ABSTRACT
Automation has enabled design of increasingly complex products, services, and systems. Advanced technology enables designers to automate repetitive tasks in earlier design phases, even high level conceptual ideation. One particularly repetitive task in ideation is to process the large concept sets that can be developed through crowdsourcing. This paper introduces a method for filtering, categorizing, and rating large sets of design concepts. It leverages unsupervised machine learning (ML) trained on open source databases. Input design concepts are written in natural language. The concepts are not pre-tagged, structured or processed in any way which requires human intervention. Nor does the approach require dedicated training on a sample set of designs. Concepts are assessed at the sentence level via a mixture of named entity tagging (keywords) through contextual sense recognition and topic tagging (sentence topic) through probabilistic mapping to a knowledge graph. The method also includes a filtering strategy, the introduction of two metrics, and a selection strategy for assessing design concepts. The metrics are analogous to the design creativity metrics novelty, and level of detail. To test the method, four ideation cases were studied; over 4,000 concepts were generated and evaluated. Analyses include: asymptotic convergence analysis; a predictive industry case study; and a dominance test between several approaches to selection of high ranking concepts. Notably, in a series of binary comparisons between concepts that were selected from the entire set by a time limited human versus those with the highest ML metric scores, the ML selected concepts were dominant.

INTRODUCTION
Design concept ideation often consists of multiple phases. Studies have identified a correlation between development of a large quantity of ideas and the identification of highly creative ideas [1]. One desirable approach to assist designers, design teams, and organizations is thus to generate many ideas in initial search strategies. When crowdsourcing is deployed to support early stage concept generation, preliminary concepts may number in the thousands. Designers, organizations and design researchers are often faced with a daunting task of categorizing these potential solutions, tagging keywords, and identifying the most innovative concepts to carry forward. To date, research in computational ideation has demonstrated the use of graph topologies to depict high level morphological differences in such idea sets [2], to automatically classify concepts using deep learning [3], or make new recommendations based on proximal solutions to a candidate solution [4,5] in structured networks such as the patent database. Previous theoretical work demonstrated mathematical foundations for calculating creativity scores from structured functional terms [6,7]. The study reported in this paper extends from these foundations and related work by the authors [2,8,9] to report an approach for the automated assessment of design metrics informed by machine learning models. Testing is applied to human generated concepts expressed in natural language. These early stage conceptual categorizations and selections could feed into other generative design tools such as those for surrogate and evolutionary prototyping environments, or even agent based design [10–14].

Design creativity metrics are used to quantify the output of an ideation session, where the design of metrics influence the way decisions are made [15]. However, it takes training and
significant human effort to evaluate these measures. Automated assessment of design metrics such as novelty - a measure of how unique a concept is relative to others [16], level of detail - a measure of how clearly the design is described [17], and feasibility - a measure of performance satisfaction and how easily a concept could be implemented [16], could enable a broader audience to assess design concepts quantitatively. A successful implementation would also provide the design research community with a benchmarking tool for repeatable assessment of design ideation outcome that would allow for the integration of findings across individual studies or institution’s results. The current process is executed through human expertise; therefore, results will vary during implementation and the many variations of extant creativity metrics. This work explores one possible approach to automate the assignment of metrics. Two metrics and a selection strategy are developed. They are analogous to novelty, level of detail, and support search for more feasible concepts, respectively. Preliminary verification tests are applied to evaluate metrics.

![Diagram](image.png)

**Figure 1:** Research approach, the primary objective is to evaluate a candidate method for ML based ideation metrics

### RESEARCH APPROACH

This study presents a method for measuring novelty, level of detail, and feasibility in a large set of design concepts through machine learning (ML) inferred ontology. The research strategy is outlined in Figure 1 (above), and aims to address the following questions:

**Regarding the ontological metrics proposed herein:**
1. What is the agreement between human rated measures of novelty and level of detail scores with the machine determined ontological metrics?
2. How does the ontology-based selection performance compare to time-limited human selection?
3. Can the application of these metrics lead to identifying innovative technologies?

