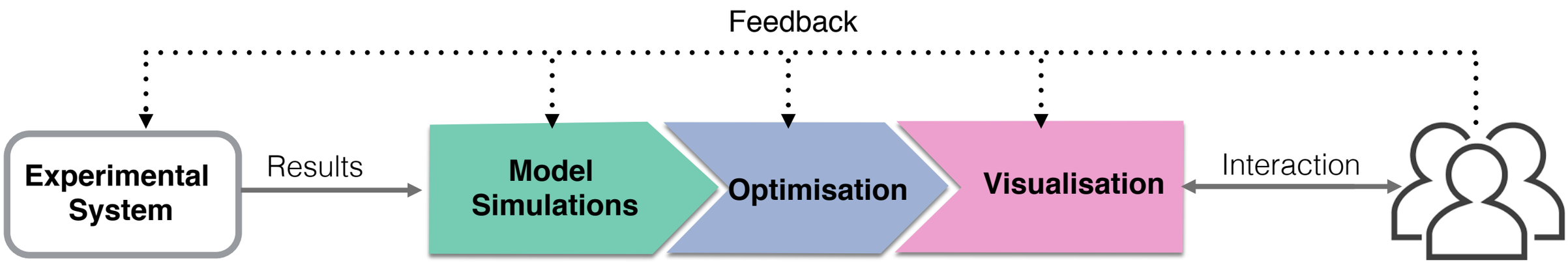
Why human in the loop in modelling ?

- to integrate valuable experts knowledge that may be hard to encode directly into mathematical or computational models.

- to help resolve existing uncertainties as a result of, for example, bias and error that may arise from automatic machine learning.

- to build trust by making humans involved in the modelling or learning processes.

Our approach to Interactive Model Exploration



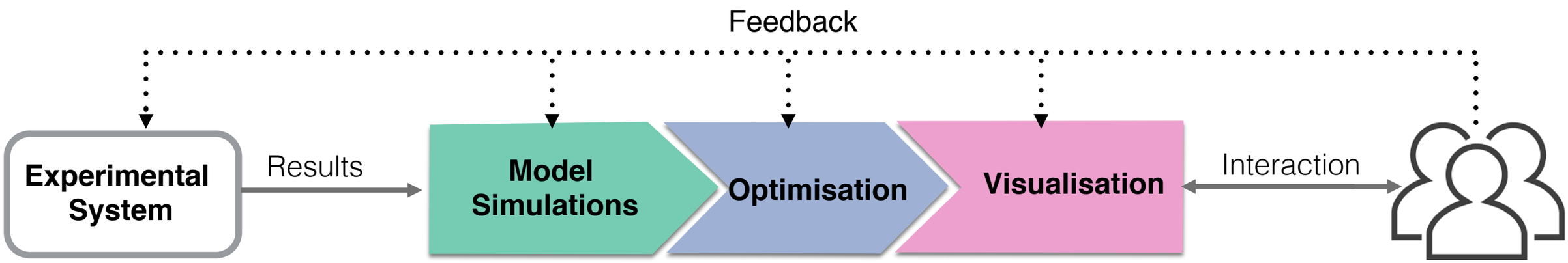
Application

Multiple Computational Stages

Multiple Actors

Boukhelifa, Nadia, et al. "An Exploratory Study on Visual Exploration of Model Simulations by Multiple Types of Experts." Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 2019.

Our approach to Interactive Model Exploration



Application

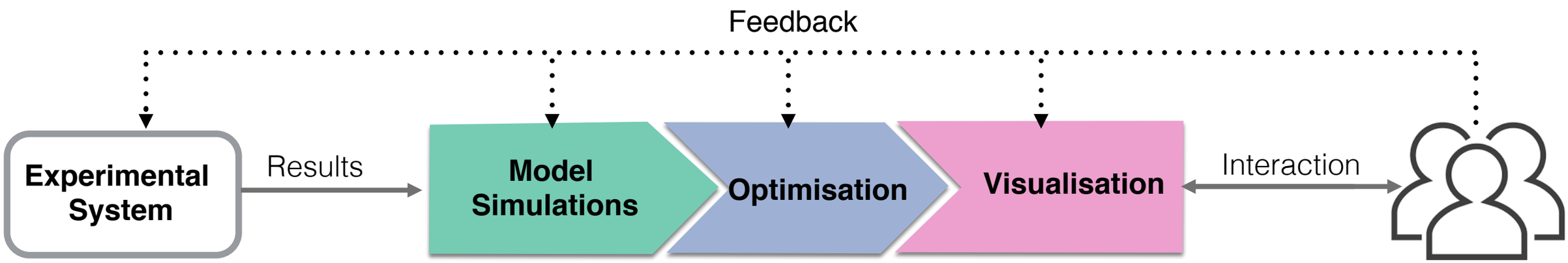
Multiple Computational Stages

Multiple Actors

Collaborative setup

Boukhelifa et al., CHI2019

Our approach to Interactive Model Exploration



Application

Multiple Computational Stages

Multiple Actors

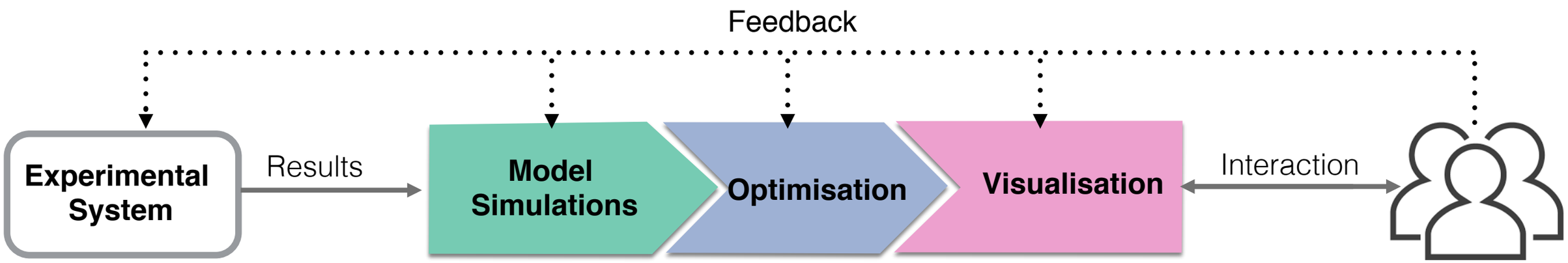
Collaborative setup

- Models written by third parties
- Experts specialists in parts of the modelled process

=> Co-located multiple expertise : domain, model, optimisation, visualization.

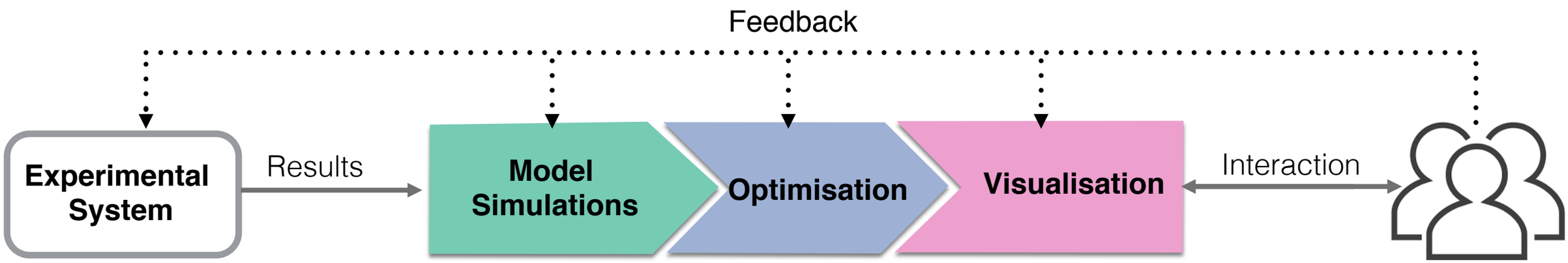
Boukhelifa et al., CHI2019

Our approach to Interactive Model Exploration



Data & Models

Our approach to Interactive Model Exploration



Application

Multiple Computational Stages

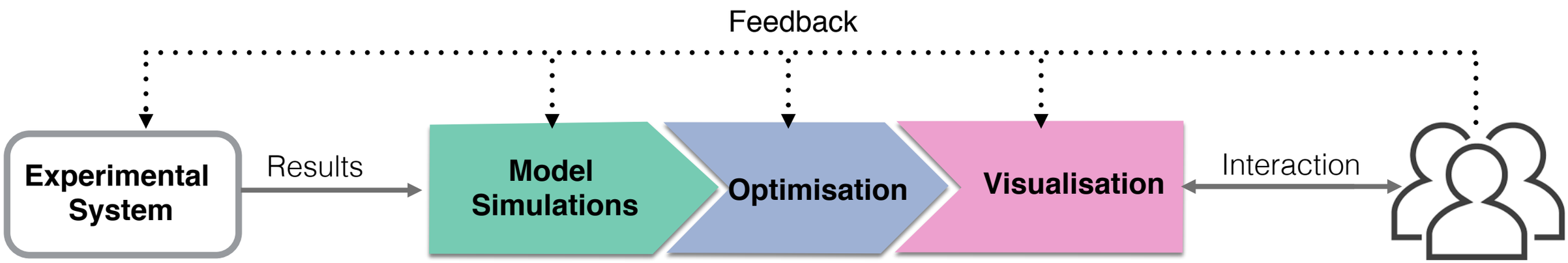
Multiple Actors

Trade-offs

Data & Models

Boukhelifa et al., CHI2019

Our approach to Interactive Model Exploration



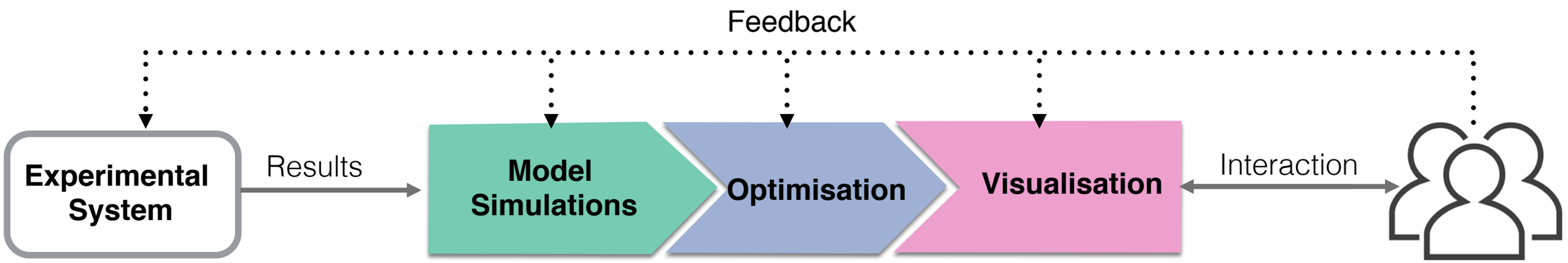
Data & Models

Pareto Fronts

set of **non-dominated compromise points**, where no objective can be improved without sacrificing at least one other objective.

Boukhelifa et al., CHI2019

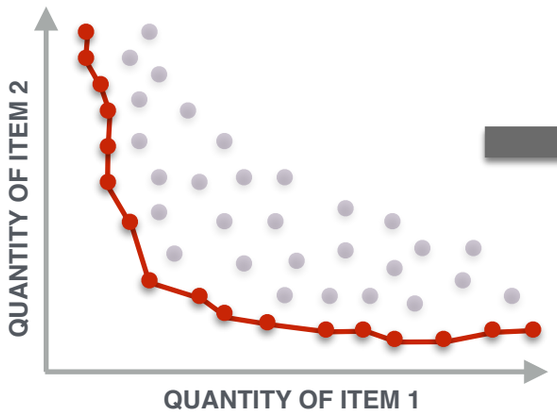
Our approach to Interactive Model Exploration

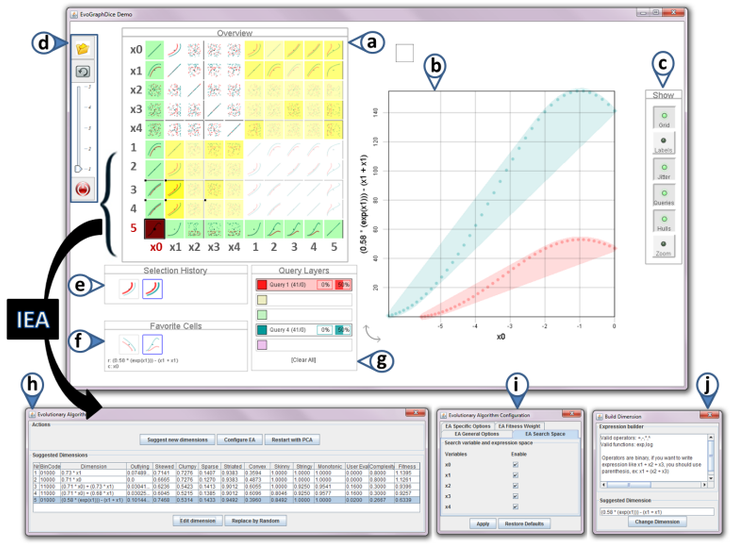


Application

Multiple Computational Stages

Multiple Actors





Data & Models

Pareto Fronts

Visualization System

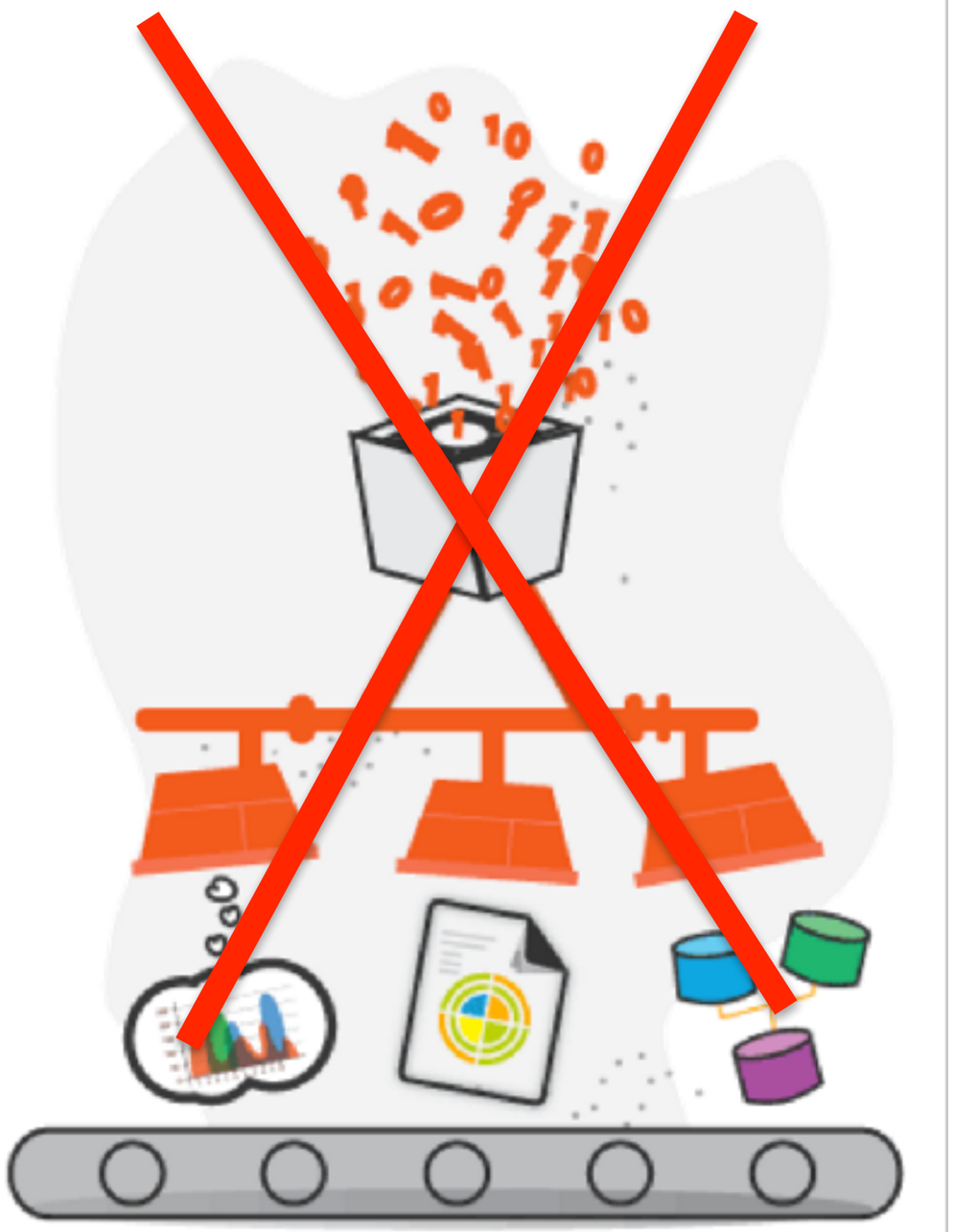
Definition #1: Visualization

“a way to generate pretty images from data”

