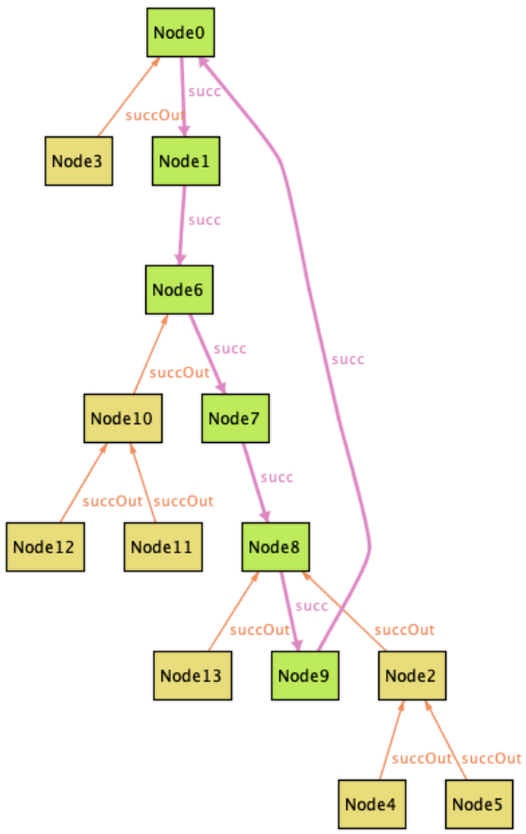
EMPIRICAL USER STUDIES

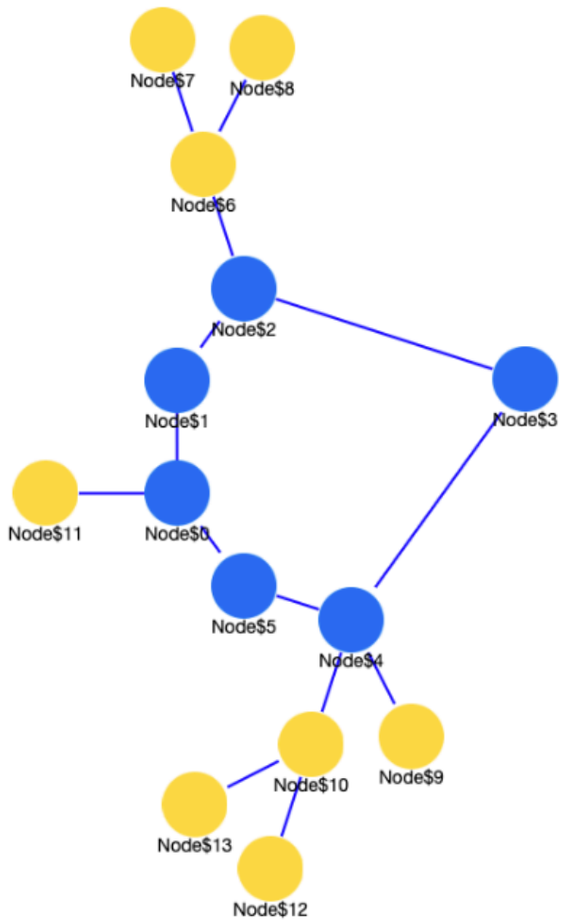
Evaluation of the Anima tool

Rui Couto • José C. Campos

OUTLINE

- The problem
- Objective specification
- Study design
- Elaboration
- Evaluation
- Publication
- Case study: OutSystems Learnability model



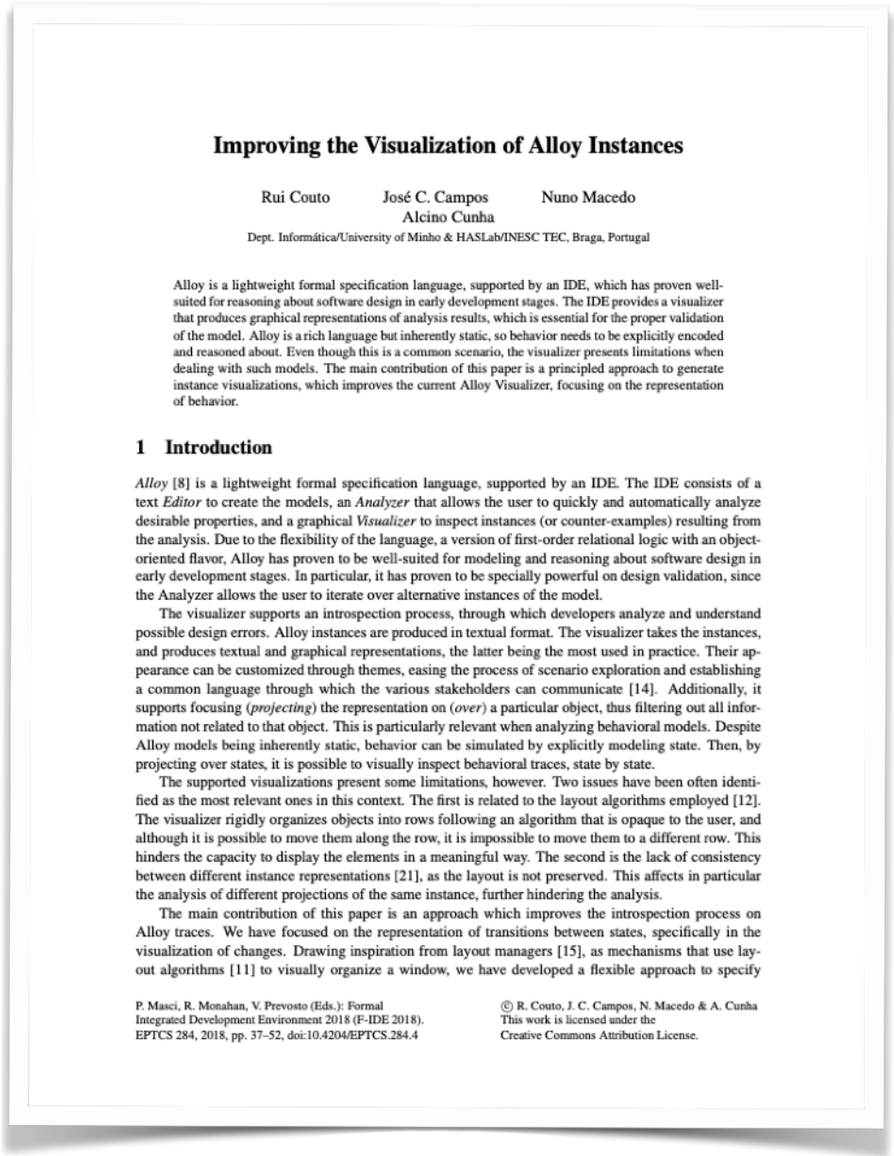


THE PROBLEM

- The motivation behind a user study is an existing problem in a given software product
- In our case, we have a concrete problem:
  - The Alloy visualiser is hard to use, not appealing, and we believe it leads to errors.
- Our solution was to develop Anima, a new visualiser which improves the Alloy visualiser, with features to tackle the identified issues

THE PROBLEM

- Since Anima was designed to solve the Alloy visualiser issues, we believe to have a better solution[1].
- However…
  - We don't know for sure if such is true, or if it is only our perception - we have developed the tool
  - A small group of users, which know the problems in beforehand isn't a significative group
  - Solving the identified issues is not a guarantee to have a better solution: introduction of new problems, solving false problems, etc.