Four ideation conditions were analyzed as the experimental data set for the study. There are at least one thousand crowdsourced concepts in each case (total n = 4000). In order to determine whether this is a sufficiently large number of concepts, asymptotic convergence analysis was applied. The objective is to determine whether the design space has been well explored, using the primary key word tagged in each concept as a “unique identifier.” Next a sample set of 50 ideas was taken randomly from each condition and expert raters evaluated novelty and level of detail according to a rubric. They iterated to reach full interrater agreement. As a first verification, these ratings were compared to the automatically determined scores. As a second verification, a subset of ten ideas is selected in four different ways: (1) a control set from the classical text Homer’s Odyssey [18], (2) a second control, randomly selected ideas, (3) top ten as scored by the ontology or ‘ML’ metrics, and (4) top ten as selected by a human with a fixed time constraint. These are then compared in a series of binary comparison permutations with replicates to determine dominant ideas. After thousands of comparisons, a probabilistic dominance matrix is developed. This second test evaluates whether the metrics can be used to identify dominant concepts. A final test compares results to industrial development.

### DESCRIPTION OF THE METHOD

This section reports key functions of the method used to assess design concepts. First, an input string, i.e., an expressed concept, is tokenized, spellchecking is performed, words are stemmed to the lemma, and part of speech is detected. These are common procedures available in most language processing toolkits.

Several exploratory works on natural language processing effectively stop at this point. Lemma are assigned a cosine similarity value using word2vec or other comparable functions, and then ‘bag of words’ similarities are computed [2,3]. These approaches have demonstrated substantial merit; however, they cannot capture concepts which rely on the order, hierarchy, or sense of terms.

In order to classify unique entities, and their contextual dependency at the sentence level, a dependency parse tree is established for the lemma of each term. This parse tree establishes part of speech and, through association of subject and object, enables training of an artificial neural network to probabilistically infer the sense of a word. For instance, dependency parsing enables a trained hierarchy to differentiate between ‘Ford’ the motor car company and ‘ford’ a shallow...
segment of a river through contextual information, e.g. the sentence also contains ‘motor’ in a given clause. These tagging requirements were met by leveraging the Textrazor platform [19–22].

In the recommended approach, three characteristics of a given concept string are determined. These are: entities or one of millions of disambiguated unique keywords; topics or a unique entry in an evolving database of higher level categories, (e.g. the hierarchical topical categories of Wikipedia) and categories which are a standard hierarchical classification which is fixed (e.g. IPTC media codes) [19,23]. Details of the categorization procedure and insight or purpose for performing each of these categorizations is further detailed in Table 1).

In crowdsourced data, some responses are poorly formed or incomprehensible. Yet, these results can sometimes be successfully tagged by one of the three unique neural networks used in tag extraction (see Table 1). In order to reduce noise, concepts which were not successfully tagged at all three levels are removed. Failure to be categorized by one of the networks indicates that the concept may have been incoherent or poorly structured. This concept is adapted from fault tolerant control strategies. It may be referred to as voting based fault tolerance or ‘byzantine’ fault tolerance [24]. Other works have explored the use of multiple artificial neural networks to improve accuracy of predictions [25]. In principle, the theory is that if multiple distinct networks or agents agree on a conclusion (in this case whether to filter the result or leave it ‘untagged’), this conclusion can be considered more reliable. Furthermore, the tagging involves assessment of individual keywords and classification of the entire entry as belonging to a category of knowledge. A contextual demonstration of this approach, using a negative and positive example from the actual data are presented in Table 2. Table 2 also provides examples of the output for entity, topic, and category tagging results.

**Definition of Metrics**
Any unfiltered entry should be coherent. If it was not, it is unlikely a category would be tagged [19,26], and the concept would be filtered. Furthermore, ambiguous concepts are stringently filtered by the Byzantine fault strategy (see Table 2). It is critical that incoherent concepts are filtered. Once the entries are tagged and filtered, ML metric scores are evaluated.