http://www.seguetech.com/

Definition #1: Visualization

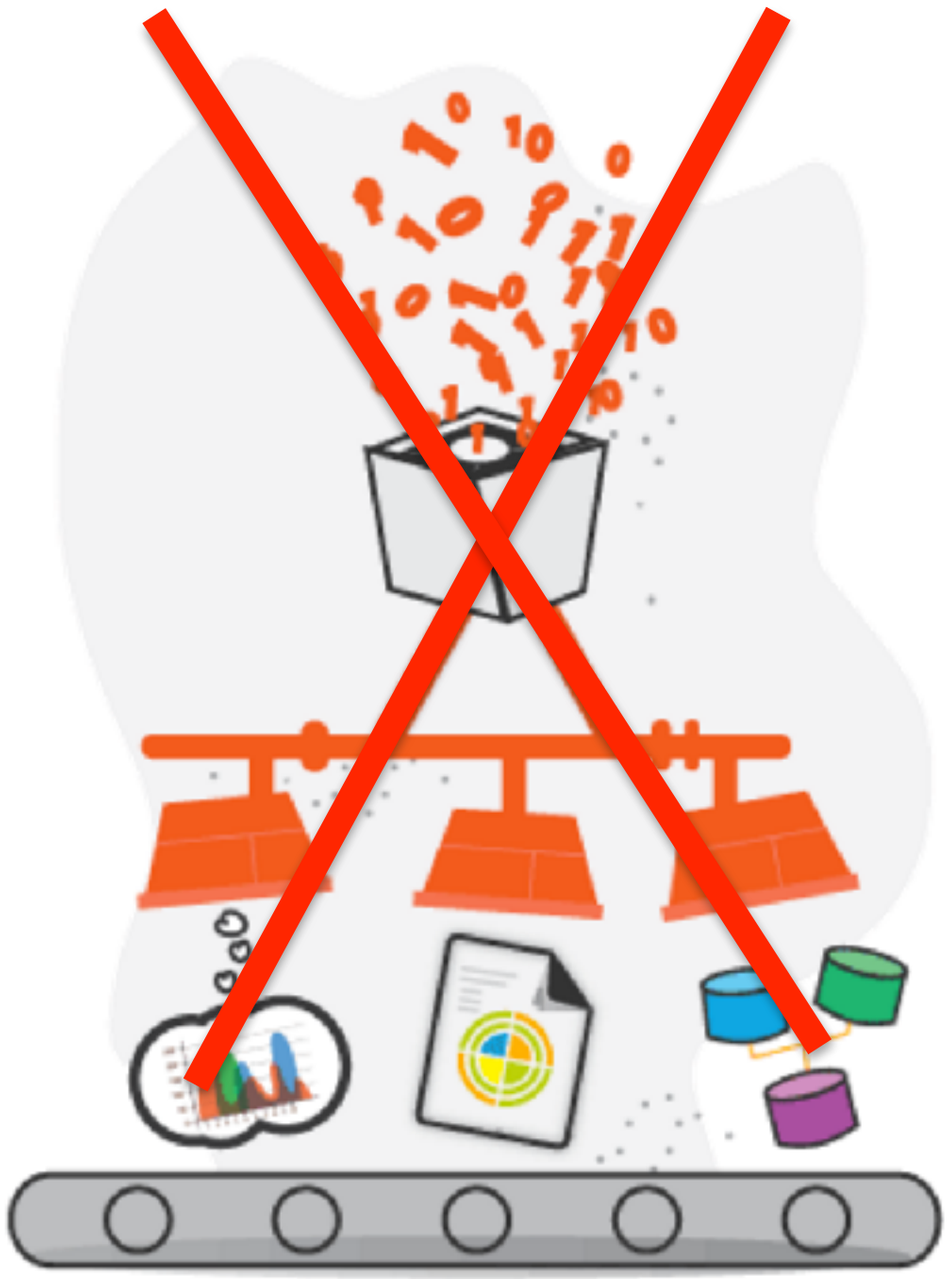
The purpose of visualization is **insight**, not pictures !



http://www.seguetech.com/

Definition #1: Visualization

The purpose of visualization is **insight**, not pictures !



http://www.seguetech.com/

Information Visualisation

> ***"The use of computer-supported, interactive, visual representations of abstract data to amplify cognition."***
> [Card et al., 1999]

Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman. "Using vision to think." Readings in information visualization. Morgan Kaufmann Publishers Inc., 1999.
Card, S.

Why visualise your data

Raw Data from Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Frank Anscombe

Statistical analysis

Raw Data from Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Statistical Properties

Mean of *x*	9.0
Variance of *x*	11.0
Mean of *y*	7.5
Variance of *y*	4.12
Correlation between *x* and *y*	0.816
Linear regression line	$y = 3 + 0.5x$

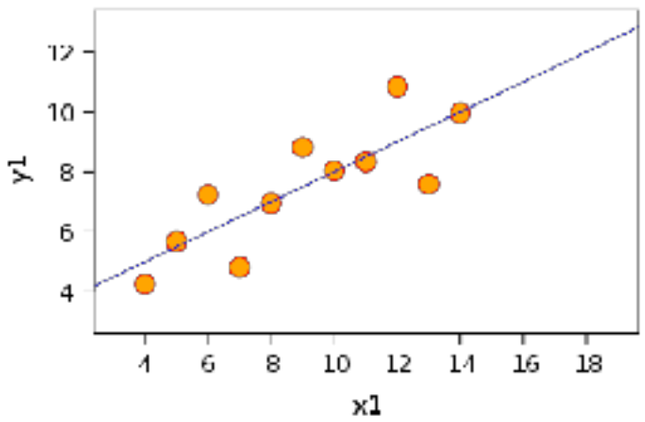
Visual representation of the data

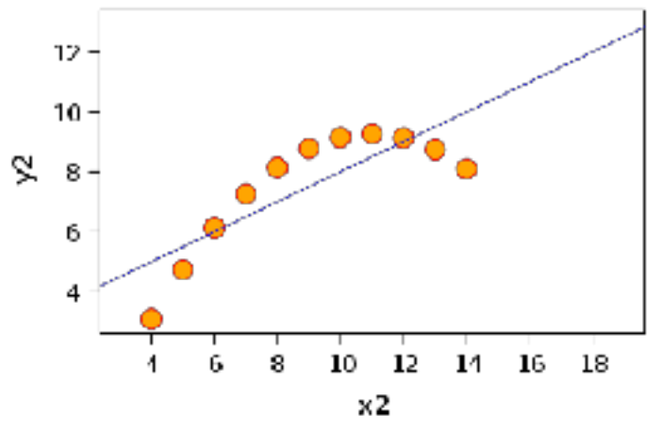
Raw Data from Anscombe's Quartet

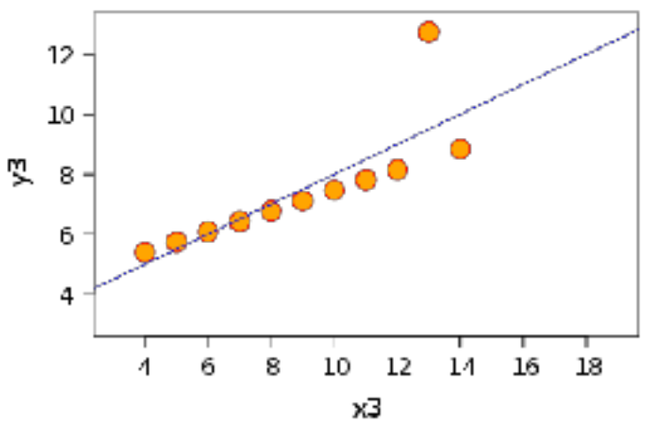
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

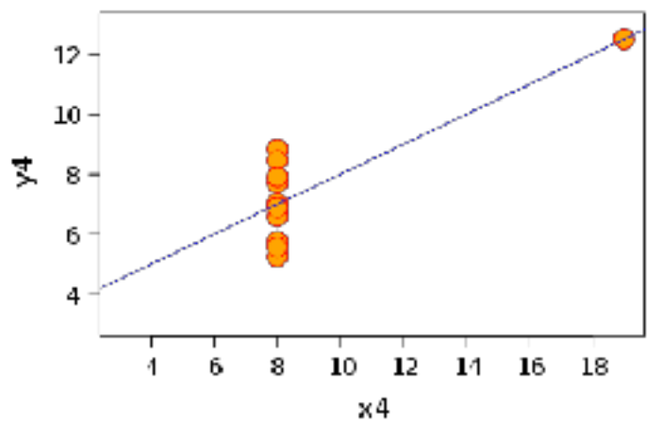
Visual Properties

Same stats, different graphs!



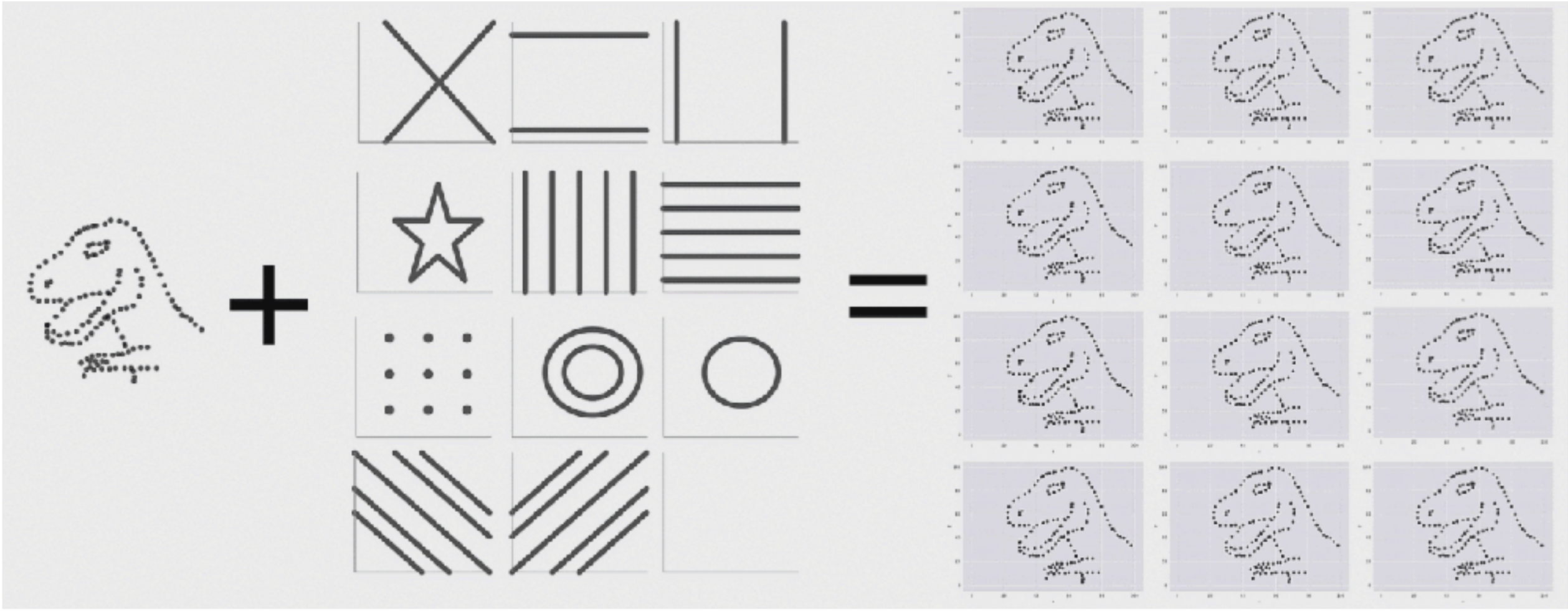






Visual representation of the data

Never trust summary statistics alone; visualise your data !



Datasaurus Dozen Dataset, Autodesk Research

https://www.autodeskresearch.com/publications/samestat

Information Visualization

Communicate visually

Explore interactively

Evaluate

Definition #2: Human-Computer Interaction

> “Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them.”

Hewett, T. et al. ACM Curricula for Human-Computer Interaction. 1992;

Definition #3: iML

"Users Are People, Not Oracles"

Amershi et al., Power to the People: The Role of Humans in Interactive Machine Learning , 2014

Taking into account human factors

"algorithms that can **interact** with both computational agents and **human agents** (in active learning: oracles) and can optimize their learning behavior through these interactions."

Andreas Holzinger. Interactive machine learning (iml). Informatik Spektrum, 39(1), 2016.
Daniel Kottke, Interactive Adaptive Learning, Intelligent Embedded Systems, University of Kassel, 2018

Definition #4: IVML

Our definition:

> "In interactive visual machine learning (IVML), a human operator and a machine collaborate to achieve a task mediated by an interactive visual interface."

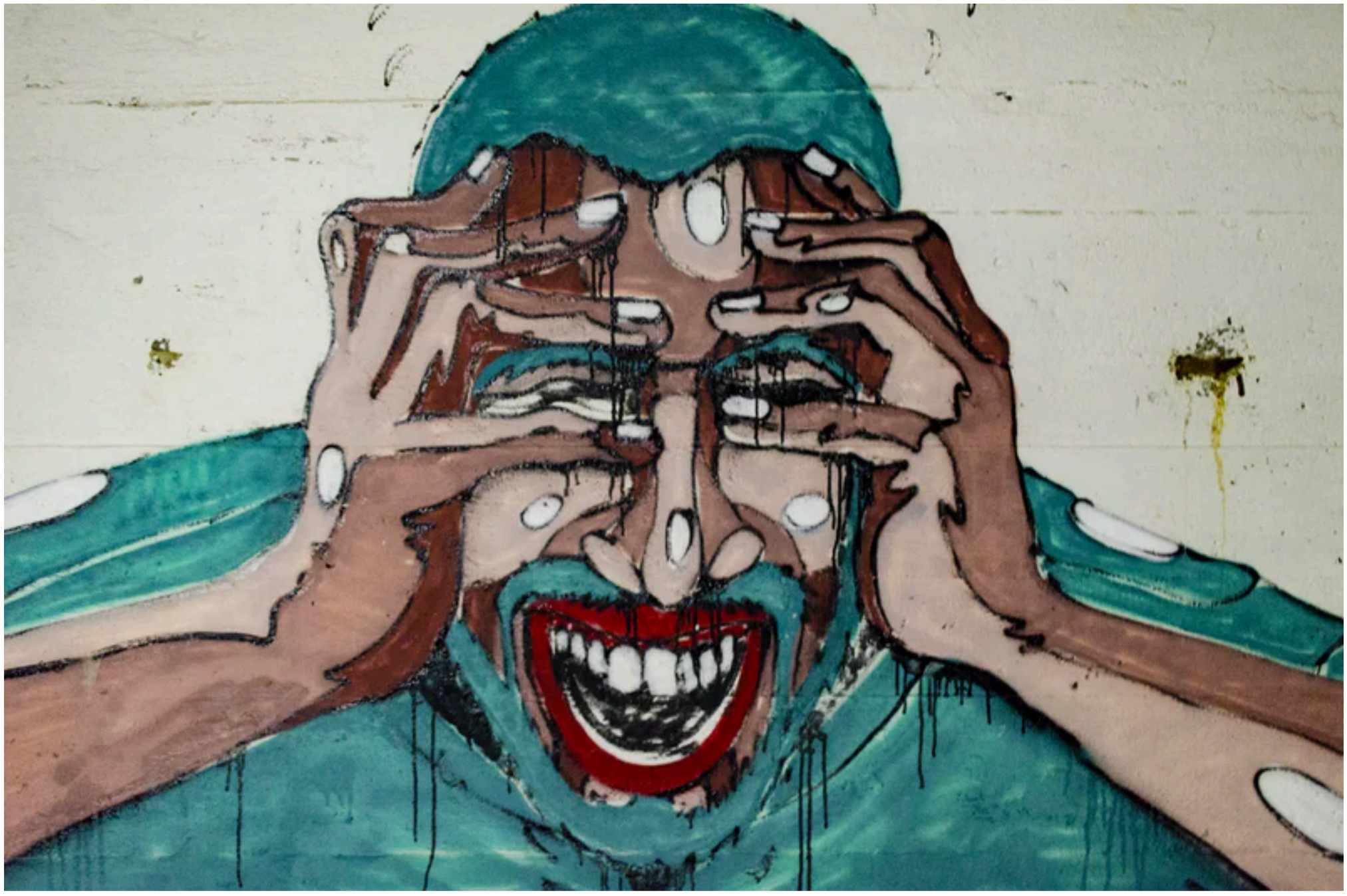
N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, 2018.

Evaluation of Interactive Systems

We know how to assess:

- ✔ User performance
- ✔ User experience
- ✔ Algorithmic performance

Evaluation of IVML Systems



https://unsplash.com/photos/VBegzj-JHBs

Evaluation of IVML Systems

Our experience with EVE - Evolutionary Visual Exploration

N. Boukhelifa, A. Bezerianos, I-C Trelea, N. M Perrot, and E Lutton. An Exploratory Study on Visual Exploration of Model Simulations by Multiple Types of Experts." ACM CHI Conference on Human Factors in Computing Systems, **2019**

N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, **2018**.

N. Boukhelifa, A. Bezerianos, W. Cancino and E. Lutton. Evolutionary Visual Exploration: Evaluation of an IEC Framework for Guided Visual Search. Evolutionary Computation Journal, MIT press, **2017**.

N. Boukhelifa, A. Bezerianos and E. Lutton. A Mixed Approach for the Evaluation of a Guided Exploratory Visualization System. EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3) **2015**.

W. Cancino, N. Boukhelifa, A. Bezerianos and E. Lutton. Evolutionary Visual Exploration: Experimental Analysis of Algorithm Behaviour. GECCO workshop on Genetic and Evolutionary Computation (VizGEC **2013**).

N. Boukhelifa, W. Cancino, A. Bezerianos and E. Lutton. Evolutionary Visual Exploration: Evaluation With Expert Users. Computer Graphics Forum (EuroVis 2013), Eurographics Association, **2013**, 32 (3).

