Gamification as a Paradigm for the Evaluation of Visual Analytics Systems

Nafees Ahmed
Stony Brook University
Computer Science
Stony Brook, NY 11790
(+1) 631 403-0880
nuahmed@cs.stonybrook.edu

Klaus Mueller
Stony Brook University
Computer Science
Stony Brook, NY 11790
(+1) 631 632-1524
mueller@cs.stonybrook.edu

ABSTRACT
The widespread web-based connectivity of people all over the world has yielded new opportunities to recruit humans for visual analytics evaluation and for an abundance of other tasks. Known as crowdsourcing, humans typically receive monetary incentives to participate. However, while these payments are small per evaluation, the cost can add up for realistically-sized studies. Furthermore, since the reward is money, the quality of the evaluation can suffer. Our approach uses radically different incentives, namely entertainment, pleasure, and the feeling of success. We propose a theory, methodology and framework that can allow any visual analytics researcher to turn his/her evaluation task into an entertaining online game. First experiences with a prototype have shown that such an approach allows tens-of-thousands of evaluations to be done in a matter of days at no cost which is completely unthinkable with conventional methods.

Categories and Subject Descriptors
H.5.2 [Information Interfaces and Presentation]: User Interfaces – Evaluation/Methodology

General Terms
Measurement, Human Factors, Verification

Keywords
Visual Analytics, Evaluation, User Studies, Gamification

1. INTRODUCTION
Visual analytics (VA) has become increasingly important for its ability to amplify human cognition of complex relationships. By converting information into pictorial representations and allowing humans to interact with these visuals it can overcome the limitations of human mental capacity and so enable human analysts to gain deeper insight into their data faster. Visual analytics has been successfully used in a wide span of domains and contexts, ranging from science to business, economics, medicine, and industry.

The best evaluator of a visual analytics method is the human him/herself, but human evaluators are difficult to recruit and this hampers progress in visual analytics research tremendously. A key observation is that human require incentives to participate in evaluation studies. We propose gamification as a paradigm to overcome this bottleneck. It allows visual analytics researchers to recruit highly motivated human evaluators with ease, at no runtime cost. Gamification engages humans into evaluation tasks by appealing to their intrinsic motivation. This can be implicit in form of design elements invisible to the human, such as the progress bar in LinkedIn, or it can be explicit utilizing applications that are obviously game-like. In explicit gamification humans acknowledge they are playing a game and oftentimes need to opt into playing. Explicit gamification gives rise to purpose-driven games where the purpose is the evaluation task and the game appeals to satisfaction of curiosity or need for entertainment. The intrinsic motivation model of gamification sets it apart from crowdsourcing which targets predominantly extrinsic motivation by providing rewards. While a game can have extrinsic outcomes as well, such as leaderboards and badges, its inherent difference to crowdsourcing is the existence of gameful elements in the solution path.

In this paper, we describe an infrastructure and a set of guidelines that can allow a visual analytics researcher to derive an entertaining game specifically purposed to evaluate, and even optimize, the visual analytics system at hand. Our paper is structured as follows. Section 2 presents related work. Section 3 discusses the elements of gamification. Section 4 describes a prototype we have developed and which has been published in [1]. Section 5 presents our ideas how one could map base visualizations as well as complete visual analytics systems into a gamified evaluation platform. Section 6 closes with conclusions.

2. RELATED WORK
Evaluating a visualization technique has always been considered a challenging task [40][8][28][39] since due to the lack of quantifiable intrinsic quality measures [9], the only acceptable solution towards measuring success of an algorithm is to do a user evaluation. An extensive study by Lam et al. [32] recently revealed, that out of the 850 paper published at the major visualization venues (EuroVis, IVS, IEEE InfoVis, and IEEE VAST) between 2002 and 2012, only 361 of these (42%) reported at least one evaluation. Isenberg et al. [28] did a similar, but slightly enhanced study on papers published in IEEE SciVis between 2006 and 2012, as well as 2003, 2000, and 1997. They found that 76% of the studied 581 papers received some kind of evaluation, but only 15% or 8% of these gauged actual user experience or performance, respectively. The vast majority of evaluations focused on algorithm performance (speed, memory). But in any event, these numbers still say little about the effectiveness of these evaluations. To this end, Ellis and Dix [21] conducted a similar overview, albeit of lesser scope, and found that some of the evaluations were “fishing for results”, arrived at “foregone conclusions”, or were “the wrong sort of experiment”. While these observations might be subjective, a fact is that visual
analytics evaluations typically rarely involve more than a dozen users, which is too low.