Recent research has identified a feature of novel designs may be a mixed integration of both near and far domain analogies [27] or common and uncommon combinations in historical co-occurrences [8].

In the tagging process, each entity is assigned a relevance score to the central topic. The relevance of each entity to the core topic is reported based on the network distance, or path length [28], between these two topics in the Freebase knowledge graph [29]. Thus, proposed metric to be used as a proxy for novelty, at the concept level is ‘Span’. Span is taken herein to be the sum of distances of each entity in the concept to the central named entity, as measured by minimum spanning path within the

---

**Table 1: Tagging types extracted. Each tag has a unique associated neural network trained for detection of that type of tag. Each tag is identified separately using a dedicated neural network.**

<table>
<thead>
<tr>
<th>Tag</th>
<th>Motivation</th>
<th>Method of Extraction</th>
<th>ML Training Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity</td>
<td>• Identify keywords</td>
<td>• Probability of co-occurrence</td>
<td>• DBpedia, Wikipedia, Newswire (unique ID number)</td>
</tr>
<tr>
<td></td>
<td>• Disambiguate the correct sense</td>
<td>• Determined by Word2vec similarity score and Part of Speech tag, via syntactic dependency P.O.S. tagging trained with recursive neural networks</td>
<td></td>
</tr>
<tr>
<td>Topic</td>
<td>• Identify the overarching sentence topic</td>
<td>• Machine learning is applied to identify patterns between combinations of specific entities and hierarchical topics within the Wikipedia database</td>
<td>• Wikipedia (hierarchical evolving categories)</td>
</tr>
<tr>
<td>Category</td>
<td>• High level categorization of subject</td>
<td>• Machine learning is applied to identify patterns between combinations of specific entities and the IPTC media codes</td>
<td>• Freebase, Wikipedia (Standardized IPTC media codes)</td>
</tr>
</tbody>
</table>

**Table 2: Example-based demonstration of Byzantine fault tolerant filtering strategy**

<table>
<thead>
<tr>
<th>Sample Concept</th>
<th>Tag Type</th>
<th>Result</th>
<th>Filtering Result</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The second new type of teleportation that I can think of is a flying car. It would either have wings or a hover pad on it. It would be able to go to landing zones and be able to go to places faster than what is normally possible.”</td>
<td>Entity (first listed)</td>
<td>Teleportation</td>
<td>Filtered, no Category tag was detected.</td>
<td>It is unclear whether this is a solution about teleportation or flying cars.</td>
</tr>
<tr>
<td>“Since rail is very expensive, I’d like to see more express bus routes, hopefully with their own lanes to keep them flowing faster than traffic and give an incentive to use them. Make the buses as comfortable as possible, again as an incentive for use.”</td>
<td>Topic (first listed)</td>
<td>Aircraft</td>
<td>Passed, all three tags successfully detected.</td>
<td>The solution is about applying buses to reduce the economic burden of increased public transit</td>
</tr>
<tr>
<td></td>
<td>Category (first listed)</td>
<td>(N/A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Entity (first listed)</td>
<td>Bus</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Topic (first listed)</td>
<td>Vehicles</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Category (first listed)</td>
<td>Economy, Business and Finance&gt;Transport</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
knowledge graph. This is consistent with extant theories of novelty in previous work where both total functions and their distinctiveness (or domain distance) are used to evaluate novelty [6,8]. It is calculated as follows:

\[ \text{Span} = \sum_{i=1}^{n} (1 - R_i) \]  

(1)

where \( i \) is an individual entity, \( n \) is the number of entities in the solution, and \( R_i \) is the relevance score for each individual entity. For instance if there were three entities in the concept with relevance score 0.5 and one with relevance score 0.1, the Span score would be: \((1-0.1)*1+(1-0.5)*3\)

Highly detailed design concepts are considered to be easier to interpret [30]. However, it requires more effort to create such designs [31]. In this case, effort of creation is a sunk cost, and the objective is identifying those ideas of the highest detail in the set which was already created.