Evaluation of IVML Systems

Our experience with EVE - Evolutionary Visual Exploration

w. Cancino A. Bezerinaos E. Lutton

N. Boukhelifa, A. Bezerianos, I-C Trelea, N. M Perrot, and E Lutton. An Exploratory Study on Visual Exploration of Model Simulations by Multiple Types of Experts." ACM CHI Conference on Human Factors in Computing Systems, **2019**

N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, **2018**.

N. Boukhelifa, A. Bezerianos, W. Cancino and E. Lutton. Evolutionary Visual Exploration: Evaluation of an IEC Framework for Guided Visual Search. Evolutionary Computation Journal, MIT press, **2017**.

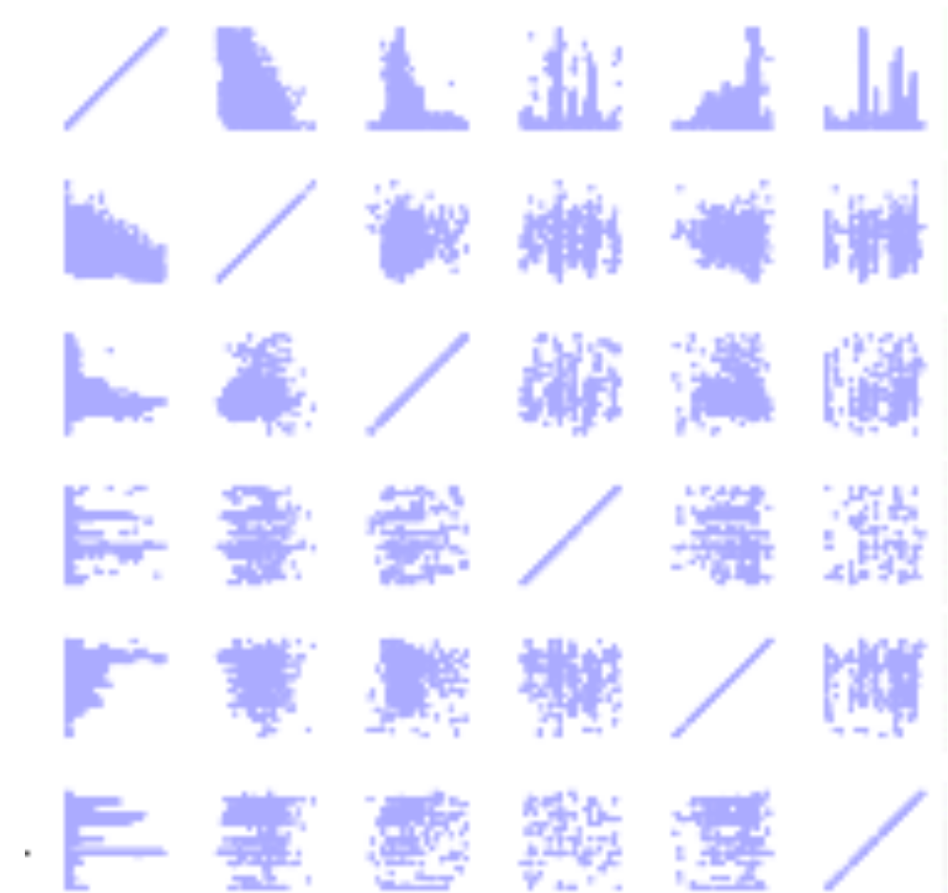
N. Boukhelifa, A. Bezerianos and E. Lutton. A Mixed Approach for the Evaluation of a Guided Exploratory Visualization System. EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3) **2015**.

W. Cancino, N. Boukhelifa, A. Bezerianos and E. Lutton. Evolutionary Visual Exploration: Experimental Analysis of Algorithm Behaviour. GECCO workshop on Genetic and Evolutionary Computation (VizGEC **2013**).

Evolutionary Visual Exploration

https://unsplash.com/photos/YTbFHT9_IhY

Need to explore combined dimensions



Top_Heat	Bottom_Heat	Humidity	Weight	Height	Colour
0	0	43	1.9	4.369	42.5
416.67	315.11	22	-39.22	4.3788	45.08
1148.6	831.88	87	-45.516	4.4064	49.292
1563.7	1173.1	30	-20.043	4.4271	51.297
1791.7	1365.5	18	-8.2444	4.4511	52.636
2098.5	1602.1	59	-8.5867	4.4852	54.369
2348.4	1810.2	10	-8.4489	4.5279	56.253
2560.8	1952.8	20	3.2	4.5768	57.316
2583.8	2010.9	99	21.787	4.6411	57.871
2692.3	2151.6	81	35.856	4.721	58.48
2750.1	2311	90	33.311	4.819	59.133
2891.2	2459.7	27	36.122	4.9469	59.853
2918.8	2604.4	60	33.102	5.1004	60.642
3067.1	2769.1	13	25.54	5.2774	61.382
3110.3	2865.2	44	18.553	5.4751	62.187

How to explore a huge search space of possible combined dimensions ?

How to explore a huge search space of possible combined dimensions ?

Evolutionary Visual Exploration

Why IEAs ?

Suitable for exploratory visualization :

- ✔ can combine objective & subjective measures
- ✔ support exploration & exploitation
- ✔ adapt to user change of interest

Boukhelifa et al., EuroVis 2013

IEA

EvoGraphDice Demo

Overview

Selection History

Query Layers

Favorite Cells

Show

Grid

Labels

Jitter

Queries

Hulls

Zoom

Query 1 (41/0) 0% 50%

Query 4 (41/0) 0% 50%

[Clear All]

(0.58 * (exp(x1))) - (x1 + x1)

x0

Evolutionary Algorithm

Actions

Suggest new dimensions | Configure EA | Restart with PCA

Suggested Dimensions

N°	BinCode	Dimension	Cullying	Skewed	Clumpy	Sparse	Striated	Convex	Skinny	Stringy	Monotonic	User Eval	Complexity	Fitness
1	01000	0.73 * x1	0.07489...	0.7111	0.7276	0.1407	0.9383	0.3584	1.0000	1.0000	1.0000	0.0000	0.8000	1.1395
2	10000	0.71 * x0	0.0	0.6655	0.7276	0.1270	0.9383	0.4873	1.0000	1.0000	1.0000	0.0000	0.8000	1.1261
3	11000	(0.71 * x0) + (0.73 * x1)	0.03041...	0.6236	0.5423	0.1413	0.9012	0.6055	1.0000	0.9250	0.9541	0.1600	0.3000	0.9396
4	11000	(0.71 * x0) + (0.68 * x1)	0.03025...	0.6045	0.5215	0.1385	0.9012	0.6006	0.8048	0.9250	0.9577	0.1600	0.3000	0.9257
5	01000	(0.58 * (exp(x1))) - (x1 + x1)	0.10144...	0.7458	0.5314	0.1433	0.9492	0.3960	0.8492	1.0000	1.0000	0.0200	0.2667	0.6339

Edit dimension | Replace by Random

Evolutionary Algorithm Configuration

EA Specific Options | EA Fitness Weight | EA General Options | EA Search Space

Search variable and expression space

Variables	Enable
x0	☑
x1	☑
x2	☑
x3	☑
x4	☑

Apply | Restore Defaults

Build Dimension

Expression builder

Valid operators: +,-,*,/

Valid functions: exp,log

Operators are binary, if you want to write expression like x1 + x2 + x3, you should use parenthesis, ex: x1 + (x2 + x3)

Suggested Dimension

(0.58 * (exp(x1))) - (x1 + x1)

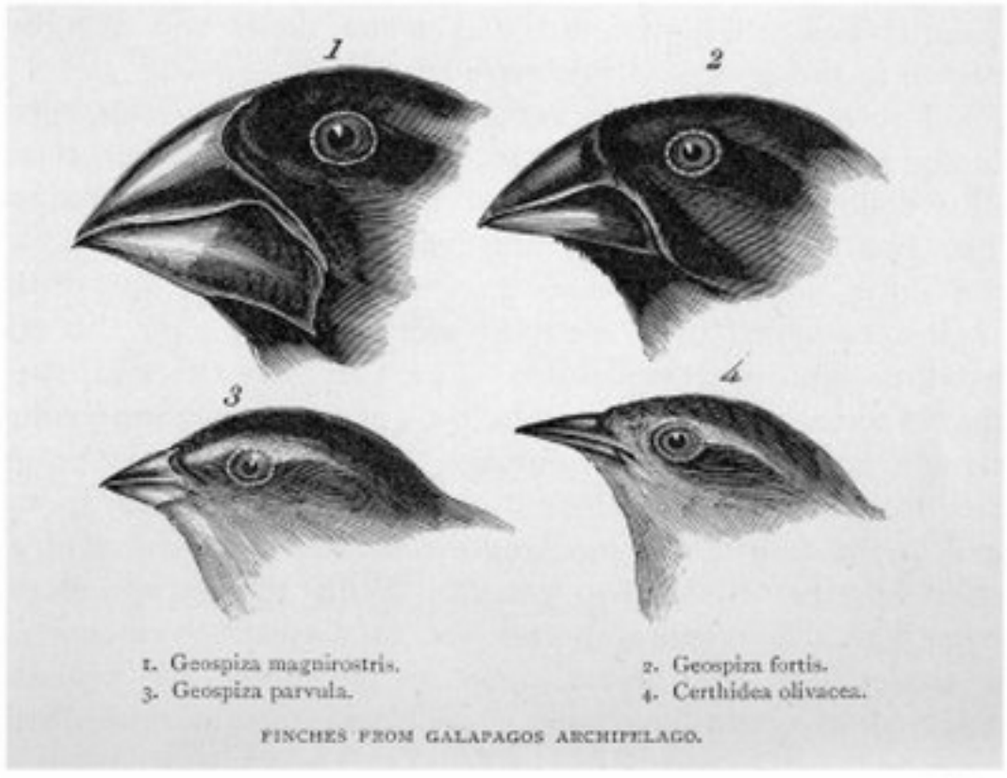
Change Dimension

Evolutionary Visual Exploration with EvoGraphDice

N. Boukhelifa, W. Cancino, A. Bezerianos, E. Lutton
INRIA, Université Paris Sud 11, INRA

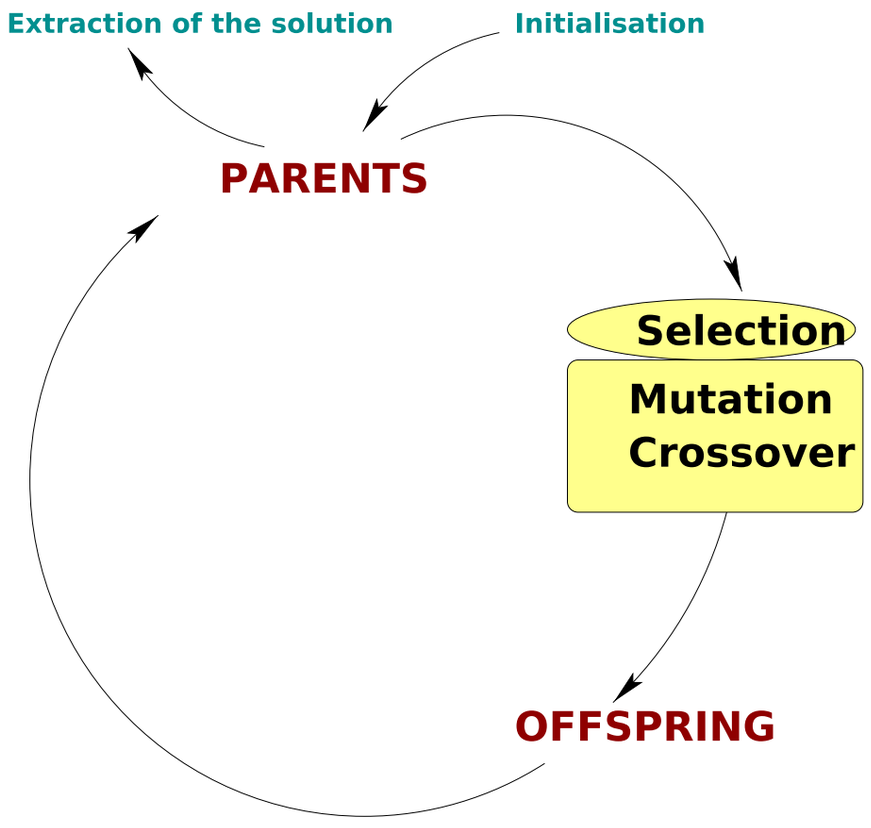
Artificial Evolution

Natural Evolution

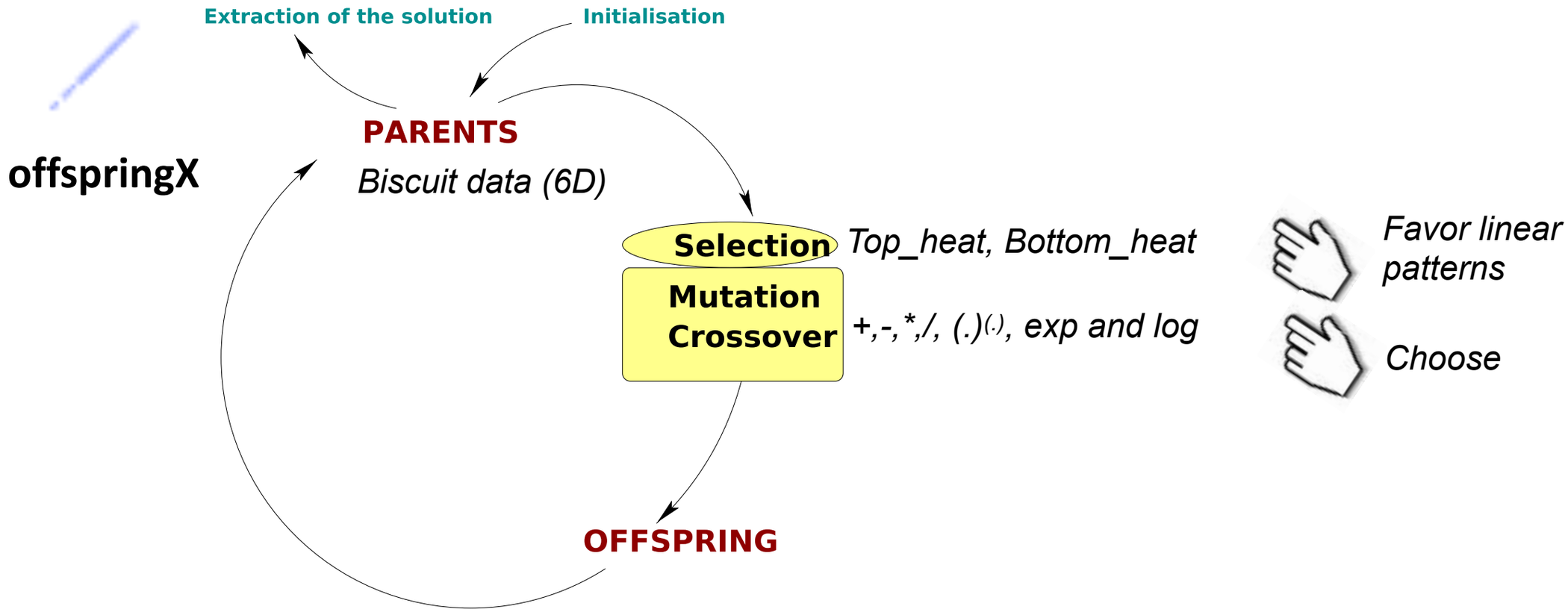


Wikipedia

Evolutionary Algorithms (EAs)



EVE: Creating New Projections



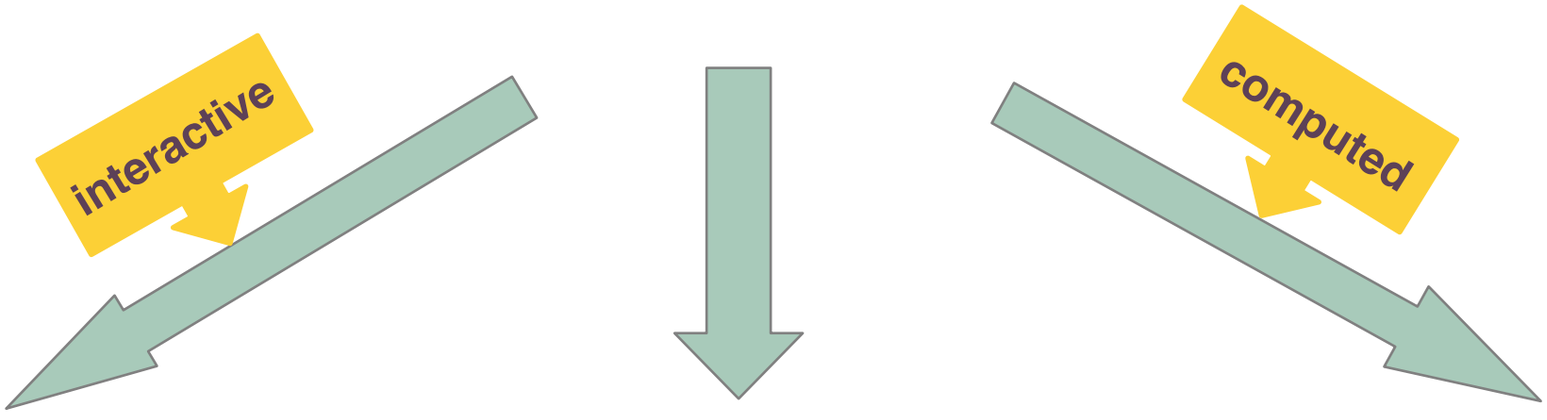
offspring1: 0.2 * Top_heat + 0.7 * Bottom_heat

offspring2: Exp(Top_heat)

offspring3: 0.99 - (0.69 * log(Top_heat))