Improving the Visualization of Alloy Instances

Rui Couto José C. Campos Nuno Macedo
Alcino Cunha
Dept. Informática/University of Minho & HASLab/INESC TEC, Braga, Portugal

Alloy is a lightweight formal specification language, supported by an IDE, which has proven well-suited for reasoning about software design in early development stages. The IDE provides a visualizer that produces graphical representations of analysis results, which is essential for the proper validation of the model. Alloy is a rich language but inherently static, so behavior needs to be explicitly encoded and reasoned about. Even though this is a common scenario, the visualizer presents limitations when dealing with such models. The main contribution of this paper is a principled approach to generate instance visualizations, which improves the current Alloy Visualizer, focusing on the representation of behavior.

1 Introduction

Alloy [8] is a lightweight formal specification language, supported by an IDE. The IDE consists of a text *Editor* to create the models, an *Analyzer* that allows the user to quickly and automatically analyze desirable properties, and a graphical *Visualizer* to inspect instances (or counter-examples) resulting from the analysis. Due to the flexibility of the language, a version of first-order relational logic with an object-oriented flavor, Alloy has proven to be well-suited for modeling and reasoning about software design in early development stages. In particular, it has proven to be specially powerful on design validation, since the Analyzer allows the user to iterate over alternative instances of the model.

The visualizer supports an introspection process, through which developers analyze and understand possible design errors. Alloy instances are produced in textual format. The visualizer takes the instances, and produces textual and graphical representations, the latter being the most used in practice. Their appearance can be customized through themes, easing the process of scenario exploration and establishing a common language through which the various stakeholders can communicate [14]. Additionally, it supports focusing (*projecting*) the representation on (*over*) a particular object, thus filtering out all information not related to that object. This is particularly relevant when analyzing behavioral models. Despite Alloy models being inherently static, behavior can be simulated by explicitly modeling state. Then, by projecting over states, it is possible to visually inspect behavioral traces, state by state.

The supported visualizations present some limitations, however. Two issues have been often identified as the most relevant ones in this context. The first is related to the layout algorithms employed [12]. The visualizer rigidly organizes objects into rows following an algorithm that is opaque to the user, and although it is possible to move them along the row, it is impossible to move them to a different row. This hinders the capacity to display the elements in a meaningful way. The second is the lack of consistency between different instance representations [21], as the layout is not preserved. This affects in particular the analysis of different projections of the same instance, further hindering the analysis.

The main contribution of this paper is an approach which improves the introspection process on Alloy traces. We have focused on the representation of transitions between states, specifically in the visualization of changes. Drawing inspiration from layout managers [15], as mechanisms that use layout algorithms [11] to visually organize a window, we have developed a flexible approach to specify

P. Masci, R. Monahan, V. Prevosto (Eds.): Formal Integrated Development Environment 2018 (F-IDE 2018). EPTCS 284, 2018, pp. 37–52, doi:10.4204/EPTCS.284.4

1 Couto, Rui, et al. "Improving the Visualization of Alloy Instances." *arXiv preprint arXiv:1811.10817* (2018).

THE PROBLEM

- The solution to address this problem is to:
  - Perform a user study with independent participants (reducing bias)
  - Measure different dimensions such as performance, time, and effort in both visualisers (gather objective data)
- Compare the results (achieve concrete results)

OBJECTIVES

- The user study is shaped according to the objectives to achieve
- The objectives define the dimensions to be measured
  - E.g. How to evaluate a compiler? And a visualisation tool?
    - Are we measuring single performance, or comparing tools?
- In this case, we are evaluating if Anima was able to improve the analysis process, when compare with Alloy.

OBJECTIVES

- We define that, in order for a visualiser to be better than the other, the following aspects should be considered:

Time - required to perform tasks

Number of errors - resulting from interpretation mistakes

Effort - required to perform the tasks

Overall perception - that the users had for both tools

- Having shaped the objectives, we were able to design the study accordingly.

OBJECTIVES

Time

- If a user takes less time to perform a task, then the tool is more effective
- More time spent in a task, might result in fatigue, and increase the number of errors

Number of errors

- These tools have the objective of showing possible errors in systems
- Performing an error means the possibility for an error in the final solution.
- Having fewer errors is one of the most relevant objectives

OBJECTIVES

Effort

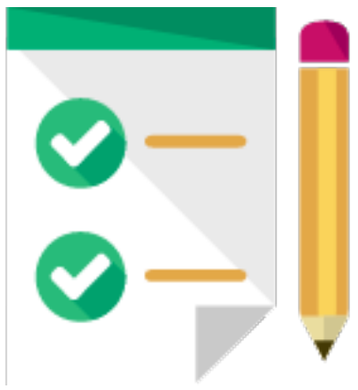
- A higher effort is associated with a higher probability for interpretation errors
- A lower effort in the interpretation means the user can focus in the interpretation of the problem itself

Overall perception

- While not an easily quantifiable measure, if the users have a bad perception of a tool, they will be less prone to use it
- A higher perception might result in a higher adoption ratio

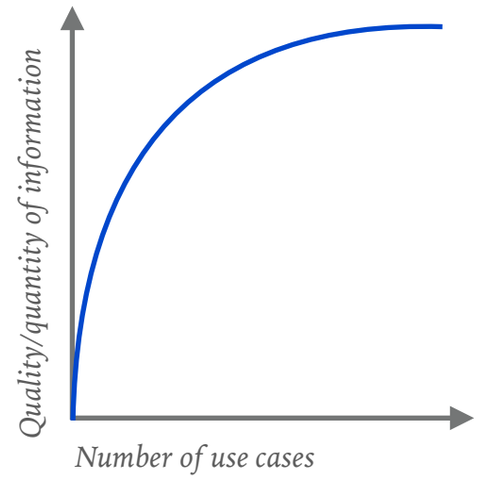
STUDY DESIGN

- Designing the study consists in:
  1. Defining the case study to be performed
  2. Defining how to measure the defined dimensions
  3. Predict possible problems
  4. Performing pilot studies



STUDY DESIGN – 1 DEFINING THE CASE STUDY TO BE PERFORMED

- The case study should be relevant, i.e. illustrate real usage scenarios
- If we design a use case based in the weaknesses/ strength of our tool, we will be hindering/promote results of the other tool
- Higher number of studies provides more uniform results
  - Time to perform a study is limited, after ~1.5h users will start to show signs of fatigue, and results will not be useful
- Lower number of results provides less objective results
  - Results from two use cases can be drastically different