The proposed proxy metric for level of detail [31] is simply ‘Detail’. Detail is taken herein to be a basic count of named entities within the concept. Note that a key precursor in applicability of this metric is the fact that the concept must be recognizable as pertaining to a distinct high level category due to the fault tolerant filtering strategy (see Table 2). The sum of entities is simply executed as follows:

\[ \text{Detail} = n \]  

(2)

where \( n \) is the number of entities in the solution, where this is a simple entity count.

A design, if useful, should be feasible to implement. Feasibility is often defined as the ability of a design to meet requirements. In some cases, the design requirements may not be strictly embedded in the design problem statement and automated evaluation of such is challenging in early stage conceptual design. However, some solution avenues may be recognizable by the designer as more likely to be feasible.

Leveraging this insight, a strategy is provided to assist designers in rapid down selection. Each design concept is tagged with a Core Entity. The Core Entity represents the central topic of the design concept. The notion is that designer could browse the solution set first using the Core Entity to identify potentially feasible avenues then select a high novelty concept using the metrics above from within the subset of concepts tagged by a given Core Entity. Textrazor includes a dependency parse which establishes a hierarchy of functional terms within a sentence. One of which is tagged as the core or most relevant unique entity. The Core Entity is a string, not a number.

\[ R_{\text{core entity}} = \max (R_i, R_n) \]  

(3)

where \( R_{\text{core entity}} \) is the relevance of the Core Entity. The relevance of the Core Entity is the maximum relevance score of any entity in the solution.

Taken altogether, these two metrics and strategy provide a way of estimating, how unique is an implementation (Span), how well it is described (Detail), and a guide for selecting a desirable solution category (Core Entity). A contextualized example is as follows:

Table 3: Exam Entry and metric calculation

<table>
<thead>
<tr>
<th>Entry</th>
<th>Core Entity</th>
<th>Relevance</th>
<th>Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charging station; 0.3151</td>
<td>&quot;Charging station&quot;</td>
<td>Electric car; 0.1534</td>
<td>0.1534</td>
</tr>
<tr>
<td>Electronic device; 0.09151</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

EXPERIMENTAL ASSESSMENT

In order to assess the proposed metrics, several tests are executed as shown conceptually in Figure 1. Details will of each test are provided below. Testing includes the collection of ideation data from four conditions. Evaluation of coverage or saturation is tested through asymptotic convergence analysis of the results from each condition. Human inter-rater testing of novelty and level of detail are reported. The human raters then reconciled to develop a final score. Inter-rater agreement of the final human scores and ML metrics is then tested as a preliminary evaluation. Then the metrics are used to select high scoring ideas from the entire set. These are rated against other selection procedures which are: ideas selected in a controlled time period by a third human (not one of the raters); random; and control (source text from an antique novel). The second test is to determine which selection approach was dominant. This test would indicate whether the metrics can be ‘forward applied’ and actually serve to identify ideas from a massive concept set. Finally, a comparison is made between the ML based filtering and an industry case. This last activity aims to see if there is any connection between the results and the industrial state-of-the-art. Details of each test are described below.

Problem Statements

Four ideation conditions are employed. These are outlined in Table 4. The objective is to explore differing types of design opportunities-problems, in particular technology push versus opportunity-problem driven, and ideation by designers in a group setting, though acting individually, versus online distributed crowdsourcing.