…

Evaluation of Projections



User assessment

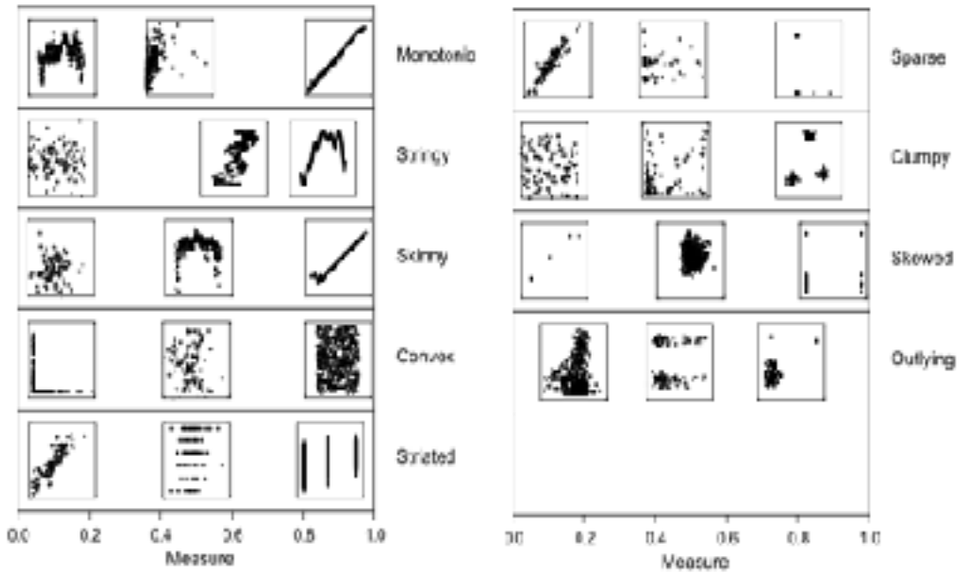
Surrogate function

Complexity

$$f_{sc}(y_i) = \sum_{k=1..9} w_k \left(\max_{j-1..n} SC_k(y_i, x_j) \right)$$

learned

$$f_c(y_i) = \left(1 - \frac{nvars(y_i)}{n}\right) \times \frac{1}{depth(y_i)}$$



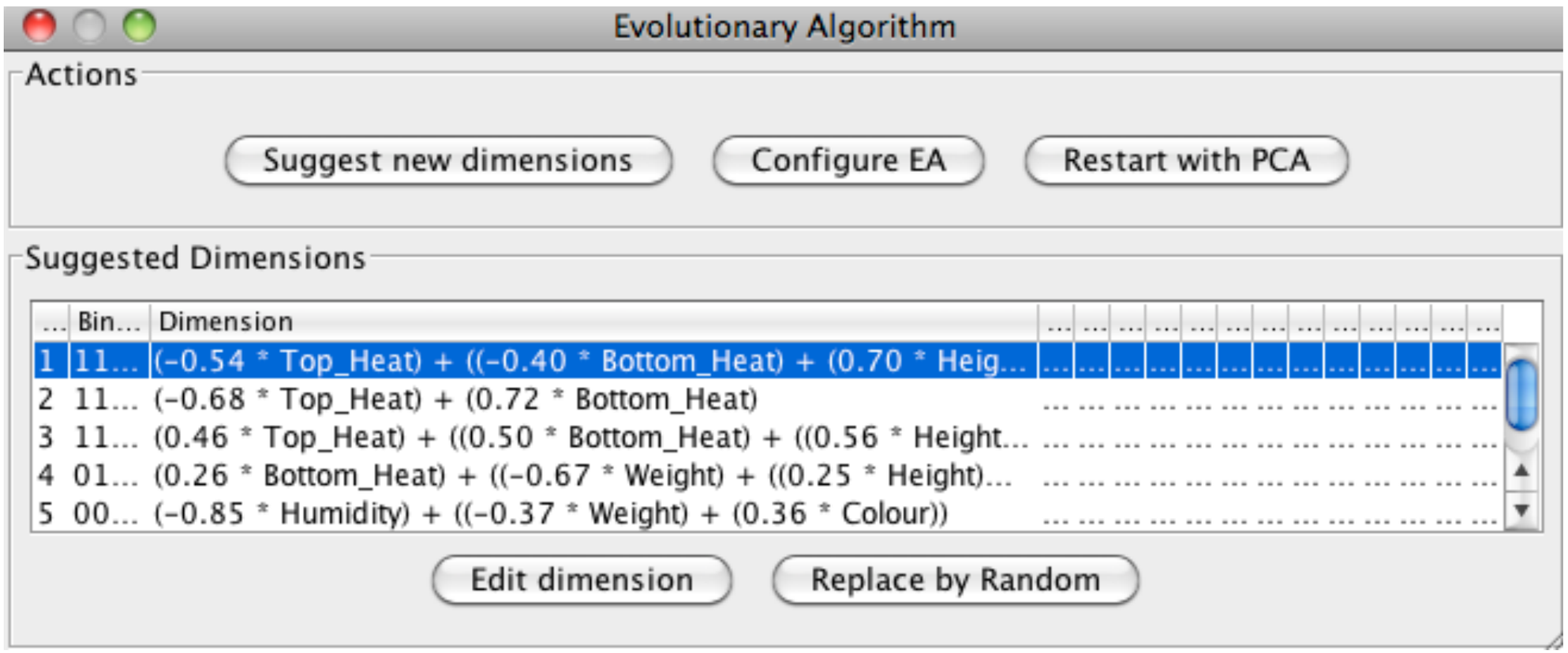
Scagnostics, Wilkinson and Wills, 2008

The Search Space

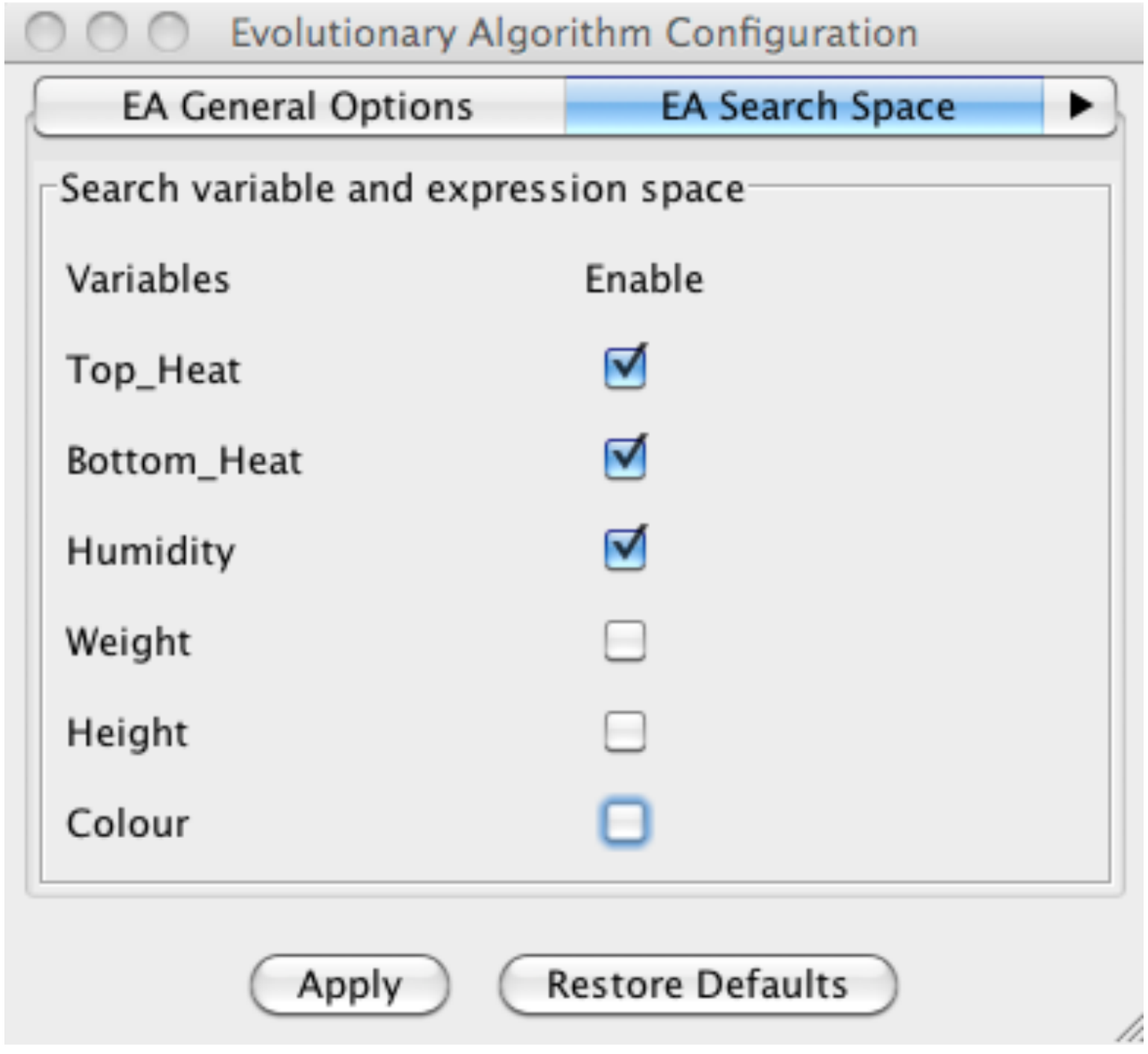
The Set of all dimensions encoded as trees

(GP framework, Koza92)

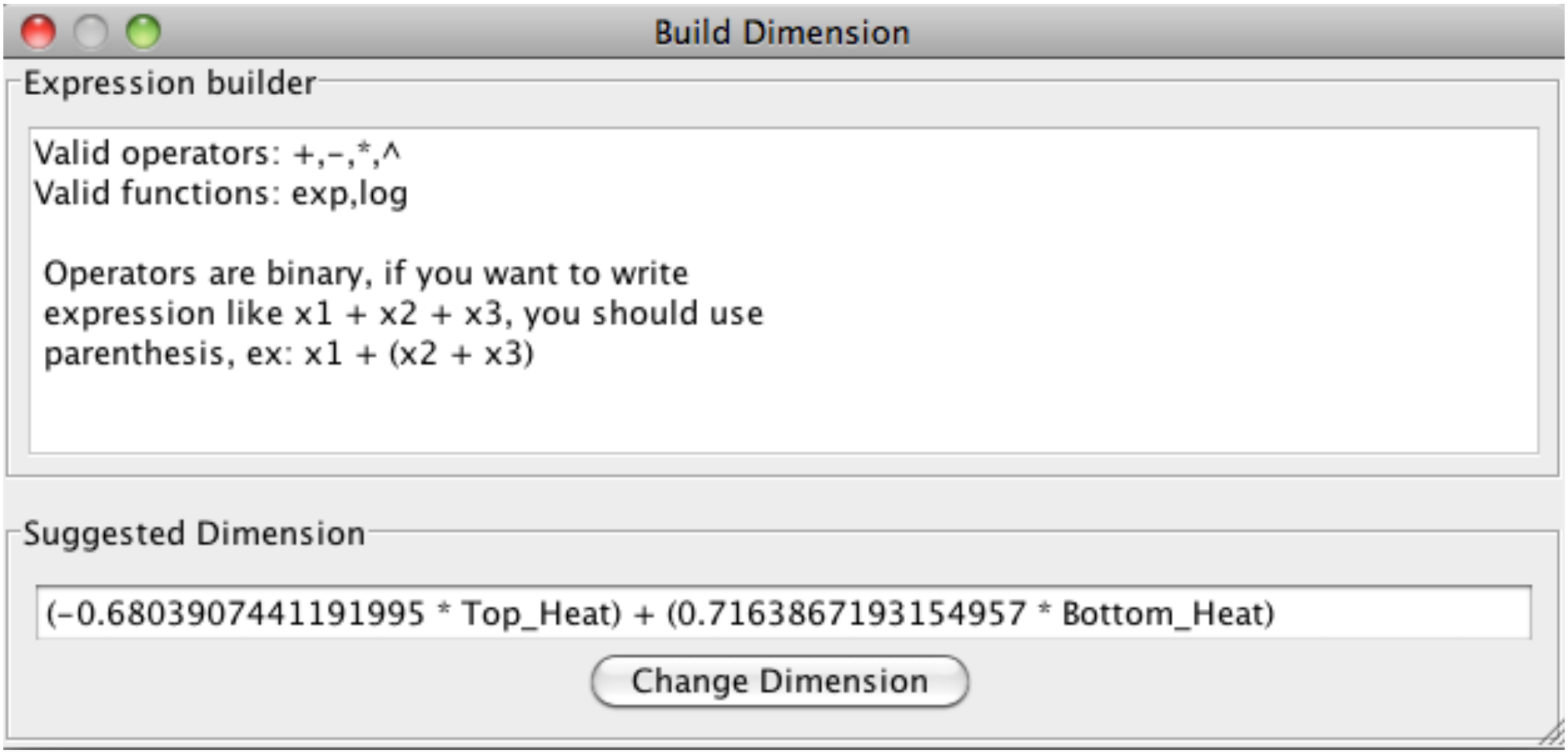
initial dimensions, operators, constants



Subspace Exploration



Subspace Exploration



Key features of EVE

Intuitive

Interactive

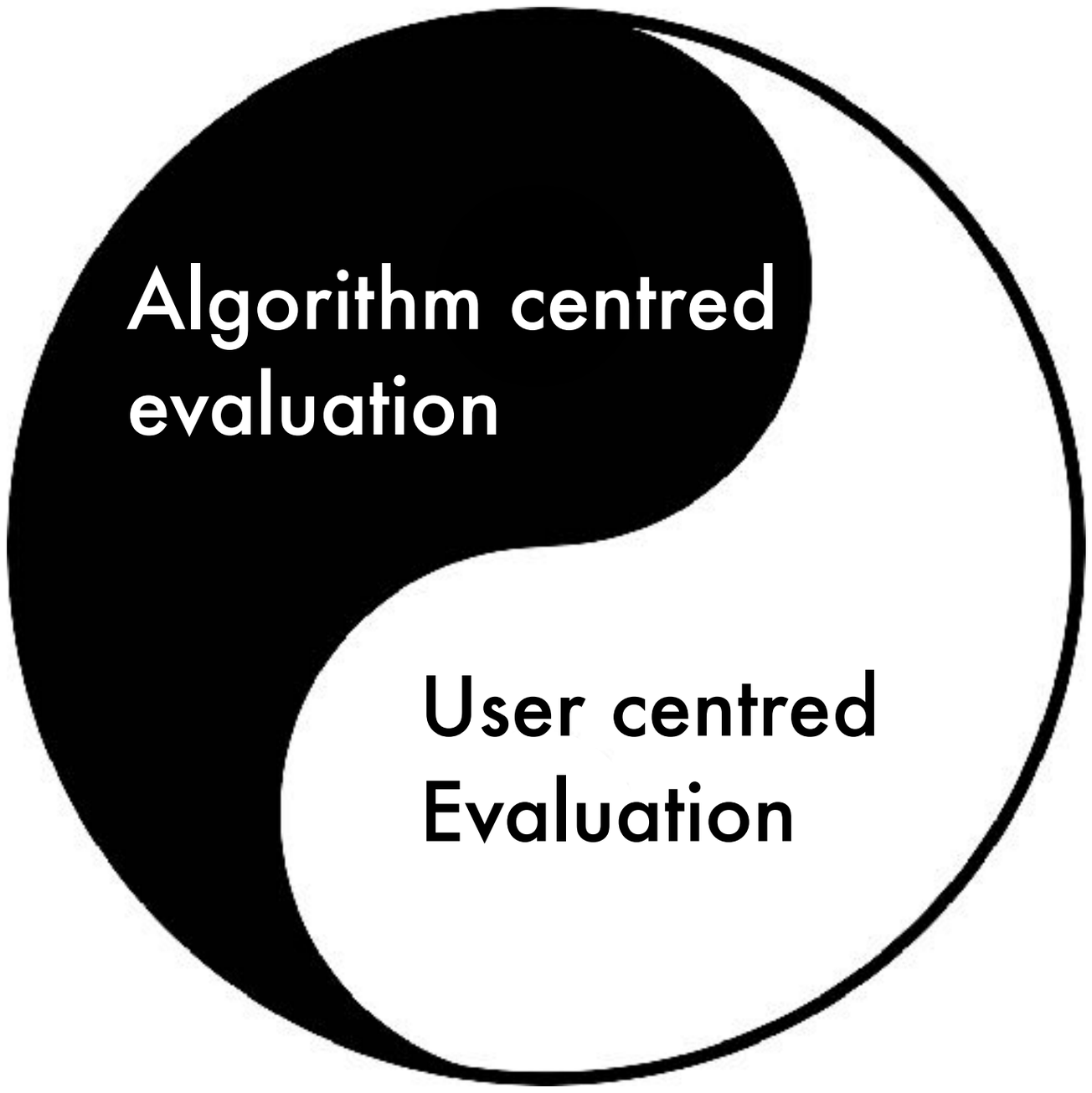
Is it ?

Adaptable

Flexible

Evaluating EVE

A mixed approach



Is the human-machine cooperation producing the right results?