2.1 Monetary Incentives

Fortunately, the now widespread web-based connectivity of people all over the world allows for a more scalable human subject recruitment and numerous efforts to engage the wisdom of the crowds into collaborative work have emerged, both for general applications (e.g. [29][3][5][24][10]) and for evaluating visualizations (e.g. [25][30][6][18]). Money is a popular incentive and Amazon Mechanical Turk has become the dominant market place. However, this type of material reward can compromise the evaluator in focusing more on profit and less on experimental accuracy. Therefore, numerous methods have been proposed to motivate “Turkers” [42][44][37], partition their workload [53], assess the outcome of their work [7][19][27][54] and filtering out delinquent workers [22]. Dismissing results, however, wastes money and can also bias the study. Also, even though Turkers are paid only a small reward ($0.02 – $0.04) per HIT (Human Intelligence Task), given a large enough parameter space this can still amount to a considerable sum of money, albeit much less than a lab study. Finally, due to the growing ubiquity of crowdsourcing, less attractive tasks are quickly superseded by more attractive ones and may never get taken, leaving the experiment unanswered. All this has led us to attack the problem from a different angle – gamification.

2.2 Entertainment-Based Incentives

In gamification each problem instance is mapped into an entertaining gaming activity. Players play these games for fun and solve the tasks for free. Gamified systems as defined are often referred in the literature as purpose-driven games and categorized as a sub-field of Human-Based Computation (HBC). As Wikipedia puts it, they are “programs that extract knowledge from people in an entertaining way”. In 2004, Luis von Ahn devised the first purpose-driven game “ESP” [46] which utilized human observation for labelling digital images, showing the power of computing with humans in solving an important problem in computer vision. A series of similar works [34][47][48][49] followed which culminated in a book [33]. HBC started off with purpose-driven games, but with the introduction of micro-task based crowd-sourcing platforms like Amazon Mechanical Turk (AMT), integrating the human processor into the flow of an actual computational process became feasible. Pioneering works are (1) “VizWiz” [5] which gives a near-real time answer to any question related to a picture taken on a cell-phone by immediately creating a task on AMT, (2) “Soylent” [3] which passes computationally hard word-processing functions to AMT workers, and “Foldit” [11] the protein-folding game which showed how even very complex scientific problems can be formulated as a multiplayer online game. Numerous other efforts have also been presented (e.g. [16][26][23][31][37]). In all of these works, the problems considered were generally simple in formulation, and thus the corresponding games had the advantage of finding simple mappings between problem statement and game parameters

3. GAMIFICATION

The advantage of gamification is that it solves the task with considerable amounts of reliability and volume, and at very minimal runtime cost. Players are fully dedicated to do the best job possible. In fact, we found that they often blame themselves and not the visualization algorithms we tested in our prototype (see Section 4) when they fail, although it might have been the latter that misled them. Based on this experience we believe that gamification is an excellent mechanism to evaluate visual analytics systems and their components, and even embed gamification concepts into deployable implementations of these. In the following we shall first present some ground rules of game design and then relate them to visual analytics system evaluation.

3.1 Ground Rules of Game Design

Gamification is the use of game thinking and game mechanics in non-game contexts to engage users in solving problems [wiki]. Here, a key observation to make is that gamers try to win, but game designers try to make gamers play. The key goal is to make gamers like the game so much that they want to play it more. This stands at the heart of intrinsic motivation and has been formally captured in Self-Determination Theory (SDT) [17]. SDT encompasses three basic needs which games must satisfy for them to be interesting and fun and, most importantly for the game designer, get played. These three needs are:

- **Competence**: The player must be challenged to acquire some kind of mastery. Overall indicators of competence are Points (scores) and Leaderboards. Powerful instruments are also Badges since they can be more specific about the type of mastery the player has. Badges can be especially meaningful in online gamer communities where they can serve as virtual status symbols and ‘tribal markers’.

- **Autonomy**: Players must be able to make choices that are meaningful to them, and these choices made must result in immediate feedback, via points and/or via explicit output produced by the game itself. Feedback communicates progress and it can be used to control the user’s priorities. Unexpected informational feedback can increase intrinsic motivation since it provides surprise.