Table 4: Overview of the four problem conditions

<table>
<thead>
<tr>
<th>#</th>
<th>Opportunity-Problem Statement</th>
<th>Source</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&quot;Design future public transportation in your city.&quot;</td>
<td>In-person Crowdsourcing</td>
<td>Problem Driven</td>
</tr>
<tr>
<td>2</td>
<td>&quot;Imagine a new transportation system in your neighborhood - Describe it!&quot;</td>
<td>Distributed Crowdsourcing</td>
<td>Problem Driven</td>
</tr>
<tr>
<td>3</td>
<td>&quot;Explore possible inventions given a new technology: GPS that is accurate to less than 1cm.&quot;</td>
<td>Distributed Crowdsourcing</td>
<td>Technology Push</td>
</tr>
<tr>
<td>4</td>
<td>&quot;Explore possible inventions given a new technology: Weather prediction that is 100% accurate.&quot;</td>
<td>Distributed Crowdsourcing</td>
<td>Technology Push</td>
</tr>
</tbody>
</table>

Data for the in-person crowdsourcing was collected at a large scale ideation workshop with over 2,000 mid-career professional design, engineering participants, see Figure 2 (case 1 in Table 3).
The remaining data was collected through Amazon’s mechanical Turk with users given the design prompt and an open ended text entry box to submit a single design solution to the prompt (the raw data is hosted and may be used with proper citation https://tinyurl.com/ML-metrics-IDC).

Saturation Evaluation
An asymptotic convergence analysis, using the algorithm outlined by Lim et al. [2], is employed. In this case, once a unique Core Entity is detected, it is not counted again if a new concept utilizes the same Core Entity. This test establishes that the sample set is justifiably complete at a high level, and the designers could be reasonably confident to cease ideation. Otherwise, it is probable that a high scoring concept could be excluded from the analysis or uncaptured. The sampling sequence is arbitrarily randomized, and multiple trials are plotted in order to establish upper and lower bounds of the saturation curve to diffuse any potential false patterns induced by a sampling sequence (see Figure 1). At approximately 1,000 entries each, all four cases were saturated according to the criteria (from [2]) in each of the sampling sequences. Thus 1,000 entries are used in testing.

An asymptotic convergence analysis was performed on the results from cases one and two, from Table 4, in a previous study. In the previous study saturation was performed by human raters [2]. The human results indicated results from case one (See Table 3) as saturating first. In Figure 3, it can clearly be seen that, for Entities, the online results (blue line) saturate first. However, results for the tag Category (see Table 1) followed the same trend, with offline or in person results saturating first. One possible explanation for this is that humans may form bins in a manner more similar to the Category tagging approach. In either case, the Category tag also saturated earlier (approx. n = 100 concepts tagged) which was also consistent with previous publications [2]. However, Entities are used herein due to the granular classification that they afford.

Inter-rater Testing
Inter-rater testing was applied to a sample set of 50 randomly selected concepts from each of the four conditions, as shown in Table 3. Pearson’s inter-rater agreement was used to determine the level of agreement between raters. The process included discussion of a small sample set, review and agreement on the rubric, followed by the actual independent rating. Once rating was complete, the raters discussed any discrepancies between raters to reach full agreement on each single rating. The objective for following this process is to increase the assessment quality. The rubrics for evaluating novelty and level of detail are as follows in Tables 5 and 6 respectively. The inter-rater agreement scores for human-to-human assessment are shown in Table 6.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Solution exists and already serves same in domain purpose, ‘everyday solution’</td>
</tr>
<tr>
<td>2</td>
<td>The technology solution combination exists in domain, but is uncommon not ‘everyday’</td>
</tr>
<tr>
<td>3</td>
<td>Single new feature (new = approach or technology from another domain)</td>
</tr>
<tr>
<td>4</td>
<td>Two new features</td>
</tr>
<tr>
<td>5</td>
<td>Three or more new features together</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Incomprehensible, or no features are well explained</td>
</tr>
<tr>
<td>2</td>
<td>One feature is described</td>
</tr>
<tr>
<td>3</td>
<td>Two features are described</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>n features are described</td>
</tr>
</tbody>
</table>
Table 7: Human to human inter-rater Pearson’s correlation agreement scores