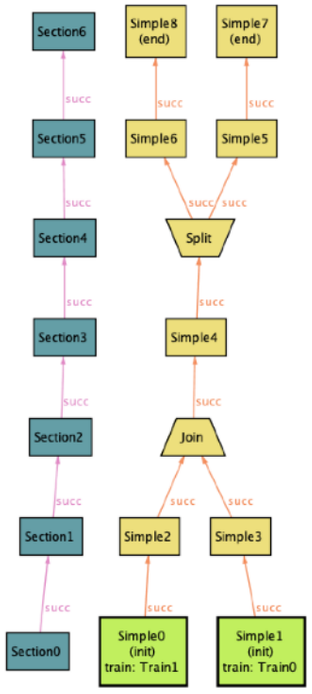


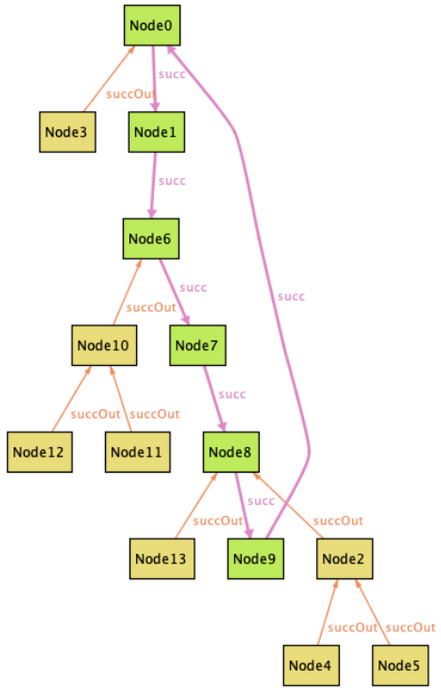
STUDY DESIGN – 1 DEFINING THE CASE STUDY TO BE PERFORMED

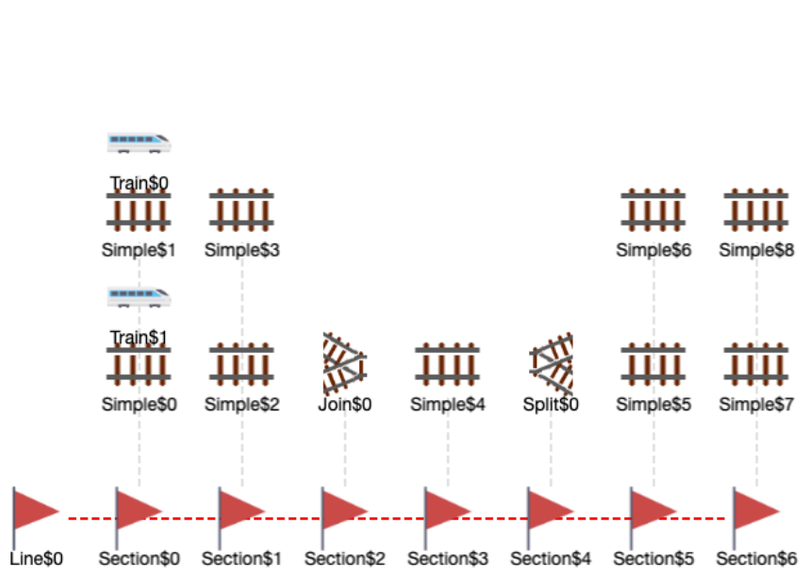
- We have used two use cases, different in nature and context
  - A higher number of use cases would be too costly
  - Fewer user cases would provide biased results towards the use case
  - We required, at least two scenarios, in order to test the customisation capabilities of Anima.

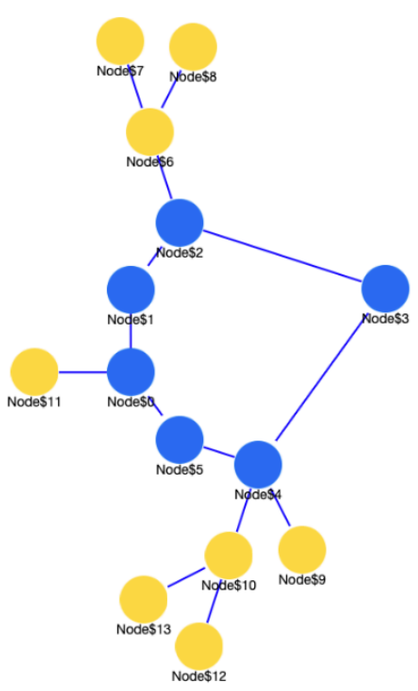
STUDY DESIGN – 1 DEFINING THE CASE STUDY TO BE PERFORMED

- We have used two existing scenarios:
  - ERTMS/ETCS level 3[1] - a model of a train line, previously published
  - Chord[2] - A well known model in the Alloy community