When we have ground-truth

Evaluating human-AI collaboration

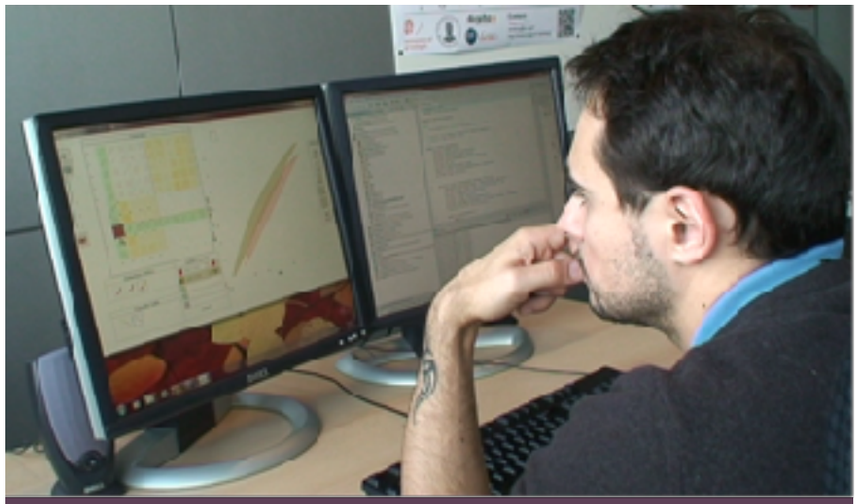
Aim

- is the exploration actually steered toward an interesting area of the search space?
- are the proposed solutions varied?

Quantitative study methodology

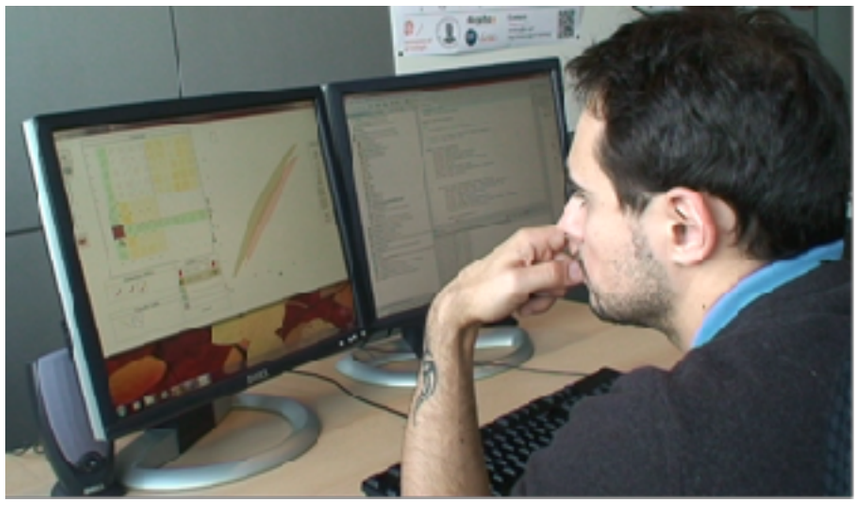
- synthetic dataset
- pre-specified task

Participants



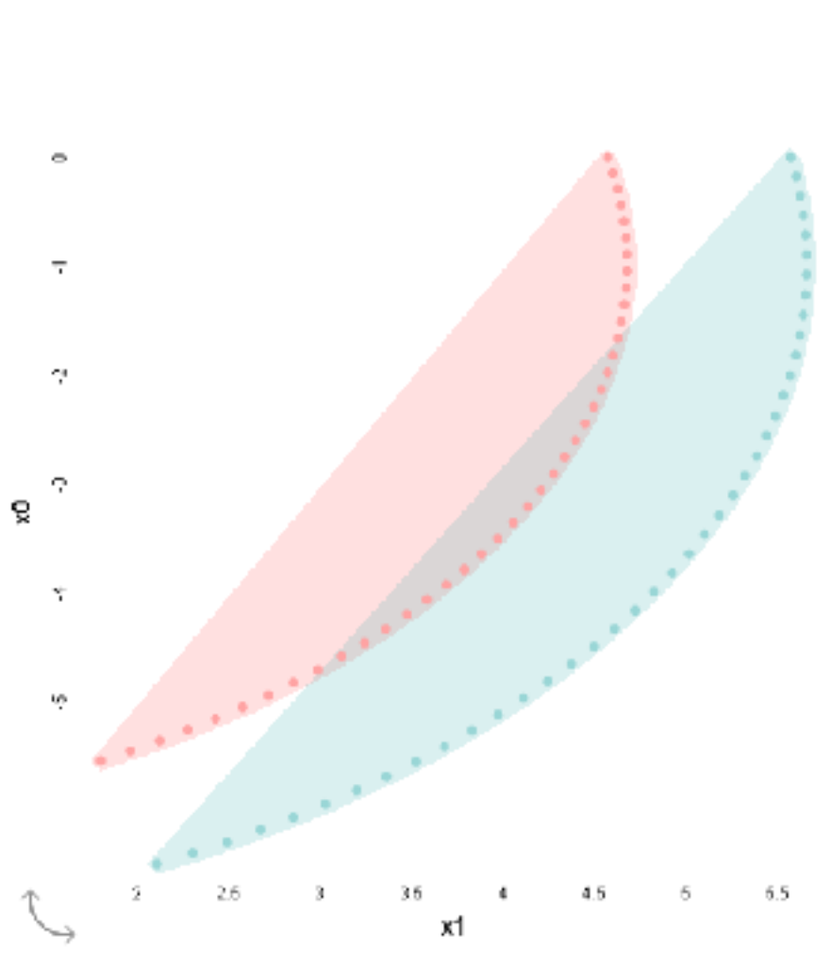
- 12 participants
- mean 28.5 years
- no experience required
- Synthetic dataset 5D
- 20 minutes