<table>
<thead>
<tr>
<th>Ideation Condition (see Table 3)</th>
<th>(1) Transport - In person</th>
<th>(2) Transport - Distributed</th>
<th>(3) GPS - Distributed</th>
<th>(4) Weather - Distributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novelty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level of Detail</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novelty/SPAN</td>
<td>0.39</td>
<td>0.41</td>
<td>0.36</td>
<td>0.47</td>
</tr>
<tr>
<td>Level of Detail/Detail</td>
<td>0.51</td>
<td>0.55</td>
<td>0.21</td>
<td>0.69</td>
</tr>
</tbody>
</table>

While the agreement between human and ML-based metrics is not as high as human to human scoring, it is important to understand that this is a completely unsupervised approach, trained on generic knowledge graphs and thus broadly applicable. Secondly, the input data is fed as natural language, the interpretation of which can be ambiguous, or result in ‘multiple variant yet correct assessments’ even between experts.

The idea is to develop a system comparable to a human rater, as Pearson correlation in standard in human to human comparisons it is utilized here.

Table 8: Human to ML-based inter-rater Pearson’s correlation agreement scores

<table>
<thead>
<tr>
<th>Ideation Condition (see Table 3)</th>
<th>(1) Transport - In person</th>
<th>(2) Transport - Distributed</th>
<th>(3) GPS - Distributed</th>
<th>(4) Weather - Distributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novelty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level of Detail</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novelty/SPAN</td>
<td>0.91</td>
<td>0.88</td>
<td>0.86</td>
<td>0.73</td>
</tr>
<tr>
<td>Level of Detail/Detail</td>
<td>0.98</td>
<td>0.88</td>
<td>0.74</td>
<td>0.74</td>
</tr>
</tbody>
</table>

**Dominance matrix**

As a performance validation a selection exercise was conducted. The objective was to select ten highly novel, detailed designs to carry forward in design. See Appendix Table A1 for full listing of each concept set. The value ten was selected as it is considered the upper limit of working memory [32] and thus it is a reasonable value to assume would be selected for further evaluation in a design project. These concepts should be both novel and detailed for the given exercise which is to “identify the ten highest impact designs in the original data set”. Concepts were selected using four strategies, a human (non-expert) was given instructions to select the ten most novel and well detailed concepts in 12 minutes. Firstly, the timing was based on a study of workplace behavior that a typical uninterrupted unit of work lasts approximately 12 minutes [33]. Secondly, a non-expert was strategically chosen. An objective of this work is to extend the ability of designers to solve more complex problems and thus navigate large concept sets outside their area of expertise. The second condition involves taking those ten ideas from the entire set which scored the highest value of Span.Detail (concepts that are novel and well described) from the ML based metrics. A third strategy consisted of selecting ten design concepts at random. A fourth set consisted of ten paragraphs from the antique epic poem, Homer’s Odyssey and was meant to serve as a control.

In an online crowdsourced platform, users were presented with two ideas in a random order. Each permutation, or unique pairing of 40 by 40 concepts, was tested with 3 replicates. The average scores are reported in the format of a dominance matrix, typically used to determine rankings or performance of a team or strategy across many individual matches or comparisons [34]. Figure 4 reports results of the binary comparisons. Ontologically selected concepts were selected as more novel and detailed the most often, followed by human, random, and control consecutively. The following differences are significant using a pairwise comparison of transformed defective means for each order in the rank: Ontological Metrics > Human (p = 0.06) nearly significant; Human > Random (p = 0.01), strongly significant; Random > Odyssey (p < 0.001) strongly significant. The full sets of concepts which were compared are provided in the Appendix. Interestingly, some passages of the poem, which is in fact about exploration, do obliquely refer to navigation and were selected in a small percent of trials. An abbreviated are taken from each condition is shown below (See Table 4).