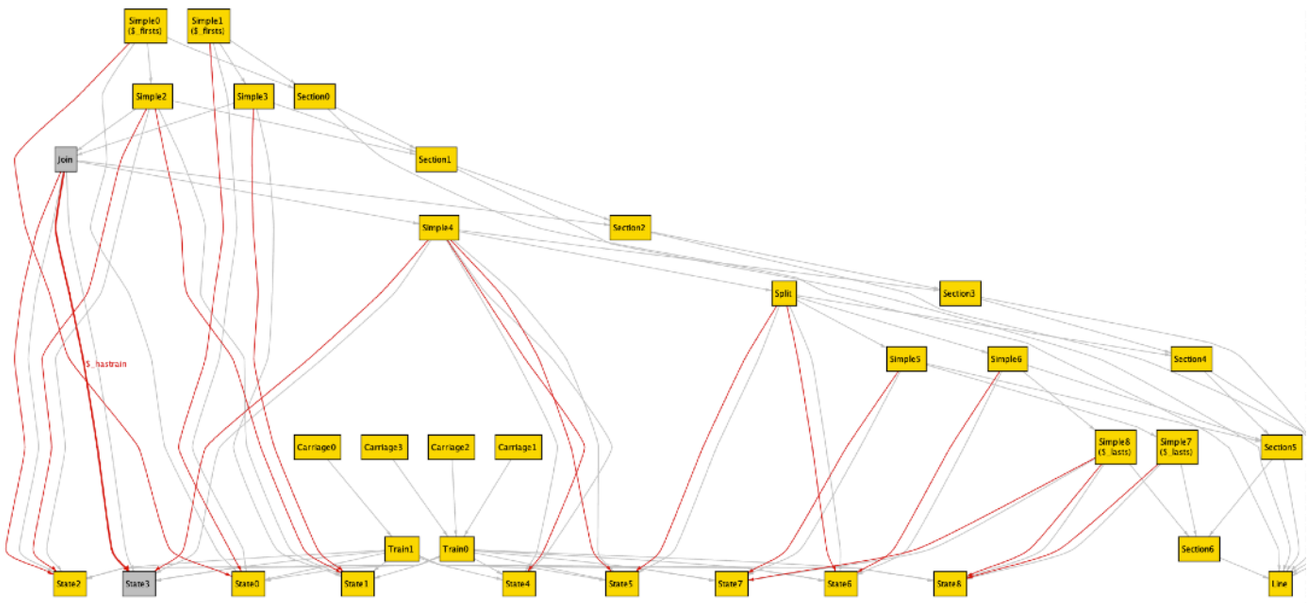


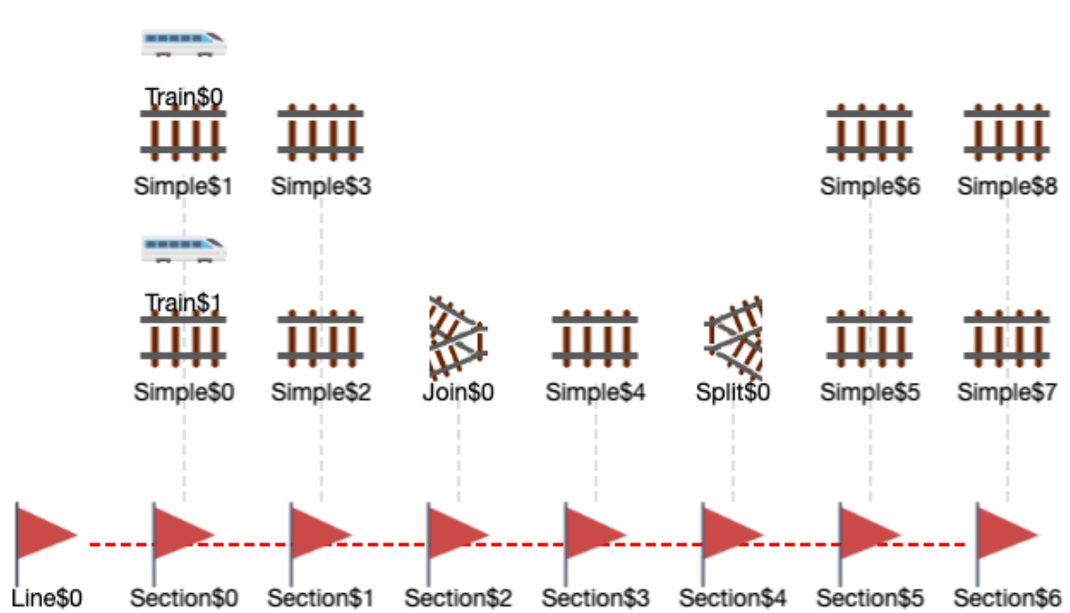


1. A. Cunha and N. Macedo. Validating the hybrid ERTMS/ETCS level 3 concept with electrum. In M. Butler, A. Raschke, T. S. Hoang, and K. Reichl, editors, *Abstract State Machines, Alloy, B, TLA, VDM, and Z*, pages 307–321, Cham, 2018. Springer International Publishing.
2. y, "Using lightweight modeling to understand Chord" (Pamela Zave; ACM SIGCOMM Computer Communication Review, 42(2):50-57, April 2012)

STUDY DESIGN – 1 DEFINING THE CASE STUDY TO BE PERFORMED

- Fairness should be considered as well.
- Both Anima and Alloy tools allow for customisation
  - We should not use a non-themed version of the Alloy representation, vs a themed version of Anima
  - We have asked an Alloy expert to create the former one





1. A. Cunha and N. Macedo. Validating the hybrid ERTMS/ETCS level 3 concept with electrum. In M. Butler, A. Raschke, T. S. Hoang, and K. Reichl, editors, *Abstract State Machines, Alloy, B, TLA, VDM, and Z*, pages 307–321, Cham, 2018. Springer International Publishing.

2. y, "Using lightweight modeling to understand Chord" (Pamela Zave; ACM SIGCOMM Computer Communication Review, 42(2):50-57, April 2012)

STUDY DESIGN – 2 DEFINING HOW TO MEASURE THE DEFINED DIMENSIONS

- In order to produce relevant information, different outputs should be gathered.
- Some are objective:
  - Time is “easily” measured, as it can be objectively observed
  - Registering the number of errors is also an objective process
- We have measured time and errors as directly observable outputs
- Other are harder to measure:
  - Is the user understanding the problem?
  - Is the user under stress?
- We have observed the users to take further conclusions.

STUDY DESIGN – 2 DEFINING HOW TO MEASURE THE DEFINED DIMENSIONS

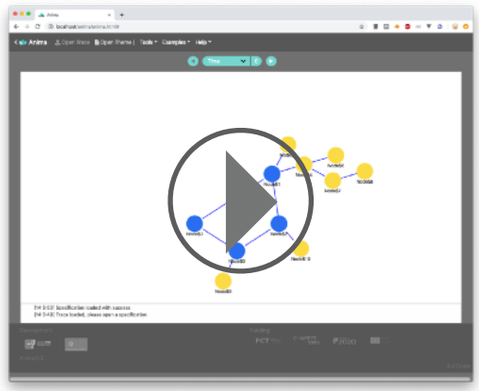
- Different techniques should be used.
- Direct observation is a non-intrusive, objective data gathering technique, e.g. time
  - For non objective data, interpretation is required, e.g. is the user understanding the problem?
- Interaction with the user helps to understand hard to observe aspects
  - Is the user really understanding the problem? Is the user struggling? Did the user make a mistake due to an interpretation error?
  - Risk of influencing the user, and affecting the final results.

STUDY DESIGN – 2 DEFINING HOW TO MEASURE THE DEFINED DIMENSIONS

- Standard questionnaires are well accepted
- They have been widely used, so a large amount of comparable information is available
- We decided to use two different questionnaires:
  - NASA TLX - measures the task load, i.e. how hard were the tasks to perform.
  - User Experience Questionnaire (UEQ) - measures the quality of the representations.
- These standard questionnaires allow us to disseminate our results in a way that other researchers understand

STUDY DESIGN – 2 DEFINING HOW TO MEASURE THE DEFINED DIMENSIONS

- Screen, audio and video are good sources for later analysis
  - They allow a deeper analysis, record statements e.g. “I like this feature”, or “This is hard to perform”
  - Can be viewed as many times as needed



- However…
  - They cause stress in the participants. Knowing that their interactions are being record, might affect their behaviour
  - Certain aspects (e.g. frustration) are easier to capture face to face
  - There are privacy concerns to be taken in consideration - some users don’t like recordings, specially video; there are legal aspects to be taken in consideration

STUDY DESIGN – 2 DEFINING HOW TO MEASURE THE DEFINED DIMENSIONS

- We tried to balance feedback record
  - We didn't record video, as we think it is the most intrusive
  - We record audio, as the participants did not seem affected by it
  - We recorded the screen, as it is the least intrusive recording media.
- Through direct observation and data logging we record remaining outputs.