- **Relatedness**: The player must be connected with the subject of the game. Sharing game-based achievements on social networks such as Facebook or in online communities, enhanced by badges, is one instance of relatedness. Another is the actual connectedness with the game’s subject – call it the deep connection – like greener living, etc. For visual analytics systems, this deep connection can be the source of the data, the story and mission behind them, and the analytical findings generated by playing the game (and shared, like in citizen science projects).

3.2 Beyond Points, Badges, and Leaderboards

But there is more than points, badges, and leaderboards (PBL) to a successful gamification. Games have three very relevant elements (in decreasing abstraction order) – dynamics, mechanics, and components – arranged into a pyramid with dynamics on the
top (Fig. 1, adapted from [52]).

**Dynamics:** These are the game’s ‘big picture’ – the constraints, story, and narrative, the player’s emotions (curiosity, competitiveness, frustration, happiness etc.) and progression, and his/her social relationships with other players.

**Mechanics:** Mechanics are ways in which the dynamics can be achieved, such as challenges that require solving, element of chance, actions, levels, aesthetics, rules, skills, competition, cooperation, feedback, resource acquisition, rewards, transactions and turns for multi-player, win objectives, and others.

**Components:** Components are specific forms mechanics and dynamics can take. They include objectives, problems, avatars, mechanisms to unlock content when an objective has been achieved, quests of predefined challenges, levels of player progression, and lastly, PBL.

### 3.3 Specific Elements of Games

Essential to games is that humans enjoy problem solving, being judged (as long as it is considered fair), get a feeling of accomplishment, and receive a surprise due to some uncertainty in the game. We need to provide mechanisms for these, and the following elements [43].

**Require and build skill:** Most games appeal to mental skills, because games are interesting when there are interesting decisions to make, which is a mental skill. Mental skills involve memory, observation, and puzzle solving. The skill a VA system requires is also a function of the data.

**Challenge:** Challenge must be continuous. Humans love a challenge, but it must be perceived conceivable. Else frustration sets in. Conversely, if the challenge is too easy, we feel bored. A player’s skills may be gradually improving, which can be coped with by introducing levels. In our VA application, we need to measure player skill to carefully tune the level of the challenges.

**Triangularity:** This is a great way to make a game interesting and exciting. It is about giving a player the choice to play it safe for a low reward, or to take a risk for a big reward. Triangularity must be balanced, that is, the rewards should be commensurate with the risks. This can be a powerful means to test data transformation shortcuts in a visual analytics system.

**Parallelism:** Players can solve small portions in the order of their choice. If one gives two or more parallel challenges at once, the player is much less likely to grow frustrated this way. Also, giving hints extend interest and fight frustration – a visual analytics game could have a hint button.

**Aesthetics:** People love to experience things of great beauty. So it helps if a visual analytics system is aesthetically pleasing. Carefully tuned music can also add to appeal

**Complexity:** There must be a balance between innate (inherent) and emerging complexity (which is more desirable because it creates player engagement). VA systems may start out in a simple configuration that gets more complex as the player explores the space followed by the solution.

**Flow:** A game must stay in the narrow margin of challenge that lies between boredom and frustration. This margin is called the “flow channel” [12]. The channel slopes upward since as the skills of the player improve, the challenges get greater (see Fig. 2). Flow is related to many of the topics mentioned above, especially those of challenge, skills, levels, and complexity.

### 4. DISGUISE – COLOR BLENDING GAME

We now present a prototype game, called Disguise [1], which we used to evaluate a set of base visualization techniques – color blending. While not explicitly designed as a gamification platform, it made use of many of its elements to entice players to help us sample the large parameter space of color combinations.

#### 4.1 Description of Disguise

Disguise was designed to evaluate four competing color blending algorithms for their ability to communicate depth orderings of two semi-transparent surfaces. In fact, of all the algorithms we tested [14][41] only one of them [50] had originally provided a detailed user evaluation. Our game had an underlying science fiction theme and a typical scenario is shown in Fig. 3. The small disks were moving across the screen, blending with the static larger disks. When meeting a large disk, a small disk could either move below it or above. Players were told they could only click (and destroy) a small disk if it was traveling on top of a large disk. If they clicked it otherwise, they would get a penalty. A given small disk would use a randomly selected blending algorithm. Therefore a less favorable blending algorithm would confuse the player – especially when the color and transparency pairing was challenging – and the player would be more likely to misjudge the depth layering and click the disk if he shouldn’t and vice versa. When designing the game, we examined the typical testing protocol – two orientation-free disks with no decoration – and incorporated this protocol into the game. This ensured that all other factors, such as object orientation, texture, etc., were controlled. Although the fast pace of the game was already quite engaging we added the following further enticements: (1) levels to keep attention high for better players and so keep them in the game, (2) additional interesting graphics rendering effects such as wobbling the large disks as a feedback mechanisms, and (3) a dramatic sound track. To enter a score board, users could login
with their Facebook account, but they could also enter anonymously – 76% of players did this.