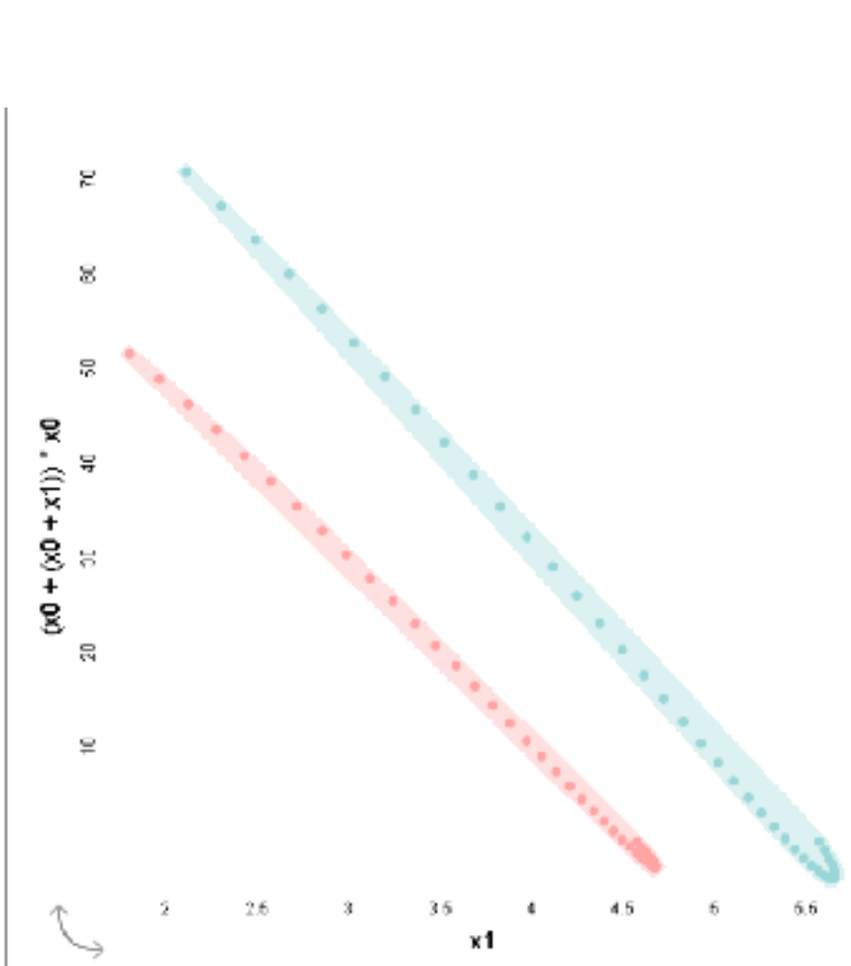
Task

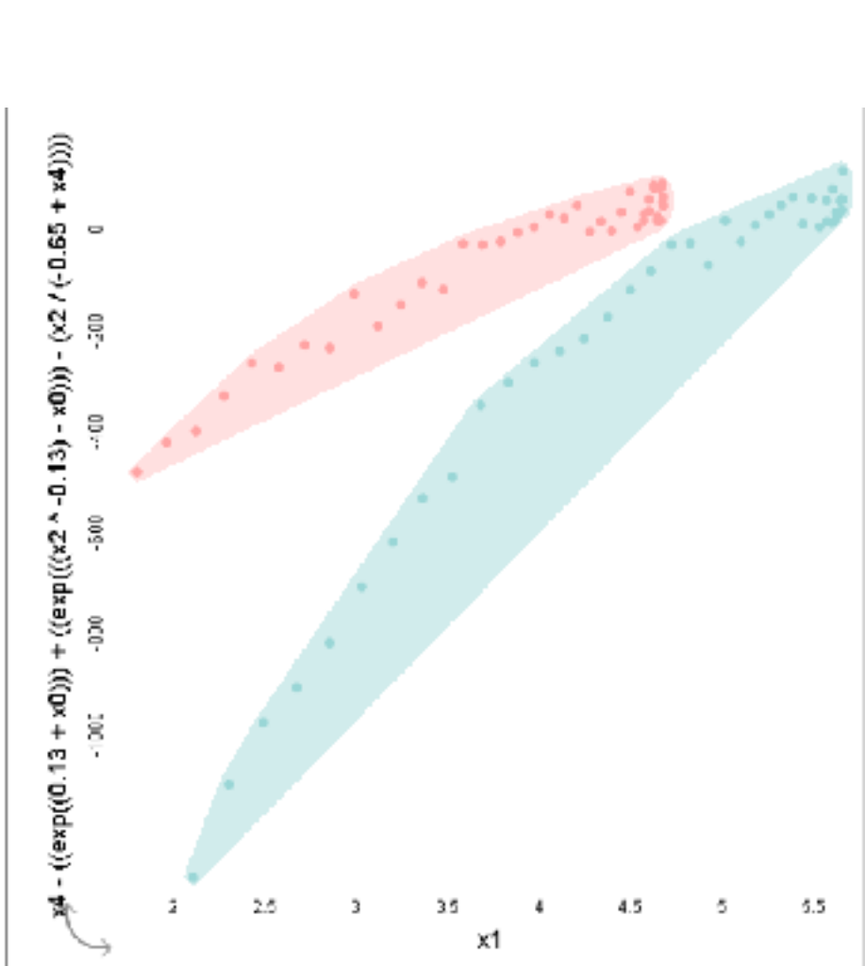


Game

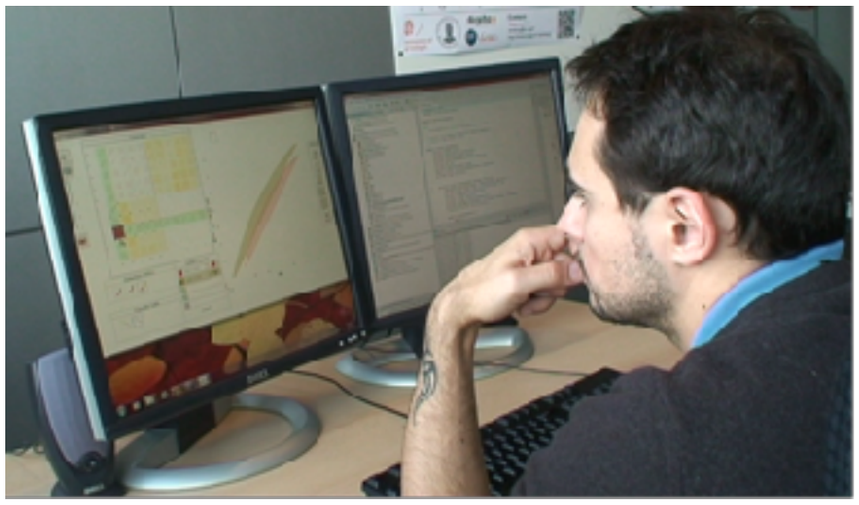
separate two point clusters



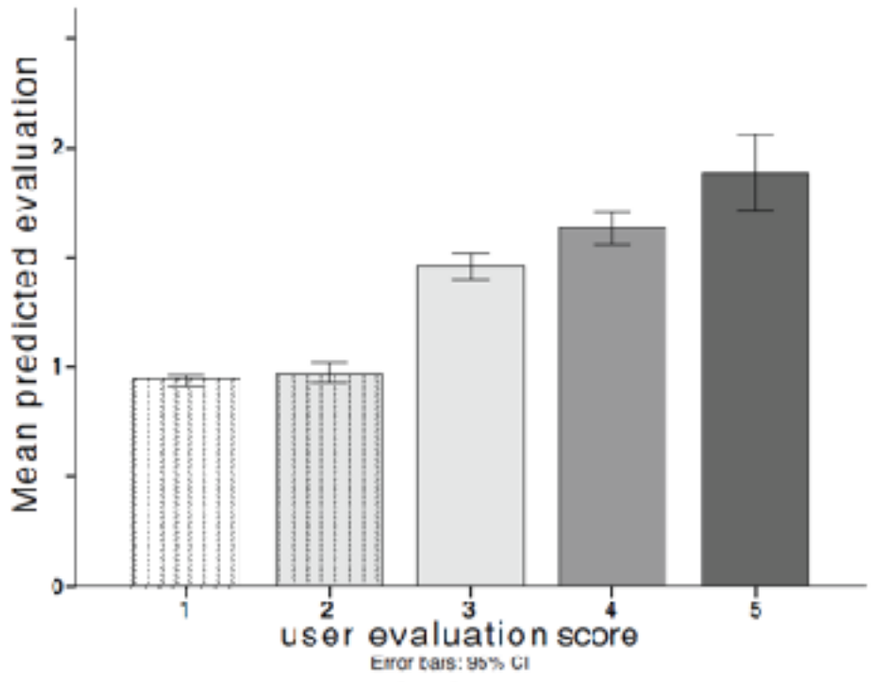


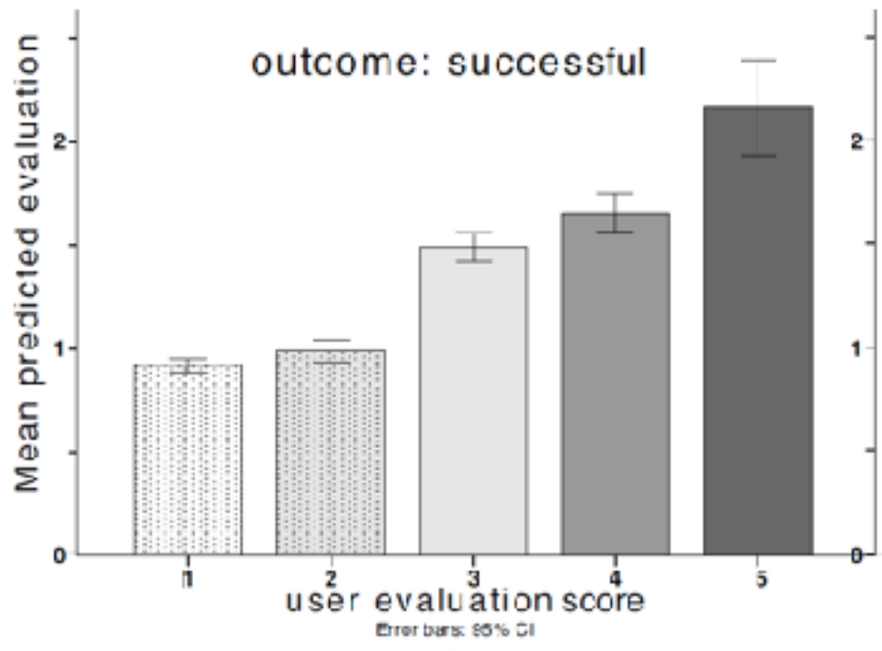


Convergence analysis

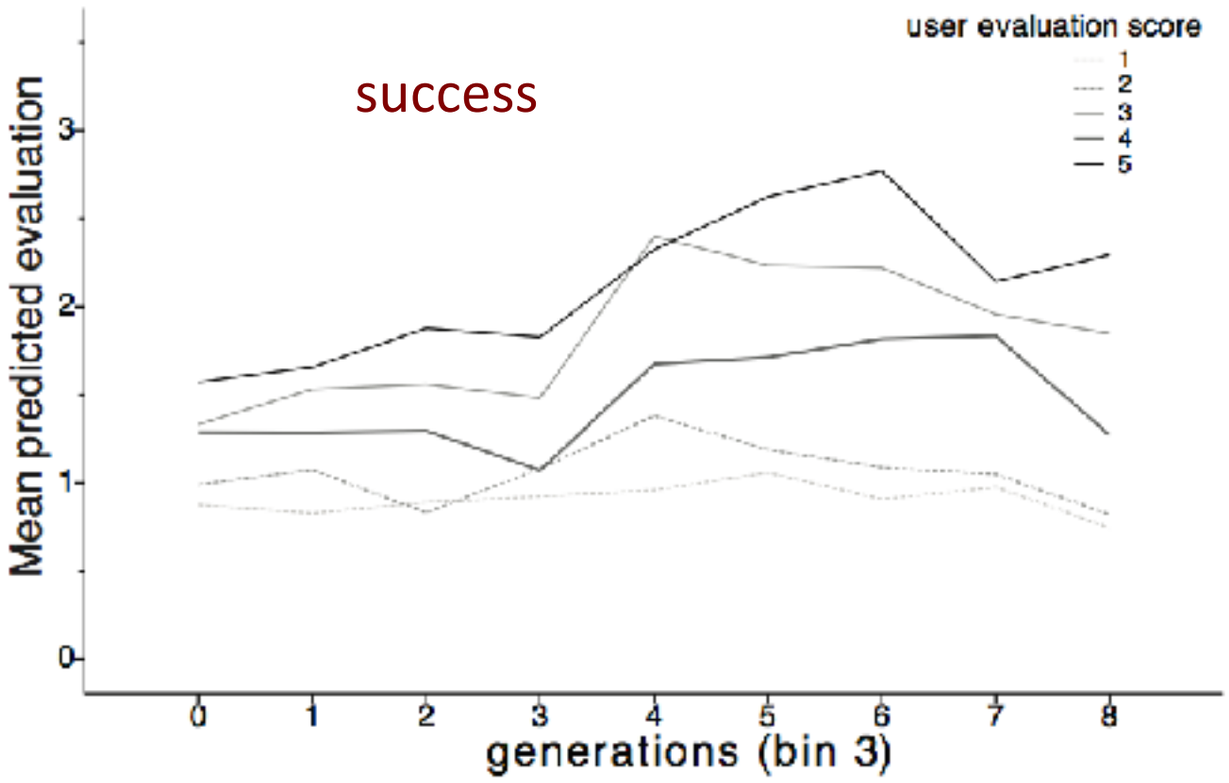


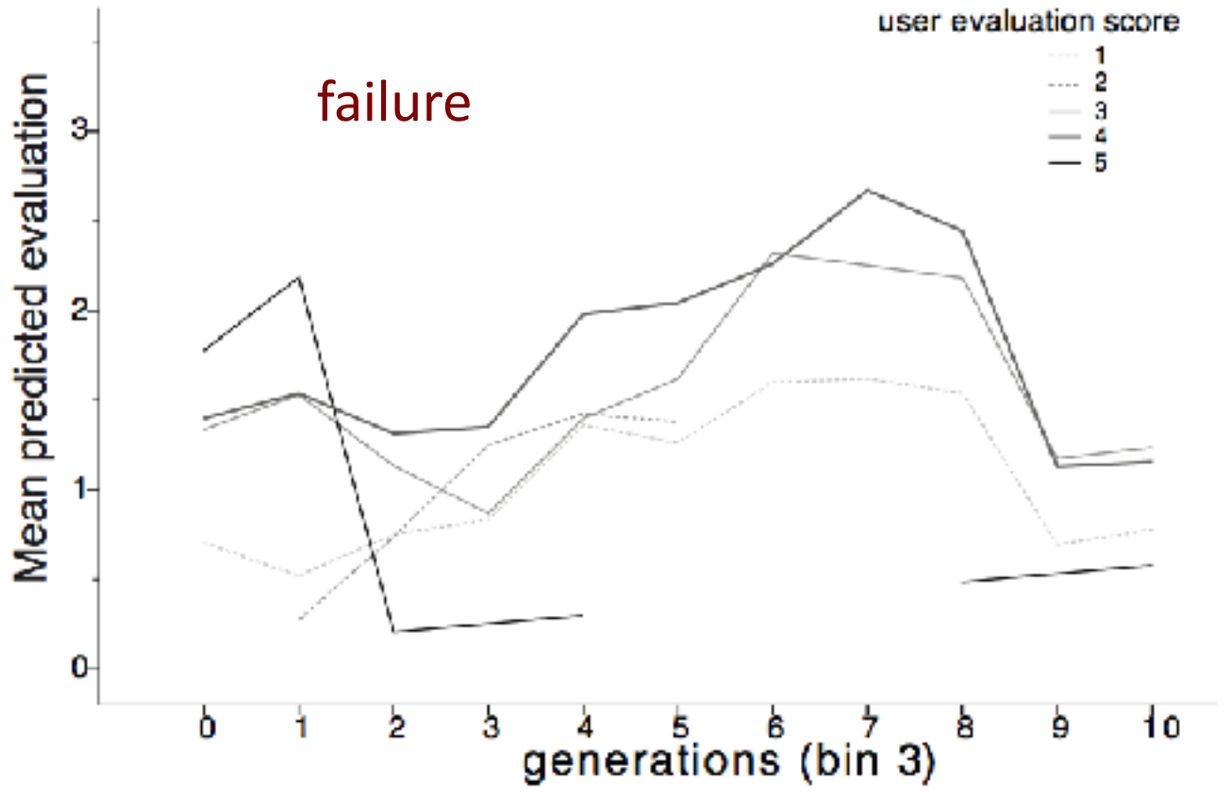
IEA is able to follow the order of user ranking

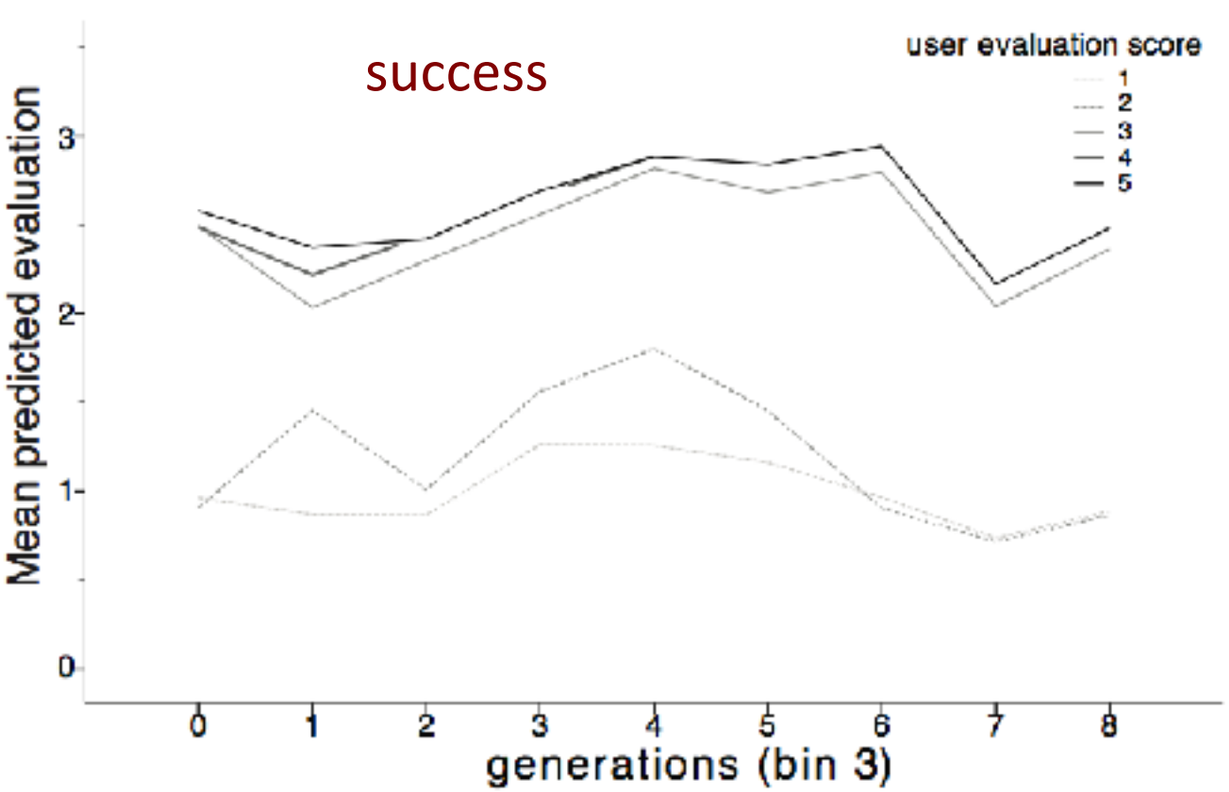


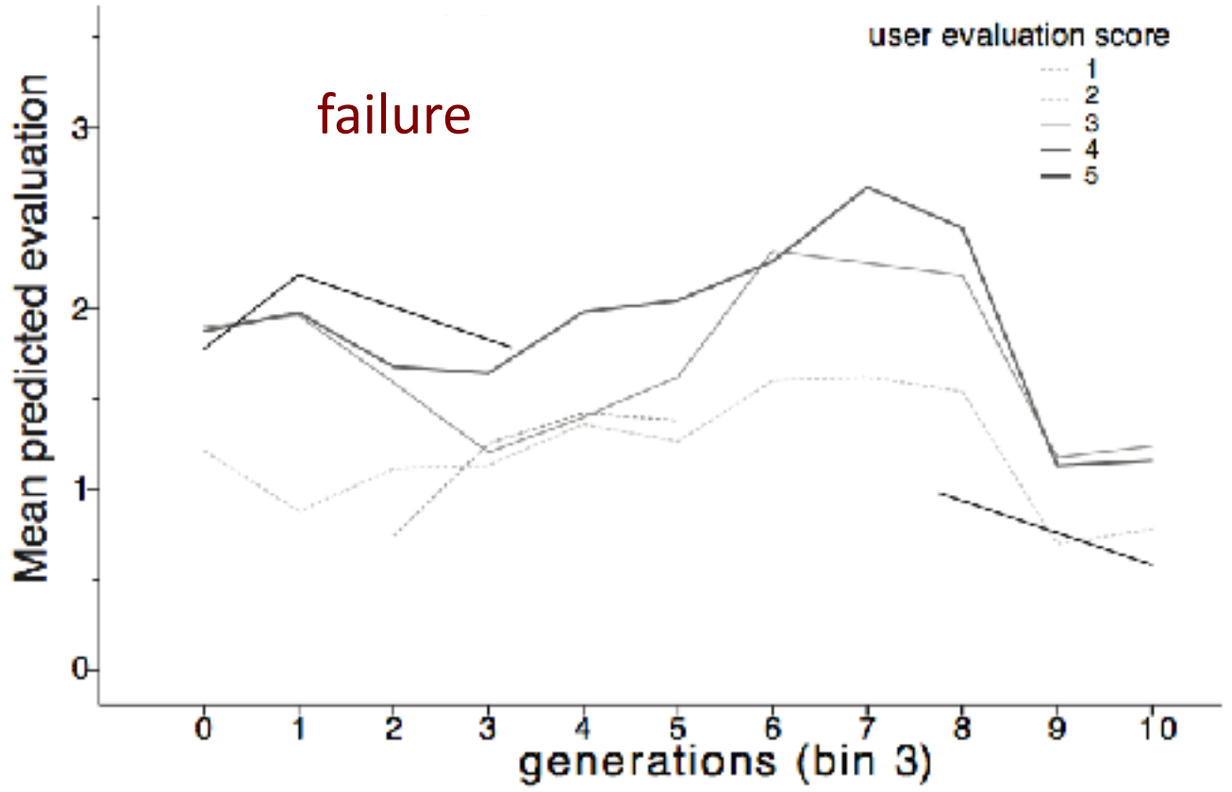


Interface evaluation-use strategies



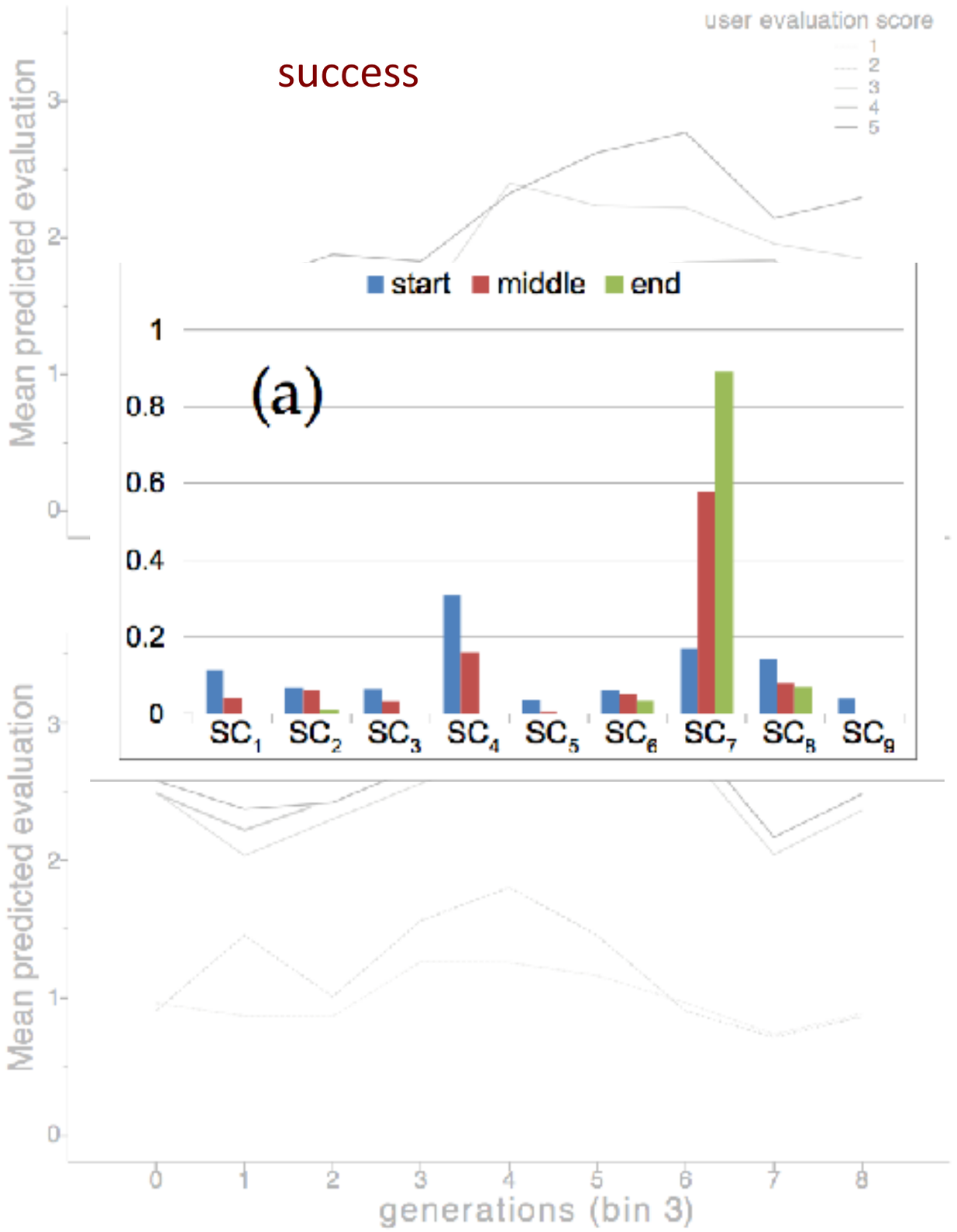




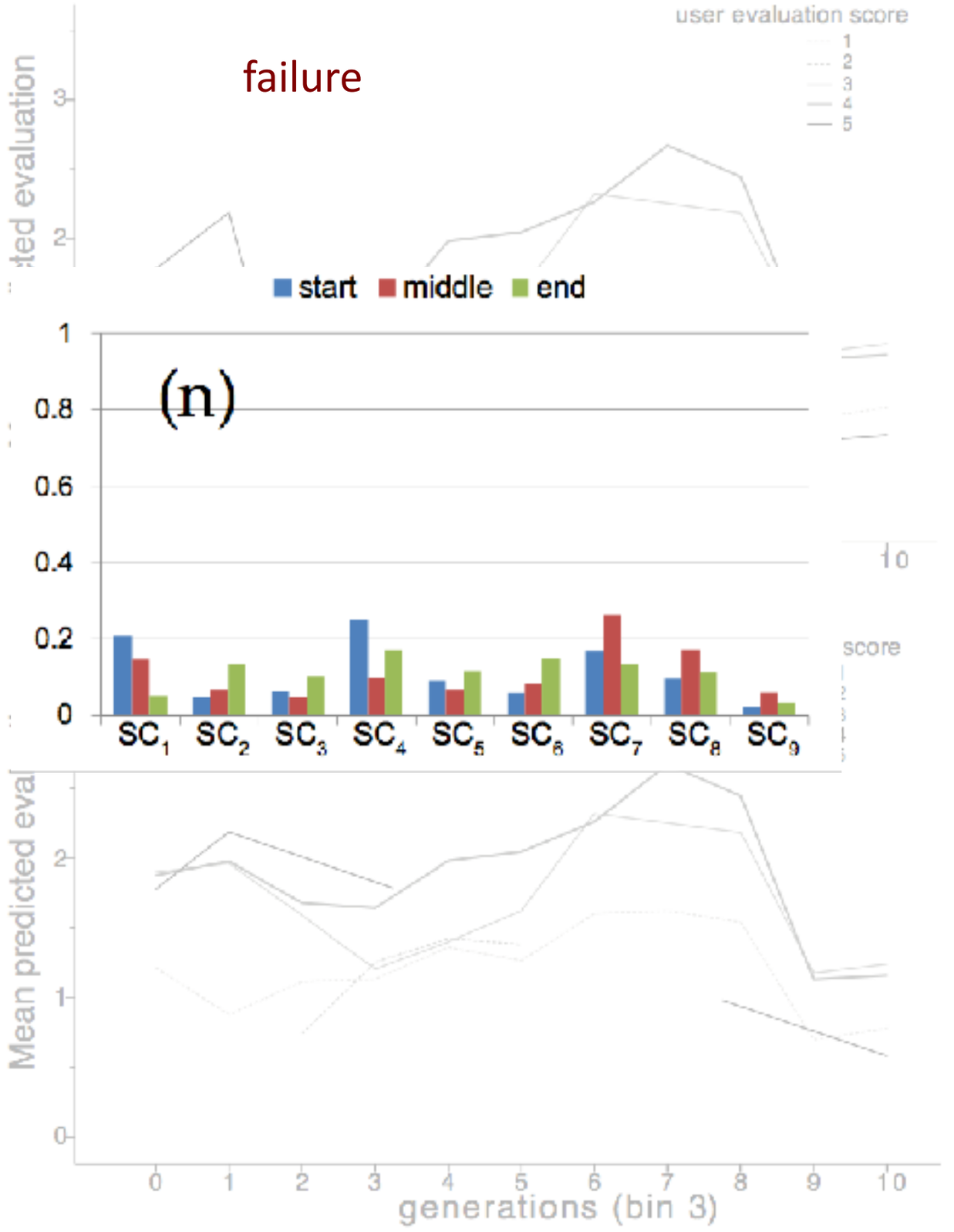


Interface evaluation-use strategies
Visual pattern selection strategies

success



failure



Key findings

Users take different search and evaluation strategies even for a simple task.