STUDY DESIGN – 3 PREDICT POSSIBLE PROBLEMS

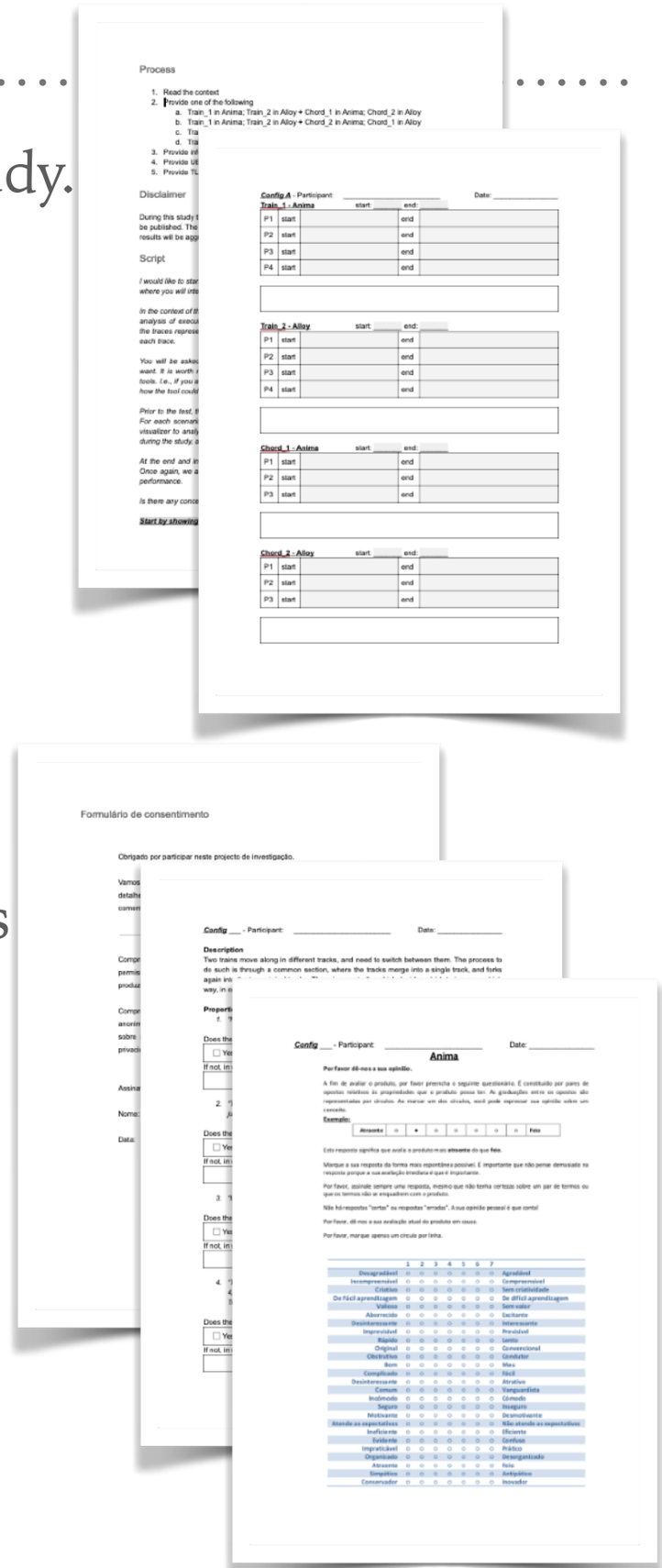
- How to gather participants?
  - Consider the study context - we asked Interactive Systems students
- Is the study adequate for the target audience?
  - Anima is used for analysis in formal methods
- Is the time adequate?
  - If the study took to long, the participants are less willing to participate
- What if the users fail to perform the tasks?
  - We decided to guide participants in case of difficulties, taking care to avoid interfering with the results

STUDY DESIGN – 3 PREDICT POSSIBLE PROBLEMS

- What are the learning effects?
  - We have created several combinations for the participants to start with different scenarios/tools
- What are the **threats to validity**?
  - Is the user focused? Is the user cooperating? (*Too much?!*) Some are out of our control
- How privacy problems can affect the results?
  - We clearly explained the context of the study, while respecting the users' privacy

STUDY DESIGN – 3 PREPARE THE MATERIAL

- There is a set of documents required to perform the study.
- We must record consents, loggings, and results
- We should also provide visual aides to guide the participants
- In this study, we have prepared the following material:
  - For the observer:
    - A script to guide the study
    - A logger sheet - to record time and other outputs
  - For each user:
    - A privacy consent - to be signed
    - A form - for the users to provide answers
    - The UEQ



STUDY DESIGN – 4 PERFORMING PILOT STUDIES

- Performing a study right away is dangerous:
  - What if it fails to fit in the predicted time?
  - What if there are errors in the scripts?
  - What if we can clearly see that our tool does not improve over Alloy?

STUDY DESIGN – 4 PERFORMING PILOT STUDIES

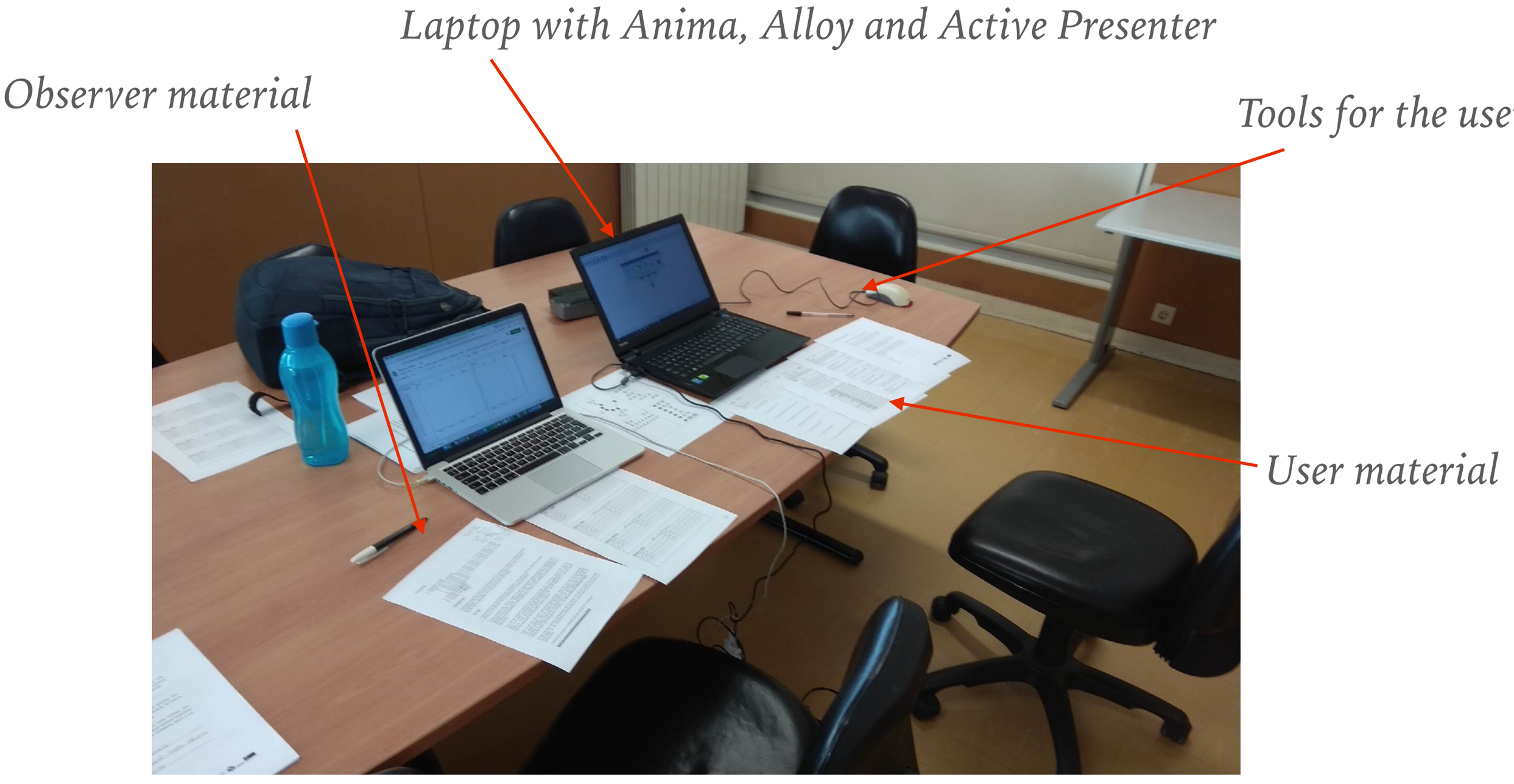
- Some studies are not repeatable - If we fail to perform them the first time, they cannot be repeated in the same context:
  - Users will be less willing to repeat the same study.
  - Users will learn about the tools, and know what to expect the next time.
    - A large amount of time is required to perform the studies!
- Pilot studies are a good approach to mitigate these problems
  - Measure the real expected time to perform the tasks
  - Identify possible errors and unexpected outcomes
  - Gather feedback (can even be considered in the final data)

ELABORATION

- Performing the studies is a laborious task, which starts with the preparation
  - All material should be printed and organised beforehand - all the time that can be saved in the study should be saved
  - All the required material should be ready, namely the laptop, pen, mouse, software, room, etc. Even missing a mouse could seriously compromise a study, as the user might not be comfortable with the touchpad.
- A flexible schedule is required, as the users preferences should be the priority - users are more willing to participate.