**Table 9: Indicative sample taken from each selection strategy**

<table>
<thead>
<tr>
<th>Source</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML Metrics</td>
<td>&quot;An automated system of … AI &quot;Horses&quot; … a short, small, bullet shaped car like you used to see on the SIDE of a motorcycle, pulling a carriage, with an AI brain controlling it... this would have the novelty and functionality of pulling people when you wanted to, or you could just use the Ai ‘sidewalk’ to run courier errands on the ground for companies that wanted fast delivery. It keeps the battery away from the passengers if there is ever a crash, too. When it's just moving small items around for parcel delivery, it’s not wasting fuel hauling an empty cab ….&quot;</td>
</tr>
<tr>
<td>Human</td>
<td>&quot;I would like to see an advanced transportation system that involved driverless cars. The idea is the same as using a taxi but without a driver. Like Uber, you would use your mobile device to order a car and control the experience. The car shows up at your door, takes you to your destination and you are billed for the ride. It would save me thousands of dollars in the cost to own, maintain, insure and license my own personal car.&quot;</td>
</tr>
<tr>
<td>Random</td>
<td>&quot;Automated driverless cars networked to navigate together and accept passengers via mobile applications&quot;</td>
</tr>
<tr>
<td>Odyssey</td>
<td>&quot;How sad,&quot; exclaimed Telemachus, &quot;that all this was of no avail to save him, nor yet his own iron courage. But now, sir, be pleased to send us all to bed, that we may lie down and enjoy the blessed boon of sleep.&quot;</td>
</tr>
</tbody>
</table>

Figure 4: Dominance matrix results reporting average scores for "Which concept is more novel and well described" for the given problem. Note that the construction is such that scores depict percent of instances that a given row defeated a given column, or in other words the probability of a superior selection being made by that approach. A total of 4,800 comparisons were made.
Feasibility Case Study
The results of the distributed ideation exercise were collected on Thursday January 16th, 2016. As some time has elapsed in developing the toolkit requisite to successfully compete against human selection processes a unique opportunity emerged, which is to explore whether any of the highly ranked concepts have since been deployed.

This is a pilot test of the predictive capability of this overall approach. The highest scoring concept was a social network for ridesharing network based on the rider’s workplace. A pilot of this concept had already been launched by MIT in 2011 [35].

Notably, the second highest scoring concept selected according to, Span*Detail and in the category with the Core Entity ‘bus’ which was the most frequent, was the following:

“...the trolley or bus could be made to target things people want to go do instead of just transportation. Like grocery shopping or night out to watch movie theatre and that way everyone is grouped and the businesses upsales, etc. can develop better because of that.” Crowdsourcing Participant 16-Jan-2016

Figure 5: Public service announcement for the recently deployed On-Demand bus new service experimental pilot in Singapore, Singapore, trial initiated 17-Dec-2018

To the knowledge of the authors there was no on demand bus service, at least in Singapore, at the time initial concepts were collected, in 2016. However, the local transit authority recently announced the deployment of a pilot project to evaluate an on demand public transportation system consisting of an app providing adaptive on demand bus routing which would cater to exactly this concept. The pilot program was announced for deployment on December 17th 18, almost two years after this entry was submitted for ideation [36].

DISCUSSION
This study provides insights from an initial foray into automating the assessment of design concepts written in natural language. The source data were from a diversity of problems. While design concepts were evaluated in this work, the intention is that this general method might serve other aspects of design research as well. This approach might potentially support evaluation of user needs data, design standards manuals, error report logs (e.g. NASA’s lesson learned library [37], or even news articles.

The method could be used in a variety of ways to highlight different types of concepts depending on context. Those taken from the categories with frequently occurring Core Entities also appear to have, for the most part, been very recently implemented. Many of the solutions that scored highest in Span*Detail have not been developed. Therefore, the work may serve as an alternative means for developing predictive technology roadmaps in concert with approaches such as the work for Chris Maegge et al. [38] in patent trend analysis pending further study.