4.2 Use of Levels
As discussed, levels are important in gamification. The Disguise game made extensive use of levels, mapping them to exceedingly difficult-to-distinguish color configurations. In these leveling, it also made use of other common measures of visual complexity, such as contrast, clutter, and obfuscation, to some extent:

**Contrast:** higher levels provide less contrast in Bertin’s [4] visual variables (size, color, shape, orientation, etc.)

**Clutter:** higher levels increase the number and density of visual primitives, and the number of visual variables

**Obfuscation:** higher levels increase interaction effects, masking, aliasing, inaccuracies.

4.3 Sampling the Parameter Space
In visualization, a computer algorithm A transforms the data into a picture which is then interpreted by the human by a cognitive response C. We claim that the ability of A to elicit a desirable C can be evaluated in an independent game, and that this setup can then be used within an iterative process for optimizing A. Disguise was extremely effective in sampling the large space of color combinations. In Disguise, each disk has a 4-D color vector (RGB and weight Alpha W) and so the blending of two disks results in an 8-D parameter space. A then transforms this 8-D space into a 3-D space of observable colors (RGB). How well this transformation works is subject to C. In Disguise, sampling the parameter space was trivial since all parameters were on a continuous and bounded scale. In the general case, however, one might randomize a large number of instances and rate them using one or more of the complexity measures listed above (and coded into levels), or define new ones. For example, Dunne and Shneiderman [20] propose several readability metrics for graphs, and Dasgupta and Kosara [15] describe quality metrics for parallel coordinate displays. These metrics can then be used to span the parameter space. Later, once the game is run, one might find that some of these metrics do not influence readability at all or only to a small extent. This, in fact, is one of the strengths of our approach – it gauges the human response directly with no need for indirect heuristic metrics. Hence, a system likes ours will be able to verify the various quality metrics that have been proposed in the literature.

4.4 Results Obtained with Disguise
The Disguise game was a sounding success. Already within 15 days of its opening we had 261 players playing the game, generating close to 30,000 data points – an order of magnitude more than with the conventional study in [50]. The average player played the game for 298 seconds producing 73 data points and 67.8% of the registered players returned, clearly indicating its attraction. During gameplay, on average a player produced 14.6 data points per minute. To give an idea how significant this is, just 1,000 players playing the game for 24 hours will already produce the massive count of 21 million data points. This tremendous number would allow large high-dimensional parameter spaces to be sampled at sufficient density and so capture any non-linear behavior in a function well. Further, the data points are also obtained entirely for free while with Mechanical Turk – even at the minimal possible payment of $0.01 per evaluation – logging one million data points will require $10,000. On top of being a cost efficient, our game also proved to be a much better source in terms of data quality. In case of well-designed purpose driven game, the players are driven purely by their sense of enjoyment and challenge. Hence, the collected data, even at large quantity, should show considerably less noise. In case of Disguise, we didn’t have a direct comparison to similar study made through Mechanical Turk to immediately compare its quality. But, comparison of our findings to existing study results on the same algorithm [50], clearly showed that gamification can indeed provide data at the comparable quality of controlled user studies. We found that the analysis of the acquired data yielded conclusive results with regards to effectiveness of the four algorithms tested, and it also produced the new result that edge blurring of a back-layer disk can enhance the perception of it being in the back. The edge blurring effect was inspired by research published in the psycho-physics domain [38]. Fig. 4 shows two of the confidence maps we constructed for each parameter combination. Specifically, we show the confidence maps for the foreground-alpha and background-alpha weight parameters in [1]. The correctness scale is on the right. The larger the circles the more evaluations were done for a given configuration. The two plots reveal that the second algorithm (plot on the right) performs better for a wider range of alpha combinations.