ELABORATION

- The setup should be established prior to the test



ELABORATION

- Three main steps occur in the elaboration:

1. The context and details should be explained to the participant
2. The participant performs the tasks, during which the observer will be taking notes
3. The observer ends the study with a debriefing

ANALYSIS

- After the first study, the results should be immediately analysed if possible
  - The observer will remember easily the experiment
    - E.g. emotions, expressions
  - Sometimes it is hard to take note of all feedback, but can be remembered afterwards
  - If a mistake occurs, it is easier to fix (e.g. wrong annotation)

ANALYSIS

- A detailed analysis process follows the study
- All the information should be represented in a proper way (mainly, spreadsheets)

	Version	Day	Hour	Age	Gender	Experience FM	Yrs Alloy	Most Attractive	Most Relevant	Which would use	Recommend
S01	A	T	12:00	21	M	0	0	Anima	Anima	Anima	yes
S02	B	T	13:30	21	F	0	0	Anima	Anima	Anima	yes
S03	C	T	15:00	22	M	0	0	Anima	Anima	Anima	yes
S04	D	W	09:00	22	M	0	0	Anima	Anima	Anima	yes
S05	A	W	09:30	21	M	0	0	Anima	same	Anima	yes
S06	B	W	10:00	21	F	0	0	Anima	Anima	Anima	yes
S07	C	W	10:30	27	M	0	0	Anima	Anima	Anima	yes
S08	C	W	11:00	22	F	0	0	Anima	Anima	Anima	yes
S09	D	W	12:00	22	F	0	0	Anima	Anima	Anima	yes
S10	A	W	13:30	22	M	0	0	Anima	depends	depends	yes
S11	B	W	14:00	21	M	0	0	Anima	Anima	Anima	yes
S12	D	W		21	M	0	0	depends	Alloy	depends	yes
S13	A	W		21	F	0	0	Anima	Anima	Anima	yes
S14	B	W		22	F	0	0	depends	Alloy	Alloy	no
S15	C	W		21	M	0	0	Anima	Anima	Anima	yes
S16	D	W		21	M	0	0	Anima	Alloy	Anima	yes
S17	A	W		21	F	0	0	Anima	Alloy	Anima	yes
S18	B	T	11:00	33	M	0	0	Anima	Anima	Anima	yes
S19	C	W	14:00	21	M	0	0	Anima	Anima	Anima	yes

ANALYSIS

- A detailed analysis process follows the study
- All the information should be represented in a proper way (mainly, spreadsheets)

	Version	Day	Hour	Age	Gender	Experience FM	Yrs Alloy	Most Attractive	Most Relevant	Which would use	Recommend
S01											
S02											
S03											
S04											
S05											
S06											
S07											
S08											
S09											
S10											
S11											
S12											
S13											
S14											
S15											
S16											
S17											
S18											
S19											

	Anima					Alloy					Anima				Alloy			
	Task 01					Task 01					Task 02				Task 02			
	Total	Total	Total	Total		Total		Total	Total		Total	Total	Total		Total	Total	Total	
	P1	P2	P3	P4	SUM	P1	P2	P3	P4	SUM	P1	P2	P3	SUM	P1	P2	P3	SUM
S01	22	16	7	10	**55**	47	19	22	21	**109**	49	10	8	**67**	139	16	10	**165**
S02	30	28	34	86	**178**	54	29	53	72	**208**	25	11	2	**38**	196	28	26	**250**
S03	17	39	28	77	**161**	101	75	95	184	**455**	72	39	9	**120**	206	30	3	**239**
S04	38	17	73	59	**187**	103	52	149	105	**409**	29	15	8	**52**	310	41	11	**362**
S05	65	136	82	33	**316**	41	65	39	108	**253**	72	21	8	**101**	208	20	18	**246**
S06	52	32	50	38	**172**	74	100	64	29	**267**	37	3	2	**42**	170	35	54	**259**
S07	103	10	37	99	**249**	80	30	128	151	**389**	126	38	17	**181**	304	27	14	**345**
S08	48	17	17	8	**90**	63	93	126	31	**313**	60	16	2	**78**	63	12	15	**90**
S09	14	28	22	32	**96**	34	39	65	63	**201**	40	10	4	**54**	48	53	13	**114**
S10	53	75	63	62	**253**	21	24	8	27	**80**	78	10	7	**95**	20	17	20	**57**
S11	32	21	42	93	**188**	17	44	35	33	**129**	65	24	2	**91**	118	56	8	**182**
S12	14	32	37	80	**163**	31	33	38	66	**168**	58	9	8	**75**	147	56	7	**210**
S13	34	10	21	11	**76**	37	12	73	24	**146**	50	29	4	**83**	182	20	8	**210**
S14	22	13	29	45	**109**	55	21	43	51	**170**	52	5	5	**62**	33	26	8	**67**
S15	24	42	75	42	**183**	54	62	34	22	**172**	27	46	4	**77**	81	22	14	**117**
S16	14	21	30	7	**72**	71	24	98	17	**210**	37	6	4	**47**	73	28	7	**108**
S17	49	39	19	16	**123**	22	19	25	17	**83**	26	7	4	**37**	57	10	11	**78**

ANALYSIS

- A detailed analysis process follows the study
- All the information should be represented in a proper way (mainly, spreadsheets)

	Version	Day	Hour	Age	Gender	Experience FM	Yrs Alloy	Most Attractive	Most Relevant	Which would use	Recommend
S01											
S02											
S03											
S04											
S05											
S06											
S07											
S08											
S09											
S10											
S11											
S12											
S13											
S14											
S15											
S16											
S17											
S18											
S19											

	Anima		Alloy		Anima		Alloy	
	Task 01		Task 01		Task 02		Task 02	

	Scenario 1								Scenario 2				
	Anima				Alloy				Anima				
	P1	P2	P3	P4	P1	P2	P3	P4	P1	P2	P3	P1	P2
S01	0	0	0	1	0	0	0	1	0	0	0	0	
S02	0	0	0	0	0	1	1	0	1	0	0	0	
S03	0	0	0	0	0	0	0	1	0	0	0	0	
S04	0	0	0	0	0	0	0	0	0	0	0	1	
S05	0	0	0	0	0	0	0	1	0	0	0	0	
S06	0	0	0	1	0	0	0	1	0	0	0	0	
S07	0	0	0	0	0	0	0	1	0	0	0	0	
S08	0	0	0	0	0	0	0	1	1	0	0	0	
S09	0	0	0	0	0	0	0	1	0	0	0	0	
S10	0	0	1	0	0	0	0	1	0	0	0	1	
S11	0	0	0	1	0	0	0	1	0	0	0	1	
S12	0	0	0	0	1	0	0	1	0	0	0	0	
S13	0	0	0	0	0	0	0	0	0	0	0	0	
S14	0	0	0	1	0	1	0	0	0	0	0	0	
S15	0	0	0	1	0	0	0	0	0	0	0	0	