Alternately, for designers who have access to a vast database and want to target review, the entity tags can also function as a simple means to rapidly categorize or search for meaningful solutions. As an illustrative example, a design team working on a carpool app might check the solutions tagged with ‘carpool’. Such as the following examples from the top ten (Span*Detail) ideas tagged ‘carpool’:

“A car-pooling app. Non-drivers and those looking for regular rides who normally carpool would join the community and hook up with drivers who are going to their general destination. There would be a monthly fee, which would lowered and even be removed if someone drives/uses their car enough. For example, it’s $5 per month, and every time YOU drive it goes down by $.50 cents. There’s incentive for both non-drivers (a small fee for rides) and drivers (free membership, maybe some gas reimbursement).”

Or

“A group carpooling smartphone app that allows people in apartment complexes or housing areas to visualize a profile of others in the area that are traveling to a nearby destination. Participants can communicate over the app, and background check and everything is done before being certified to use the service. People can post looking for ride notes, or say that they are traveling to a specific area and looking for passengers to ride in order to split costs.”

Finally, it may be possible to forgo the initial crowdsourcing and achieve fully generative concept selection by identifying sufficiently relevant databases such as by extracting relevant sentences. An example of a relevant database would be the BASE open source research database [39].

In closure, for the problem to ‘cross a river’, current tools may automatically provide augmented boat designs given the starting point of a boat; this work reaches towards solutions to enable the automatic suggestion of high level concept alternatives e.g. ‘build a bridge’ or ‘relocate to a shallow ford before crossing’.

LIMITATIONS
The work has several limitations in scope and application which future works may address through replication, expansion or pivots into related methods. Firstly, the ML creativity proxy metrics were developed through a synthesis of literature review and empirical testing (of several variations). However, extended review and more detailed competitive analysis of differing
CONCLUSIONS

This study presents a method to automatically assess design concepts. A ML package [19] is employed to extract ontological data at the sentence and keyword level, then the results are filtered using a control systems strategy and finally assessed with design metrics. Two metrics, Span, Detail, are proposed that are analogous proxies to the traditional design metrics of Novelty, and Level of Detail. A strategy for selection by Core Entity enables selection by category to support identification of desirable, or feasible clusters.

Data was sourced and evaluated from four different ideation conditions. Analysis of the resulting ideas were evaluated for asymptotic convergence to determine if the design space was well covered. Convergence is based on the core (most relevant) Entity. The convergence procedure is taken Lim et al. [2]. Unique entity tags converged in all four cases converged with less than 1,000 unique entries.

A series of tests was employed to establish performance of the method. First a comparison was made between expert human rater assessment and the ML metrics using Pearson’s inter-rater agreement. While the level of agreement was low compared to human-to-human responses, it is comparable or superior to first round human-to-human agreement when a coding rule-set is not predefined i.e. open ended encoding.

In a second test, ten ideas were selected from the online crowdsourced example problem. The strategies were: human selection, ML metric selection, and control in the form of both random idea selection, as well as text injected from another source (classic novel). Using a series of pairwise comparisons a dominance matrix was established. The dominance matrix shows that the ten ideas selected by the ML metrics (highest ranked concepts) were the most dominant.

Though it is not always the case that a human designer’s time constraint is explicitly time limited in practice, person hours are a well-established measure of project cost [40] and thus it can be argued that this result supports the case for automated concept selection as a means to save cost or to add leverage to human selection.

A third, predictive observational test explores whether any of the ideas initially reported in the concept ideation study have since appeared in the public domain. This serves as a retrospective assessment of this method’s applicability as a technology development roadmap indicator, and ties the outcome to real world developments. In short, concepts that scored well in the metrics have in fact been either recently developed or in one case developed only after the ideation session. Some are yet to be developed.

ACKNOWLEDGMENTS

This work is funded by the SUTD-MIT International Design Centre (IDC) in Singapore (https://idc.sutd.edu.sg). Any opinions, results, or conclusions in the paper are those of the authors and do not necessarily reflect the views of the sponsor.

REFERENCES