On average the surrogate function follows the order of user ranking fairly consistently.

Link between user evaluation strategy, and outcome of exploration & speed of convergence.

Promising results but …

- Real world situations have more complex datasets and tasks.

- Users are not always consistent or give detailed feedback.

- We do not always have ground truth.

User-centred evaluation

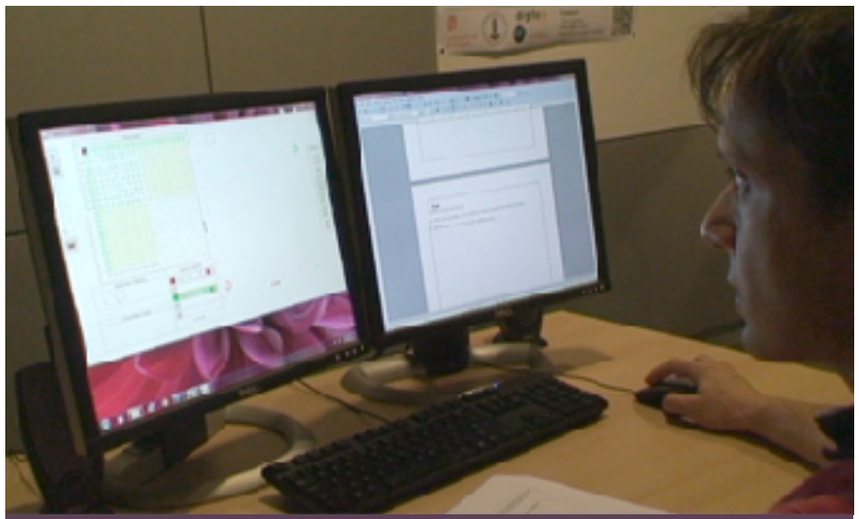
Insight and usability evaluation

- are experts able to **confirm** old knowledge?
- are experts able to gain **new** insight?

Qualitative study methodology

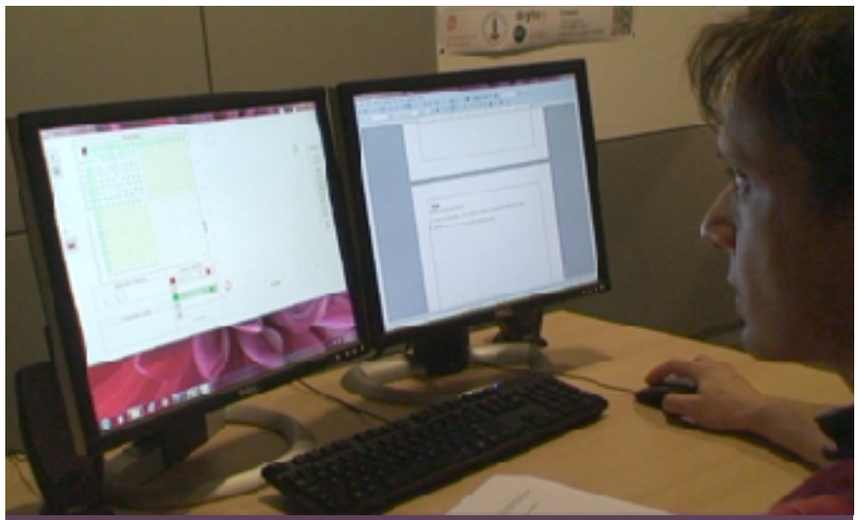
- think aloud, observe, interview & questionnaire
- videotaped and log data capture

Participants



- 5 domain experts
- mean 34.2 years
- own datasets
- pre-questionnaire
- 2.5 hours

Tasks



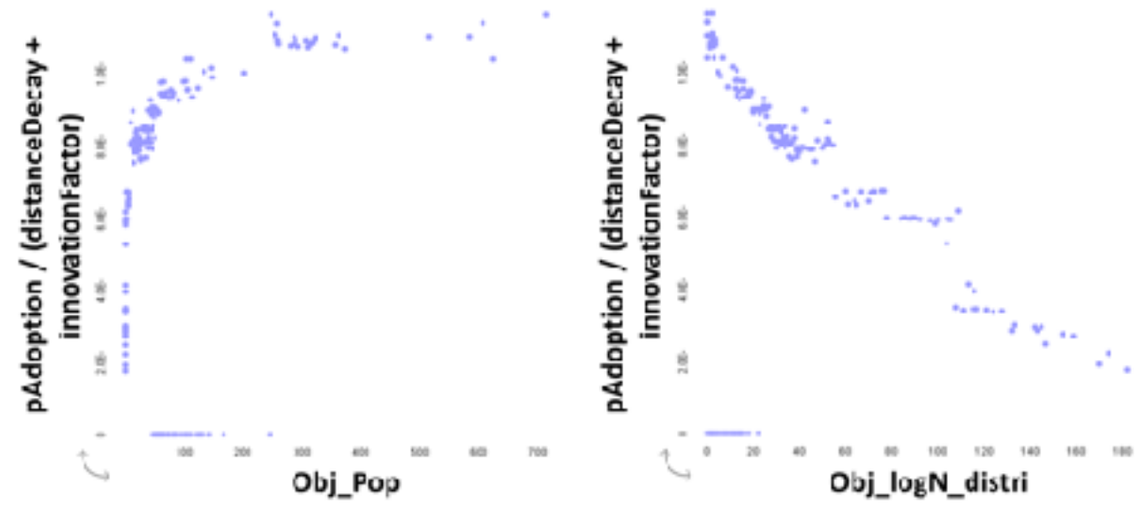
Training

T1: show in the tool what you already know about the data

T2: explore the data in light of a research question

EVE Results

Hypothesis generation

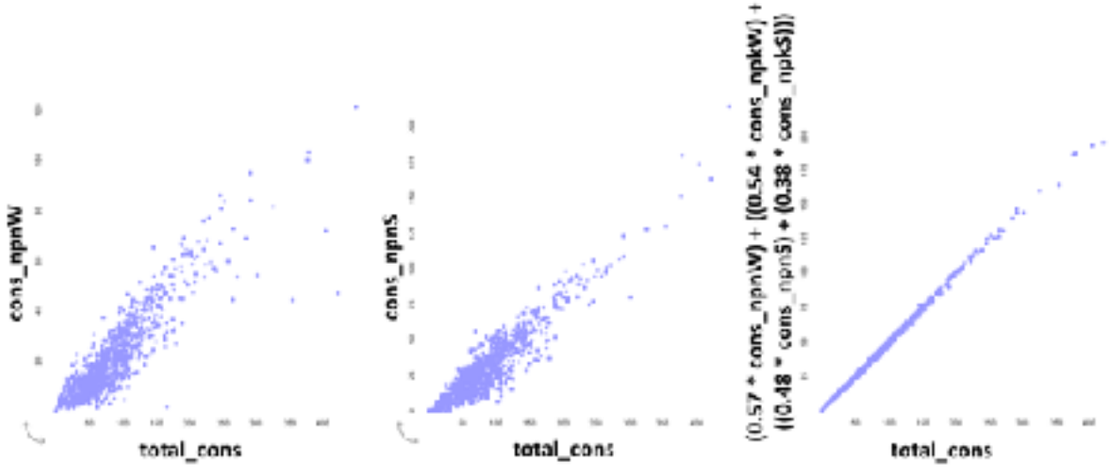


City emergence model

"this combination may be an important finding because it involves parameters that affect only one part of the simulation model ..."

EVE Results

Hypothesis quantification



Electricity consumption profiles

"we always talk about this qualitatively. This is the first time I see concrete weights …"

EVE Results

experts were able to:

- try out alternative scenarios
- think laterally
- quantify a qualitative hypothesis
- formulate a new hypothesis or refine old one
- (domain value)

But …

Evaluation of IVML is Challenging

« On one hand, humans perform unpredictable and sophisticated reasoning; on the other, artificial solvers are technically complex and adopt solving strategies which are very different from those employed by humans. Secondly, the environment from which the problem to be solved is drawn is usually uncontrollable and uncertain. Together, these factors complicate the task of designing precise and effective evaluation studies. »

G. Cortellessa and A. Cesta, AAAI

Cortellessa, Gabriella, and Amedeo Cesta. "Evaluating Mixed-Initiative Systems: An Experimental Approach." *ICAPS*. Vol. 6. 2006.

Evaluation of IVML is Challenging

Ch#1 Complex Human Factors

Ch#2 Multiple Expertise

Ch#3 Stochastic Processes

Ch#4 Co-adaptation

Ch#5 Uncertainty

Ch#1 Complex Human Factors

e.g., Need better techniques to capture user intent.

"Users Are People, Not Oracles"

Amershi et al., Power to the People: The Role of Humans in Interactive Machine Learning , 2014

Frustrated

Reluctant to give feedback

Bored

Annoyed

Fatigue

Inconsistent

Interruptibility

Bias

Ch#2 Multiple Expertise

Should IVML help groups reach consensus ?

Encourage multiple views ?

For us : build common ground and select best trade-offs

Ch#2 Multiple Expertise

Should IVML help groups reach consensus ?

Encourage multiple views ?

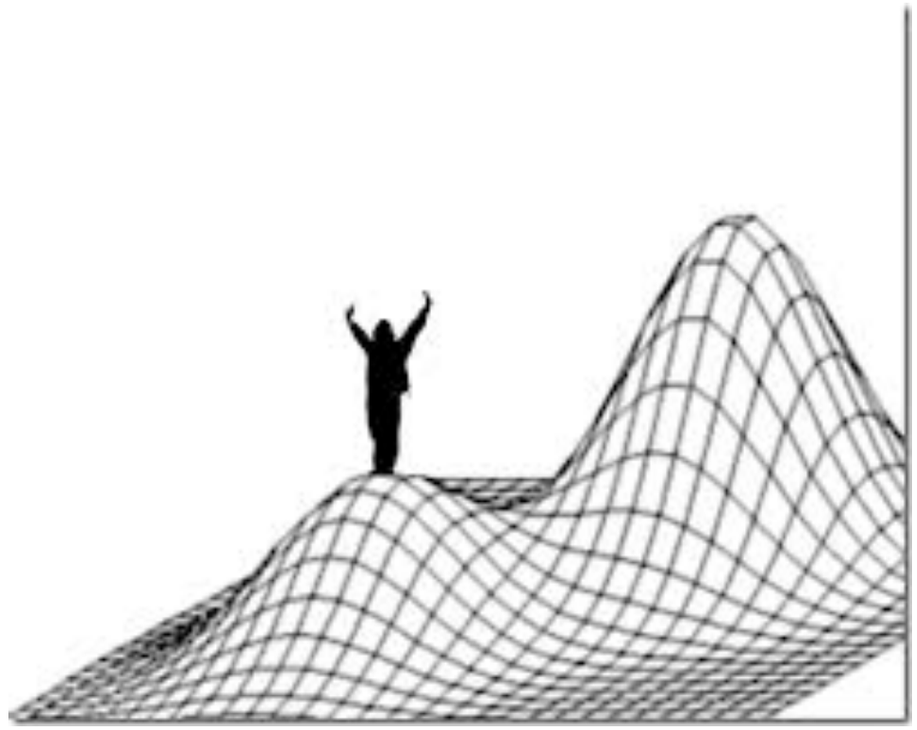
For us : build common ground and select best trade-offs

We need to cater for individual as well as collective exploration

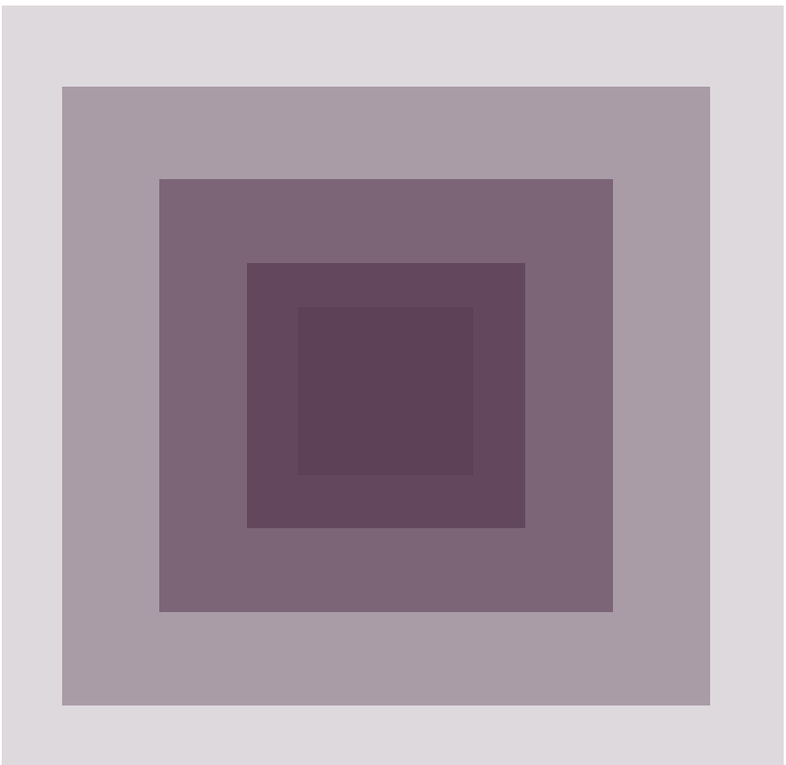
Need setups that can switch between individual and collective learning

Ch#3 Stochastic Processes

- Risk of getting stuck in local optima ?



- Effect of stochasticity on user's mental mode of the ML



Ch#3 Co-adaptation

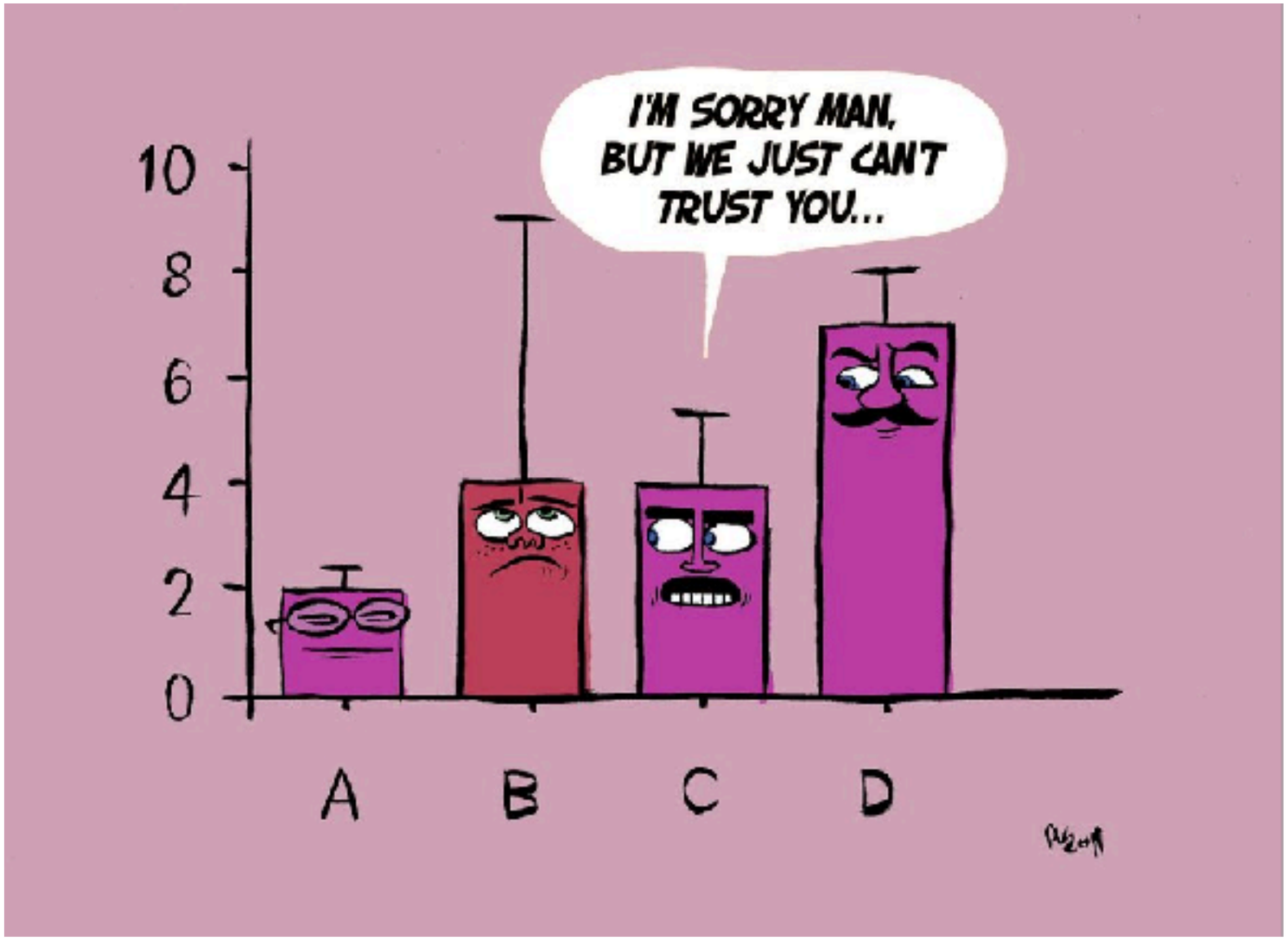
W. E. Mackay, Users and Customizable Software: A Co-Adaptive Phenomenon, 1990.