ANALYSIS

- A detailed analysis process follows the study
- All the information should be represented in a proper way (mainly, spreadsheets)

	Version	Day	Hour	Age	Gender	Experience FM	Yrs Alloy	Most Attractive	Most Relevant	Which would use	Recommend
S01											
S02											
S03											
S04											
S05											
S06											
S07											
S08											
S09											
S10											
S11											
S12											
S13											
S14											
S15											
S16											
S17											
S18											
S19											

	Anima	Alloy		Anima	Alloy
	Task 01	Task 01		Task 02	Task 02

	Scenario 1		Scenario 2	
	Anima	Alloy	Anima	

Subject_ID	First_Name	Age	Gender	MD_Rating	PD_Rating	TD_Rating	Performance_R	Effort_Rating	Frustration_Rating		MD_W
s01	s01	21	Male	0	0	15	70	35	20		0.3333
s02	1	21	Female	30	0	10	15	20	0		0.2666
s03	1	22	Male	11	0	15	3	22	0		0.2666
s04	1	22	Male	40	40	22	6	30	20		0.3333
S05	1	21	Male	15	0	0	20	52	0		0.2666
s06	1	21	Female	40	0	0	10	11	5		0.2666
s07	1	27	Male	20	20	30	20	25	10		0.1333
s08	1	22	Female	30	3	6	90	30	0		0.2
s09	1	22	Female	30	40	11	21	26	0		0.2
s10	1	22	Male	30	5	11	10	25	6		0.1333
s11	1	21	Male	11	15	70	11	14	11		0.2
s12	1	21	Male	40	30	40	30	50	35		0.3333
s13	1	21	Female	65	5	50	74	65	66		0.3333
s14	1	22	Female	27	10	30	19	40	0		0.2
s15	1	21	Male	20	0	20	15	50	15		0.3333
s16	1	21	Male	32	20	23	13	15	5		0.2
s17	1	21	Female	34	11	50	29	36	22		0.2666
s18	1		Male	36	27	29	24	45	32		0.3333
s19	1		Male	45	40	74	10	59	11		0.2
s01	2		Male	0	30	15	60	35	30		0.3333

ANALYSIS

- A detailed analysis process follows the study
- All the information should be represented in a proper way (mainly, spreadsheets)

	Version	Day	Hour	Age	Gender	Experience FM	Yrs Alloy	Most Attractive	Most Relevant	Which would use	Recommend
S01											
S02											
S03											
S04											
S05											
S06											
S07											
S08											
S09											
S10											
S11											
S12											
S13											
S14											
S15											
S16											
S17											
S18											
S19											

	Anima	Alloy	Anima	Alloy
	Task 01	Task 01	Task 02	Task 02

	Scenario 1		Scenario 2	
	Anima	Alloy	Anima	

	Subject_ID	First_Name	Age	Gender	MD_Rating	PD_Rating	TD_Rating	Performance_R	Effort_Rating	Frustration_Rating	MD_W
S01	s01	s01	21	Male	0	0	15	70	35	20	0.3333
S02	s02										0.2666
S03	s03										0.2666
S04	s04										0.3333
S05	s05										0.2666
S06	s06										0.2666
S07	s07										0.1333
S08	s08										0.2
S09	s09										0.2
S10	s10										0.1333
S11	s11										0.2
S12	s12										0.3333
S13	s13										0.3333
S14	s14										0.2
S15	s15										0.3333
S16	s16										0.2
S17	s17										0.2666
S18	s18										0.3333
S19	s19										0.2

Anima

Items																								
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
6	5	2	4	4	4	4	4	3	3	4	3	5	5	4	5	4	3	4	4	3	5	2	2	4	3
6	6	4	2	4	4	5	4	2	4	6	2	5	6	4	5	3	3	2	6	3	6	2	2	2	5
6	6	1	1	2	6	7	6	2	1	6	1	7	6	5	6	1	2	2	7	2	7	1	1	2	6
7	7	2	2	2	6	6	6	5	2	6	2	6	6	7	6	2	2	2	6	2	6	1	1	2	6
6	6	2	2	2	5	6	6	1	3	6	2	6	6	3	7	1	2	1	6	2	7	1	2	2	6
6	5	2	1	2	6	6	7	1	2	7	1	6	7	5	6	2	2	1	7	2	6	2	2	2	6
6	6	5	6	6	7	6	6	2	2	6	2	6	6	6	6	3	3	3	6	2	6	2	2	2	6
7	1	6	7	6	1	1	1	7	6	6	1	6	6	6	6	1	1	1	7	2	7	1	1	1	6
6	6	3	1	2	4	4	6	3	5	5	2	6	5	4	6	1	3	1	6	1	5	2	2	2	4
5	5	5	3	4	5	5	4	3	3	5	3	6	5	3	5	4	4	4	5	3	5	2	3	2	5
7	6	2	3	2	5	5	5	2	2	6	2	6	6	5	5	1	2	2	6	3	6	2	1	2	6
6	5	2	2	3	5	6	4	3	3	6	3	6	6	5	6	2	3	3	6	2	6	3	1	1	6
7	6	1	2	3	5	5	5	3	2	6	2	6	6	6	7	2	2	2	6	2	6	2	2	2	6
5	4	4	2	4	3	4	6	6	3	5	3	7	4	4	6	3	4	2	6	3	6	3	5	4	4

ANALYSIS

- The data should then be processed in order to answer the fundamental question:
  - Is there a difference in performance, while using both tools?
- Depending on the test, different approaches are used
  - Time & errors: mean times are directly compared
  - NASA TLX & UEQ: Use provided tools for calculating the results

ANALYSIS

- But are the differences significative?
  - Even if a tool performs 50% better than the other one, for a given dimension, it doesn't mean that there is a significative difference
- Depending on the size and distribution of the sample, the T-Test, and the Wilcoxon Signed-Rank Test can be used
- Kolmogorov-Smirnov is used to check for data normality

ANALYSIS

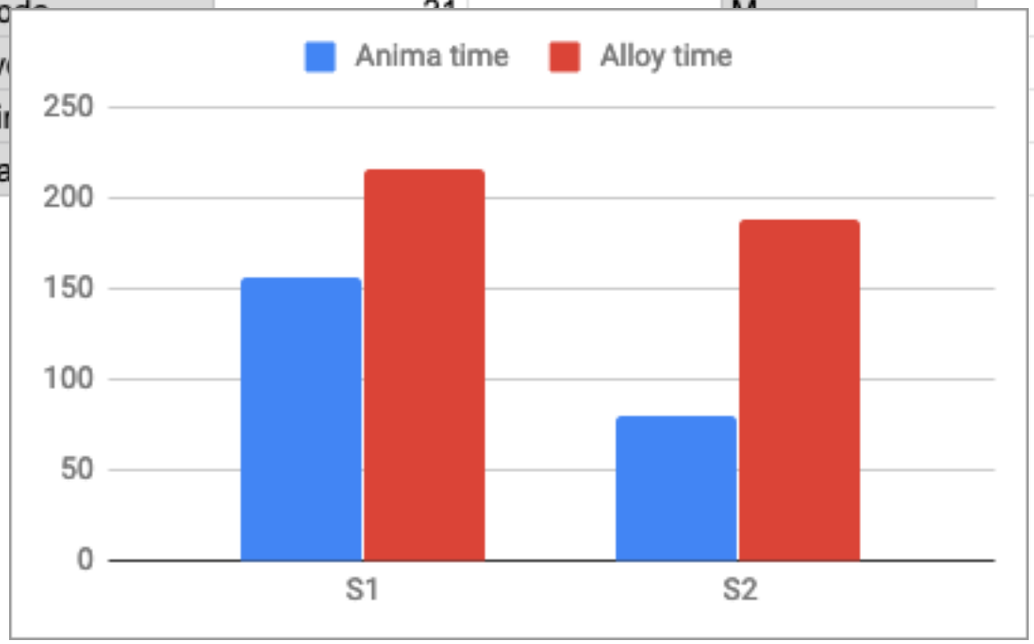
- Finally, the data is represented in a meaningful representation, such as charts