Our first effort also revealed some potential issues regarding this technique. First, designing a game from scratch, by itself is a complex design job. Even though there are no per data point cost involved, but there lies a one-time cost, both in terms of time and money, in design and implementation of the game. Second, the demographic of the online gamers is expected to be highly skewed and the design of the study should always be highly considerate of that. Third, designing a game and making it popular takes a considerable amount of time and effort. Until we can find a general technique of mapping any problem instance into a parameter of an already existing game with a large user base, this technique will only be worth pursuing for a limited set of large scale problems. For detailed discussion about the observed advantages and disadvantages of the technique, refer to [1].

5. MAPPING VA TO GAMIFICATION
Visual analytics combines three key elements – visualization, interaction, and analytics – into a symbiotic triad. In this triad, visualization amplifies human cognition of complex relationships by externalizing data and information into pictorial representations, overcoming the limitations of human working memory. Computational analytics supports the user by transforming, analyzing, and storing data and information. Finally, interaction allows users to steer and control the computational

Fig. 4: Confidence maps for the foreground-alpha and background-alpha weight parameters in [1].
analytics suite using the visualizations as feedback media. Combining these three elements yields a powerful synergistic system that nicely appeals to human creativity for deriving insight from massive, dynamic, and often conflicting data.

Using gamification principles to evaluate, and even augment, visual analytics systems and their components is a promising approach since visual analytics, just like games, appeals to human creativity and curiosity, requires human pattern recognition skills, uses imagery to communicate, and provides interactive tools to control and steer the underlying mechanisms. Furthermore, it can be collaborative (multi-player) or stand-alone (single-player). Yet, gaining an actual understanding of how this can be achieved, both in theory and in practice, is not straightforward, and deriving a scientific solution for this challenge is at the heart of our ongoing work. In the following sections we present a set of strategies informed by game design by which this can be accomplished.

5.1 Mapping Visualizations to Games

When designing a game it is helpful to first identify a visualization method’s innate tasks and then design the game such that the method’s support for these visualization tasks can be evaluated. We can identify these basic visualization tasks via a suitable taxonomy [51][55][2][45].

- **Locate:** the user points at a known item or describes it
- **Identify:** the user points or describes the item but without having known it previously
- **Distinguish:** the user is able to distinguish different objects as distinct visual items
- **Categorize:** the user is able to recognize items of different categories
- **Cluster:** the user can recognize system-identified categories of items by their links or groupings
- **Distribution:** the user points out categories and items belonging to them are distributed to them
- **Rank:** the user indicates some order of the items displayed
- **Compare:** the user is asked to compare entities based on their attributes
- **Associate:** the user can establish relations between displayed items
- **Correlate:** the user can observe shared attributes between items
- **Remember, recall:** although not part of the original taxonomy, persistence is an important goal

In the specific case of color blending algorithms the most dominant tasks were to enable the viewer to identify, locate, distinguish, categorize, rank, compare, and associate the two (or more) constituents participating in the blending. Our game Disguise tested these basic tasks.

What is also to decide is the game genre. While there are a variety of game genres, “Action” is the best fit for visualization evaluations. Action games require players to use observation, quick reflexes, accuracy and timing to overcome obstacles. These actions deeply involve human cognition of gameplay elements and so align well with the basic visualization tasks as listed above. Disguise is an action game – it requires quick responses from the player. It also has a rather dramatic soundtrack that adds to the immersion. On the other hand, the “Strategy” genre will work well for general VA systems (see Section 5.2).

Something to look out for is that the visuals used in the game are not overloaded. Visualizations encode information using one or more of Bertin’s visual variables. Likewise, games also use visuals to provide feedback, but their design mostly targets entertainment and aesthetics. We must make sure that the visuals used in the game do not conflict with visuals of the visualization. Modifications of the other visual variables are welcome if they make the game more entertaining, but only if they do not create biases in the observations of the tested visual variables. In Disguise, the tested visual variables were ‘Color’ and ‘Value.’ We kept these two intact along with ‘Shape’ – as advocated by prior studies in psycho-physics – and only experimented with ‘Position’ and ‘Size’ to create a challenging action game. Thus, the first step in game design would identify the visual variables that are allowed to change and those that are not.

5.2 Mapping Visualizations – An Example

Disguise was a game that did not directly operate in the information visualization domain, although color blending is an important operation, as shown in Fig. 5 using a layered illustrative parallel coordinate plot as an example.