“ Co-adaptive phenomena are defined as those in which the environment affects human behavior and at the same time, human behavior affects the environment. Such phenomena pose theoretical and methodological challenges and are difficult to study in traditional ways. ”

Ch#4 Uncertainty

Different sources of uncertainty arising from :

- Automatic inferences
- Analysts reasoning



https://www.facebook.com/pedromics

IML Evaluation - HCI Studies

N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, **2018**.

Paper	Classification	Clustering	Density estimation	Dimensionality Reduction	Implicit User Feedback	Explicit User Feedback	System Feedback	Case Study	User Study	Observational Study	Survey	Objective Metrics	Subjective Metrics
Co-integration [2]													
DDLite [20]													
Interest Driven Navigation [25]													
ISSE [11]													
RCLens [34]													
ReGroup [1]													
View Space Explorer [5]													
Visual Classifier [26]													
OLI [48]													
ForceSPIRE [21]													
ForceSPIRE [38]													
RugbyVAST [32]													
3D Model Repository Explorator [22]													
User Interaction Model [19]													
SelPh [30]													
EvoGraphDice [6]													
EvoGraphDice [37]													
Dis-Function [10]													
UTOPIAN [17]													

Review of CHI/VIZ publications 2012-2017

Keyword search: IML + Evaluation

19 papers from different domains

Not an exhaustive survey!

IML Evaluation - HCI Studies

N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, **2018**.

Paper	Classification	Clustering	Density estimation	Dimensionality Reduction	Implicit User Feedback	Explicit User Feedback	System Feedback	Case Study	User Study	Observational Study	Survey	Objective Metrics	Subjective Metrics
Co-integration [2]	✓				✓				✓			✓	
DDLite [20]	✓					✓	✓	✓				✓	✓
Interest Driven Navigation [25]	✓				✓	✓	✓	✓					✓
ISSE [11]	✓					✓			✓		✓	✓	✓
RCLens [34]	✓					✓	✓	✓			✓	✓	✓
ReGroup [1]	✓				✓				✓		✓	✓	✓
View Space Explorer [5]	✓					✓	✓	✓				✓	
Visual Classifier [26]	✓					✓	✓		✓		✓	✓	✓
OLI [48]		✓		✓	✓			✓				✓	
ForceSPIRE [21]		✓			✓			✓				✓	
ForceSPIRE [38]		✓			✓					✓		✓	✓
RugbyVAST [32]		✓				✓	✓	✓	✓			✓	✓
3D Model Repository Explorator [22]		✓				✓		✓	✓			✓	
User Interaction Model [19]		✓			✓		✓		✓		✓	✓	✓
SelPh [30]			✓	✓	✓	✓	✓		✓		✓	✓	✓
EvoGraphDice [6]				✓	✓	✓	✓		✓			✓	
EvoGraphDice [37]				✓	✓	✓	✓			✓	✓	✓	✓
Dis-Function [10]				✓		✓	✓		✓			✓	✓
UTOPIAN [17]				✓	✓		✓	✓				✓	

IML Evaluation - HCI Studies

N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, **2018**.

Human Feedback:

- implicit (7 papers)
- explicit (8 papers)
- mixed (4 papers):
  - implicit human feedback helps infer information to complement implicit human feedback.

Paper	Classification	Clustering	Density estimation	Dimensionality Reduction	Implicit User Feedback	Explicit User Feedback	System Feedback	Case Study	User Study	Observational Study	Survey	Objective Metrics	Subjective Metrics
n [2]	✓				✓				✓			✓	
[20]	✓					✓	✓	✓				✓	✓
[25]	✓				✓	✓	✓	✓					✓
[11]	✓					✓			✓		✓	✓	✓
[34]	✓					✓	✓	✓			✓	✓	✓
o [1]	✓				✓				✓		✓	✓	✓
r [5]	✓					✓	✓	✓				✓	
[26]	✓					✓	✓		✓		✓	✓	✓
[48]		✓		✓	✓			✓				✓	
[21]		✓			✓			✓				✓	
[38]		✓			✓					✓		✓	✓
[32]		✓				✓	✓	✓	✓			✓	✓
[22]		✓				✓		✓	✓			✓	
[19]		✓			✓		✓		✓		✓	✓	✓
[30]			✓	✓	✓	✓	✓		✓		✓	✓	✓
e [6]				✓	✓	✓	✓		✓			✓	
[37]				✓	✓	✓	✓			✓	✓	✓	✓
[10]				✓		✓	✓		✓			✓	✓
[17]				✓	✓		✓	✓				✓	

IML Evaluation - HCI Studies

N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, **2018**.

Paper	Classification	Clustering	Density estimation	Dimensionality Reduction	Implicit User Feedback	Explicit User Feedback	System Feedback	Case Study	User Study	Observational Study	Survey	Objective Metrics	Subjective Metrics
n [2]	✓				✓				✓			✓	
[20]	✓					✓	✓	✓				✓	✓
[25]	✓				✓	✓	✓	✓					✓
[11]	✓					✓			✓		✓	✓	✓
[34]	✓					✓	✓	✓			✓	✓	✓
o [1]	✓				✓				✓		✓	✓	✓
r [5]	✓					✓	✓	✓				✓	
[26]	✓					✓	✓		✓		✓	✓	✓
[48]		✓		✓	✓			✓				✓	
[21]		✓			✓			✓				✓	
[38]		✓			✓					✓		✓	✓
[32]		✓				✓	✓	✓	✓			✓	✓
[22]		✓				✓		✓	✓			✓	
[19]		✓			✓		✓		✓		✓	✓	✓
[30]			✓	✓	✓	✓	✓		✓		✓	✓	✓
e [6]				✓	✓	✓	✓		✓			✓	
[37]				✓	✓	✓	✓			✓	✓	✓	✓
[10]				✓		✓	✓		✓			✓	✓
[17]				✓	✓		✓	✓				✓	

System Feedback:

- to inform humans about the state of the machine learning algorithm, and provenance of system suggestions.
- can be visual, progressive, and can indicate uncertainty.
- most systems gave feedback.
- challenge: to inform without overwhelming the user.

IML Evaluation - HCI Studies

N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, **2018**.

Types of study:

- 12/19 studies involving users (some for of controlled study)
- Difficult to conduct:
  - potential confounding factors
  - no ground truth

Paper	Classification	Clustering	Density estimation	Dimensionality Reduction	Implicit User Feedback	Explicit User Feedback	System Feedback	Case Study	User Study	Observational Study	Survey	Objective Metrics	Subjective Metrics
n [2]	✓				✓				✓			✓	
[20]	✓					✓	✓	✓				✓	✓
[25]	✓				✓	✓	✓	✓					✓
[11]	✓					✓			✓		✓	✓	✓
[34]	✓					✓	✓	✓			✓	✓	✓
) [1]	✓				✓				✓		✓	✓	✓
r [5]	✓					✓	✓	✓				✓	
[26]	✓					✓	✓		✓		✓	✓	✓
[48]		✓		✓	✓			✓				✓	
[21]		✓			✓			✓				✓	
[38]		✓			✓					✓		✓	✓
[32]		✓				✓	✓	✓	✓			✓	✓
[22]		✓				✓		✓	✓			✓	
[19]		✓			✓		✓		✓		✓	✓	✓
[30]			✓	✓	✓	✓	✓		✓		✓	✓	✓
[6]				✓	✓	✓	✓		✓			✓	
[37]				✓	✓	✓	✓			✓	✓	✓	✓
[10]				✓		✓	✓		✓			✓	✓
[17]				✓	✓		✓	✓				✓	

IML Evaluation - HCI Studies

N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, **2018**.

Evaluation Metrics :

objective:

- time, precision, re-call, #Insights
- iML vs. baseline or variants of ML; with/out system feedback; explicit vs. implicit; impact of user feedback
- difficulty separating usability issues from task results

Paper	Classification	Clustering	Density estimation	Dimensionality Reduction	Implicit User Feedback	Explicit User Feedback	System Feedback	Case Study	User Study	Observational Study	Survey	Objective Metrics	Subjective Metrics
n [2]	✓				✓				✓			✓	
[20]	✓					✓	✓	✓				✓	✓
[25]	✓				✓	✓	✓	✓					✓
[11]	✓					✓			✓		✓	✓	✓
[34]	✓					✓	✓	✓			✓	✓	✓
o [1]	✓				✓				✓		✓	✓	✓
r [5]	✓					✓	✓	✓				✓	
[26]	✓					✓	✓		✓		✓	✓	✓
[48]		✓	✓		✓			✓				✓	
[21]		✓			✓			✓				✓	
[38]		✓			✓					✓		✓	✓
[32]		✓				✓	✓	✓	✓			✓	✓
[22]		✓				✓		✓	✓			✓	
[19]		✓			✓		✓		✓		✓	✓	✓
[30]			✓	✓	✓	✓	✓		✓		✓	✓	✓
c [6]				✓	✓	✓	✓		✓			✓	
[37]				✓	✓	✓	✓			✓	✓	✓	✓
[10]				✓		✓	✓		✓			✓	✓
[17]				✓	✓		✓	✓				✓	

IML Evaluation - HCI Studies

N. Boukhelifa, A. Bezerianos, and E. Lutton. Evaluation of Interactive Machine Learning Systems. In Human and Machine Learning, pp. 341-360. Springer, Cham, **2018**.

Evaluation Metrics :

subjective:

- aspects of user experience, e.g., reported, easiness, speed, task load, trust, and confidence.

Paper	Classification	Clustering	Density estimation	Dimensionality Reduction	Implicit User Feedback	Explicit User Feedback	System Feedback	Case Study	User Study	Observational Study	Survey	Objective Metrics	Subjective Metrics
n [2]	✓				✓				✓			✓	
[20]	✓					✓	✓	✓				✓	✓
[25]	✓				✓	✓	✓	✓					✓
[11]	✓					✓			✓		✓	✓	✓
[34]	✓					✓	✓	✓			✓	✓	✓
o [1]	✓				✓				✓		✓	✓	✓
r [5]	✓					✓	✓	✓				✓	
[26]	✓					✓	✓		✓		✓	✓	✓
[48]		✓		✓	✓			✓				✓	
[21]		✓			✓			✓				✓	
[38]		✓			✓					✓		✓	✓
[32]		✓				✓	✓	✓	✓			✓	✓
[22]		✓				✓		✓	✓			✓	
[19]		✓			✓		✓		✓		✓	✓	✓
[30]			✓	✓	✓	✓	✓		✓		✓	✓	✓
c [6]				✓	✓	✓	✓		✓			✓	
[37]				✓	✓	✓	✓			✓	✓	✓	✓
[10]				✓		✓	✓		✓			✓	✓
[17]				✓	✓		✓	✓				✓	

"One sign of success of iML systems is when humans forget that they are feeding information to an algorithm, and rather focus on synthesising information relevant to their task".

Endert et al., 2012 [38]

Evaluation Metrics :

subjective:

- aspects of user experience, e.g., reported, easiness, speed, task load, trust, and confidence.

[2]	✓				✓				✓			✓	
[20]	✓					✓	✓	✓				✓	✓
[25]	✓				✓	✓	✓	✓					✓
[11]	✓					✓			✓		✓	✓	✓
[34]	✓					✓	✓	✓			✓	✓	✓
[1]	✓				✓				✓		✓	✓	✓
[5]	✓					✓	✓	✓				✓	
[26]	✓					✓	✓		✓		✓	✓	✓
[48]		✓		✓	✓			✓				✓	
[21]		✓			✓			✓				✓	
[38]		✓			✓					✓		✓	✓
[32]		✓				✓	✓	✓	✓			✓	✓
[22]		✓				✓		✓	✓			✓	
[19]		✓			✓		✓		✓		✓	✓	✓
[30]			✓	✓	✓	✓	✓		✓		✓	✓	✓
[6]				✓	✓	✓	✓		✓			✓	
[37]				✓	✓	✓	✓			✓	✓	✓	✓
[10]				✓		✓	✓		✓			✓	✓
[17]				✓	✓		✓	✓				✓	

Guidelines for IVML evaluation

We know how to evaluate interfaces & interactive systems, but very few guidelines exist specifically for iML systems!

Jakob Nielsen's 10 Usability Heuristics for User Interface Design

G1: Visibility of system status
G2: Match between system and the real world
G3: User control and freedom
G4: Consistency and standards
G5: Error prevention
G6: Recognition rather than recall
G7: Flexibility and efficiency of use
G8: Aesthetic and minimalist design
G9: Help users recognize, diagnose, & recover from errors
G10: Help and documentation

Guidelines for IVML evaluation

Principles of Mixed-Initiative User Interfaces - Eric Horvitz, 1999

G1 Developing significant value-added automation.

G2 Considering uncertainty about a user's goals.

G3 Considering the status of a user's attention in the timing of services.

G4 Inferring ideal action in light of costs, benefits, and uncertainties.

G5 Employing dialog to resolve key uncertainties.

G6 Allowing efficient direct invocation and termination.

G7 Minimizing the cost of poor guesses about action and timing.

G8 Scoping precision of service to match uncertainty, variation in goals.

G9 Providing mechanisms for efficient agent–user collaboration to refine results.

G10 Employing socially appropriate behaviors for agent–user interaction.

G11 Maintaining working memory of recent interactions.

G12 Continuing to learn by observing.

Guidelines for IVML evaluation

New guidelines proposed, e.g., Amershi et al., 2019.

G1 Make clear what the system can do.

G2 Make clear how well the system can do what it can do.

G3 Time services based on context.

G4 Show contextually relevant information.

G5 Match relevant social norms.

G6 Mitigate social biases.

G7 Support efcient invocation.

G8 Support efcient dismissal.

G9 Support efcient correction.

G10 Scope services when in doubt.

G11 Make clear why the system did what it did.

G12 Remember recent interactions.

G13 Learn from user behavior.

G14 Update and adapt cautiously.

G15 Encourage granular feedback.

G16 Convey the consequences of user actions.

G17 Provide global controls.

G18 Notify users about changes.

No conclusions - research questions!

Q1 What aspects or components of the IVML system are most important to evaluate?

Q2 What types of tasks can be delegated to machine learning and which are best left to humans?

Q3 Who are the target users of IVML systems?

Q4 What is the role of expertise in this context?

Q5 Should we have domain experts train IVML systems?

Q6 What are the risks and benefits of introducing human expertise into the (machine) learning process?

Q7 What metrics should we use to evaluate?

Q8 Can we establish application or domain-independent metrics?

Q9 Do we establish different evaluations measures for the understanding of IVML systems and for their performance?

Q10 Can we establish benchmark data sets and test use-cases to help evaluate IVML systems?

Q11 Should we seek replication in this context, and if so, how do we support replication of results in a co-learning or adaptive environment?

Q12 How do we communicate the evaluation results to other disciplines?

iML-eval : On-going Research Topic

eviva-ml.github.io

EVIVA-ML

CONTACT US

21 October, 2019

EValuation of Interactive VisuAl Machine Learning systems

IEEE VIS 2019 WORKSHOP
OCTOBER 21, 2019 IN VANCOUVER, BC, CANADA

The goal of the EVIVA-ML workshop is to bring together visualization researchers and practitioners to discuss experiences and viewpoints on how to effectively evaluate interactive visual machine learning systems.

EVIVA-ML

Early registration deadline 20/09

CONTACT US

eviva-ml.github.io

EValuation of Interactive VisuAl Machine Learning systems

IEEE VIS 2019 WORKSHOP
OCTOBER 21, 2019 IN VANCOUVER, BC, CANADA

Organizing Committee


- Nadia Boukhelifa (INRA, FR)
- Anastasia Bezerianos (Univ. Paris-Sud and INRIA, FR)
- Enrico Bertini (NYU Tandon School of Engineering, USA)
- Christopher Collins (Uni. of Ontario Institute of Technology, CA)
- Steven Drucker (Microsoft Research, USA)
- Alex Endert (Georgia Tech, USA)
- Jessica Hullman (Northwestern University, USA)
- Michael Sedlmair (University of Stuttgart, DE)
- Remco Chang (Tufts University, USA)
- Chris North (Virginia Tech, USA)

No conclusions - research questions!

Danke!

Q1 What aspects or components of the IVML system are most important to evaluate?

Q2 What types of tasks can be delegated to machine learning and which are best left to humans?

Q3 Who are the target users of IVML systems?

Q4 What is the role of expertise in this context?

Q5 Should we have domain experts train IVML systems?

Q6 What are the risks and benefits of introducing human expertise into the (machine) learning process?

Q7 What metrics should we use to evaluate?

Q8 Can we establish application or domain-independent metrics?

Q9 Do we establish different evaluations measures for the understanding of IVML systems and for their performance?

Q10 Can we establish benchmark data sets and test use-cases to help evaluate IVML systems?

Q11 Should we seek replication in this context, and if so, how do we support replication of results in a co-learning or adaptive environment?

Q12 How do we communicate the evaluation results to other disciplines?