Age			**Gender**			**Experience**	
Mode	21		M	12		FM	0
Average	**22,29665072**		F	7		Alloy	0
Min	21						
Max	33						

ANALYSIS

- Finally, the data is represented in a meaningful representation, such as charts

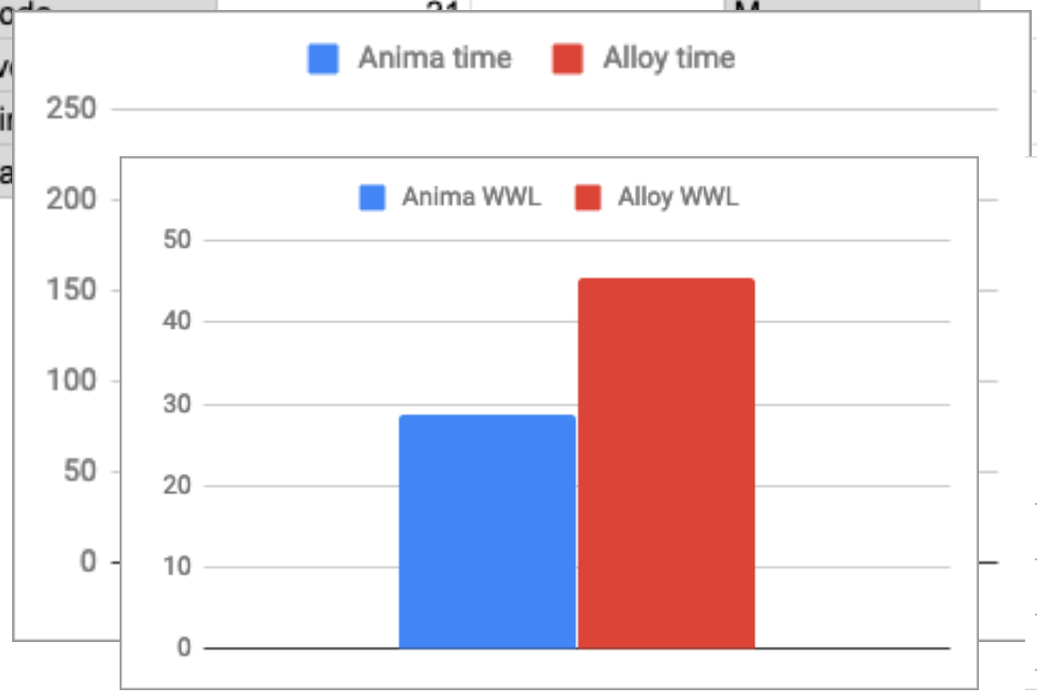
Age		Gender		Experience	
Mo	21	M	13	FM	9
Av					
Mi					
Ma					

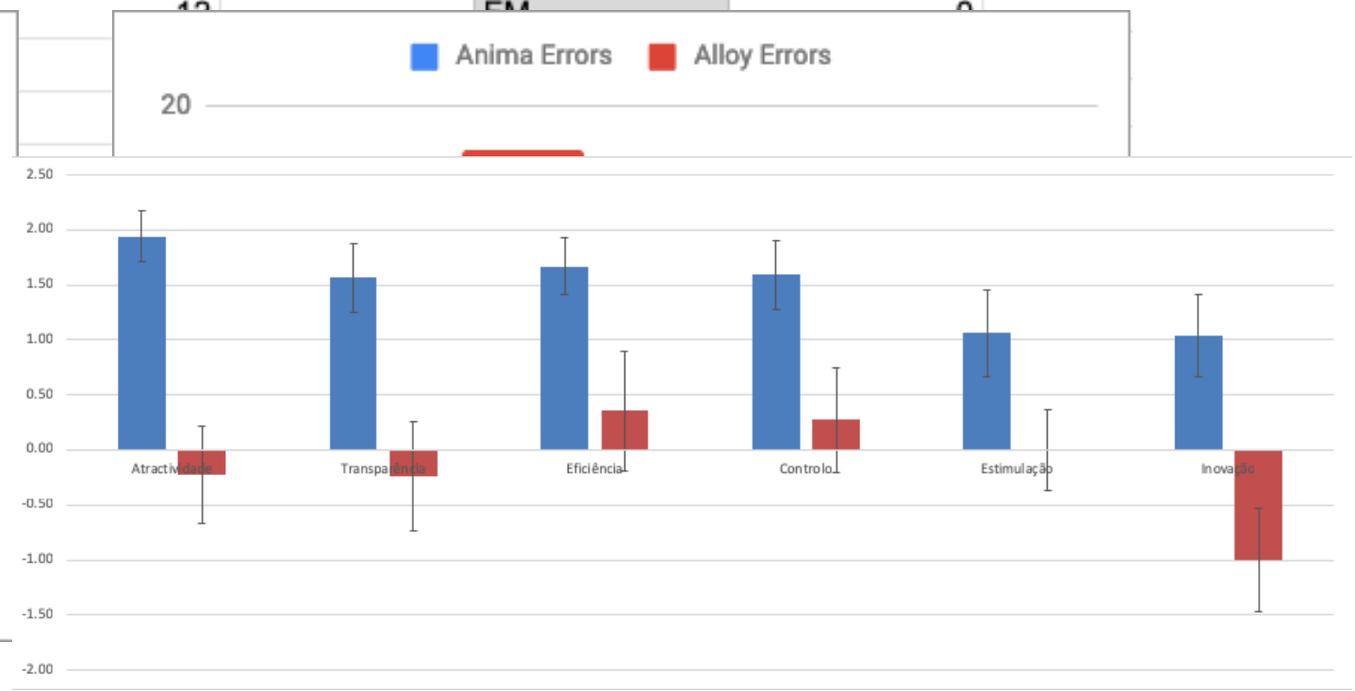


ANALYSIS

- Finally, the data is represented in a meaningful representation, such as charts

Age		Gender		Experience	
Mo	21	M	13	EM	9
Av					
Mi					
Ma					





ANALYSIS

- Finally, the data is represented in a meaningful representation, such as charts

FINAL RESULTS

- Regarding our study, the results have shown that:
  - The users were faster when using Anima
  - The users performed less errors with Anima
  - In some cases, some users did not consider Alloy representations worse
- Animations and consistency between states were pointed by the participants as the biggest improvements
  - The same is true for the images.

EMPIRICAL USER STUDIES

Evaluation of the Anima tool

Rui Couto • José C. Campos