As a task more directly related to information visualization, let’s suppose we wish to evaluate how well users can detect outliers in a scatterplot. A conventional user study would present the test subject with a random scatterplot and ask to click on the (suspected) outlier. Success is measured by accuracy and timing of the subject’s response. This also represents an ID task, which readily maps to a simple action game scenario. We would first design the visual elements. The components in this visualization are points and the visual variable that encodes information is ‘Position’. To improve aesthetic, we remove redundant visual components (e.g. axes, labels), modify other non-related visual variables (e.g., replace points with some other shape, but uniformly so) and associate a story to motivate actions (e.g., find the spy hidden in your base or locate the socially awkward person in the party). Next we define control and response. ID tasks require mouse motion and clicks from the user for selecting the target. On successful ID, the game rewards points, else zero points. To ensure user participation, we introduce time-out for each ID task. Timing out causes negative rewards and soon results in closure of the game. This point system ensures correctness of the evaluation. Finally, to introduce challenge into the game, we start with a random distribution of the visual elements (i.e., the points) and let individual elements move and settle to their destined location. With this modification, we get a twofold advantage: (1) we receive another metric for evaluation – the time required for ID and (2) the motion of the elements can be utilized for story telling (e.g. people moving around in the party scenario).
5.3 Mapping Visual Analytics to Games

Examples provided in earlier sections give an idea of how useful gamification can be in evaluation. But, it doesn’t provide a clear picture about how one can use similar methods in their own research. Generalization of gamification process for evaluation of visual analytics system would be a very desirable, but the path towards this goal is still hazy and complex. At its very core, the process of designing a game effectively is making a series of meaningful choices. From choice of elements, to their attributes, mechanics etc, each of the contributing component has infinitely many choices. Only some of those choices can have the desired impact on the targeted audience. Because of such immense design space, the task of designing a successful game from scratch is still considered to be an art, rather than definite science. Still, over the years, the study of how games have been made and how players have reacted to different commonly practiced choices in design have given us some understanding of what should be considered good practices. We intend to expand such studies to the field of visualization to better understand how gamification can be handled in this domain. A good starting point should be a simplification of the design space so that a researcher can find his way to a suitable design without having to go through such large set of possibilities.

To get an appreciation of the possibilities, it is useful to consider a simple classification of game designs in game design space $U$ (Fig. 6a), which contains all possible combinations of elements and their values. Let us define two subsets of $U$. First subset, $G$, contains all game designs that results in adequate level of fun and engagement. The second subset, $R$, contains all game designs that conform to a particular set of rules, restrictions or requirement, possibly representing our goal task. The designer of a game, generally is handed a particular requirement for a design. The requirement can come in the form of “Design a game about playing tennis” or “A game that teaches its players about personal finance best practices”. Requirements can be the theme, the story, the desired experience, the audience, targeted platform or in our case, a specific non-game purpose such as evaluation of visualization. Such restrictions reduces the set of possible game designs from $U$ to $R$. Not all of the game designs that fall into $R$ can produce the desired impact on players. The designer, in particular is interested in finding the game designs that both conforms to the requirements and are also enjoyable to play. That is, the designer’s job is to find that elusive design that falls into the set $G \cap R$. The set $G$ depends on the target audience. The set $R$ depends on the requirements set. We can think of a simple classification of game design processes (and also the games resulting from it) from the perspective of the restriction set $R$.

With such classification, we can map the design of each game onto a continuous 1-D space, bounded by two extreme cases of requirements – one being most restrictive and the other most flexible (Fig. 6b). Fig. 6c shows this landscape with various restriction regions. The most flexible, $R1$ may represent a game design task where the designer has only mild restrictions, such as the target platform, but is otherwise free to choose any story, visuals, sounds, mechanics, parameters as long as the game is fun to play. “Angry Birds” for portable touch devices falls into this region. Design $R3$ may represent design of a game where the purpose is to engage users in an otherwise scientific experiment, but the designer is free to choose his own visual elements, story, theme, interactions etc. Purpose driven games such as “fold.it” falls in such categories. On the very extreme is $R4$ – the region of VA systems – where the game design is strongly restricted by the VA system’s private set of visual encodings and interactions. However, the designer still has the freedom to choose other orthogonal mechanics such as points, levels, achievements, time restrictions, story/themes, audio feedback that can provide gamification without conflicting with the purpose of the evaluation. Such strong restrictions has two fold impact on the final game design, it eliminates many good tested choices of mechanics that makes the commercial games or even some purpose driven games successful. But, at the same time, it also greatly reduces the design space to a much smaller, manageable set, leaving some handful decisions to make before a good game can be designed.

Section 3 discussed a set of popular game mechanics, which can be translated into VA systems to make them more engaging. An essential mechanic is level which we can support by using data that are progressively more complex. The mechanic flow and surprise can be introduced for example, by injecting further data that create additional complexities. Another essential mechanic is parallelism where in addition to increasing the levels we can also change the flavor of the challenge for a while to keep players interested about the upcoming levels. On the other hand, the mechanic parallelism can be realized by providing several paths of progress in the game, such as asking players to temporarily change the focus in the analytics task, like switching from looking for outliers to removing noise.

The triangularity mechanic has some interesting applications as it can be used to test specific VA system features. Players might use an optimize button to quickly arrive at a solution, but they

---

Fig. 6: (a) The design space of game designs. (b) $Rx$ in terms of restriction level. (b) Overlap of $Rx$ with fun games $G$. 
would earn more points by manual search. In fact they might even gain additional rewards by finding better or even novel insight which the optimizer could not catch. Additionally, the optimizer button could also be used for other mechanics: (1) as a hint button for frustrated players to keep them in the game, and (2) as an unlock reward, allowing players who have collected enough experience points to unlock secret configurations of the data. These ‘secret’ configurations might be high-quality starter configurations far away which are obtained by optimizing at a wide range.

Finally, the story of the game is also important. It can be associated with the data themselves or it can be fictional. If we gave data – the game’s actors – that are just simple points in space, we have much freedom to attach semantics to them. We can even think of replacing the points by dense pictorial representations fitting the theme of the game, to enhance its aesthetics and make it more engaging. Many fictional stories are possible. For example, one may have a secret agent theme in which the player must catch an evil terrorist (an outlier) and round up like-minded targets (the clusters) for some interrogation. Studying the space of commercial games to identify good themes for VA system gamification will be helpful.

5.4 Evaluation of the Game
Gamification creates an additional layer in the VA system’s design process and hence it should be made part of the nested mode validation process [39]. The following two perspectives apply:

Verification of the design: The verification process asks if the game is capable to yield the desired VA system evaluation. Here, the game mechanics solicit certain user responses which can be manifold – mouse clicks, moves, selections, delineations, and many more. Their relevance will depend on the evaluated task. We note that this will not mean that the players are able to solve the objective of the game (i.e. the visual analytics task) – reaching these objectives would be a verification of the visual analytics system itself since it provides the tools for this task. Rather, we are verifying the game itself. Of course, the game will not perform well when the objectives cannot be met, which we discuss next.

Performance of the game: In the space of all games U, many offer a verified design as defined above, but not all of these belong to the space of fun games G (see Fig. 6). A key aspect of successful game design is balancing the game mechanics and parameters for ensuring the flow. We can achieve this by a gradual deployment of the game. In the first phase, the game might only be tested within a small trusted group to see the direct impact of the design choices. Then, in the second phase, the game might be deployed to an internet-based crowd, but still small enough to retain some control. We can collect amp data about playing behavior and responses, enabling solid decisions about both the game’s verification and performance. Finally, once all choices have been solidified, a full evaluation platform can be made available to everyone on the web.

6. Conclusions
We have proposed gamification as a new methodology for recruiting human subjects for the evaluation of visual analytics algorithms. Convincing humans to volunteer for these purposes has always been a significant obstacle, making this phase of the development process a traditional bottleneck and, as a result, slowing down progress in visual analytics research as a whole. Any attempt of automating this process by machine observers is futile since human perception and cognition are far from fully understood – visual analytics is purposed for humans and thus must be tested with such. To overcome this fundamental chasm we have described the mechanisms needed for a gamified evaluation platform to appeal to human motivation – intrinsic motivation such as enjoyment and interest in the mastery of a subject, and extrinsic motivation such as winning, social recognition, and rewards.

In this paper we have distinguished between the gamification of low-level base visualization tasks and visual analytics scenarios. While we have already developed and demonstrated an example for the former, which was highly successful, we are currently working on an implementation of an example for the latter. Initial results are promising.

ACKNOWLEDGMENTS
This research was partially supported by NSF grant 1117132 and the Korean Ministry of Science, ICT and Future Planning Korea under the IT Consilience Creative Program supervised by NIPA. The material is based upon work supported by the Department of Energy under Award Number DEEE0000220. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacture or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES


