# Coalescing the Vapors of Human Experience into a Viable and Meaningful Comprehension

**Tomer D. Ullman (tomeru@mit.edu)**
Department of Brain and Cognitive Science, MIT

**Max Siegel (maxs@mit.edu)**
Department of Brain and Cognitive Science, MIT

**Joshua B. Tenenbaum (jbt@mit.edu)**
Department of Brain and Cognitive Science, MIT

**Samuel J. Gershman (gershman@fas.harvard.edu)**
Department of Psychology, Harvard University

**Abstract**

Models of concept learning and theory acquisition often invoke a stochastic search process, in which learners generate hypotheses through some structured random process and then evaluate them on some data measuring their quality or value. To be successful within a reasonable time-frame, these models need ways of generating good candidate hypotheses even before the data are considered. Schulz (2012a) has proposed that studying the origins of new ideas in more everyday contexts, such as how we think up new names for things, can provide insight into the cognitive processes that generate good hypotheses for learning. We propose a simple generative model for how people might draw on their experience to propose new names in everyday domains such as pub names or action movies, and show that it captures surprisingly well the names that people actually imagine. We discuss the role for an analogous hypothesis-generation mechanism in enabling and constraining causal theory learning.

**Clerk:** Occupation?
**Comicus:** Stand-up philosopher. I coalesce the vapors of human experience into a viable and meaningful comprehension.
**Clerk:** Oh, a bullshit artist!
(*History of the World, Part 1*. Dir. Mel Brooks)

## Introduction

How do people come up with new concepts, causal models or theories, to make sense of their world? Whether it is scientists trying to explain the natural world with formal theories of physics or chemistry, or children and adults trying to explain their experience with common-sense theories such as folk biology or folk psychology, the question of how fundamentally new ways of thinking unfold over time and change in response to evidence has long been of interest to cognitive science (Carey, 2009; Gopnik & Wellman, 2012; Schulz, 2012b).

This paper is about the more everyday aspects of this process: New thoughts of on-the-fly explanations and causal models to make sense of everyday problems and puzzles. The difference between radical conceptual change and prosaic concept generation is one of degree, like the difference between coming up with a general explanation for how objects balance and fall – a difficult process that may take months or years (Baillargeon, 2008) – and coming up with an off-the-cuff explanation for why soda tastes fizzy, or a plausible name for a new action movie.

While everyday creativity is not categorically different from scientific creativty, the specific challenge of explaining creativity in everyday idea generation has been pointed out by Schulz (Magid, Sheskin, & Schulz, 2015; Schulz, 2012a). Schulz called for both empirical and theoretical research on the issue. On the theoretical side, everyday creative thought poses a problem for computational models that have been developed to capture more radical theory change (Goodman, Ullman, & Tenenbaum, 2011; Ullman, Goodman, & Tenenbaum, 2012). Briefly put, these models see humans as reasoning over a hypothesis space ("theory space") that contains explanations and concepts. People do not know in advance what the right concepts and theories are, so they must stochastically search through these large (potentially infinite) spaces, adjusting and discarding their concepts as they go. If a newly proposed concept or theory better fits the data, and is more likely under a general prior favoring such things as 'simplicity', then that newly proposed concept will be accepted.

Such a process might make sense for theory-change that takes years to unfold, but surely (Schulz argues) it is underconstrained when it comes to everyday thought. The search spaces are too large, and the process does not take into account the specific constraint of a task. For example, suppose a friend asks you to come up with a name for their hip new Thai restaurant, located near a cluster of start-ups. You may never have been faced with such a problem before, but after some thought you might come up with "Thai-Tech". It is not a particularly good name, but it is at least a relevant one. It is better than "Mummified Ragdoll Fifteen" or "Croatian delight", or any of the other infinite combinations of words that language affords. Such proposals would not just be rejected if proposed, they would not be thought of in the first place.

That people do not waste time with nonsense (in the sense of proposing gibberish answers used above) seems almost trivial, and suggesting a concept-production algorithm that runs through all combinations in the English language for any given task seems like a clear non-starter. And yet, many stochastic search algorithms for hypothesis generation suffer from exactly this problem. How can such algorithms be amended to not consider 'obviously wrong' proposals without actually proposing them first?

The challenge of everyday thinking is the challenge of reducing hypothesis spaces quickly and on-the-fly, in response to a task that might never have been considered before. The reduced hypothesis spaces might still be large, and they might still contain 'bad' ideas and concepts. But these concepts and ideas would at least be relevant to the task. They would be capable of being wrong, in the sense that a human would recognize them as a bad or good response to the task, as opposed

to completely unconnected.

In this paper we take up the computational challenge of everyday reasoning, and propose a structured approach for narrowing hypothesis spaces by inferring a generative model over relevant examples taken from memory. The resulting "Bounded-Space" model (or, BS model, for short) can generate reasonable proposals for learning or problem solving within a domain, but it does not exempt a thinker from stochastically searching within the reduced space and evaluating the proposals. Ideas generated in the reduced space may be bad ones. They may even be, for lack of a more polite term, mostly bullshit. But by quickly cutting down the space of possible thoughts from 'mostly nonsense' to, at worst, 'mostly bullshit', everyday thinking and learning can be powerfully and usefully constrained.


Figure 1: A space of possible names for things. Only a small subregion contains names relevant for a given task (dark area, in this case pub names). A stochastic search over the entire space would result in many examples that are not even bad.

We develop and test our BS model using the relatively simple task of coming up with new names for items, specifically movies of different genres and pubs (this is similar to a task proposed in Magid et al., 2015). We compare its proposals to those of people, and to real world data. We find that the basic BS model fares worse than people, but is already within a reasonable range. A cognitively plausible extension to the BS model that evaluates a small number of samples before selecting one is surprisingly consistent with the range of people's novel name generation behavior. The BS model provides an initial proposal for capturing everyday thought, but extending it beyond the (relatively) simple task of proper names to tasks such full-blown explanations is not trivial, a point that we take up in the discussion.

## Modeling on-the-fly thought

Generating new thoughts and concepts can be seen as a stochastic search through a hypothesis space, guided by some metric of success or quality (Ullman et al., 2012). Ideally, the hypothesis space should be large enough to encompass any possibly correct or useful theory or concept. Infinite hypothesis spaces that fit this bill can be easily specified through grammars over conceptual primitives (Piantadosi, Tenenbaum, & Goodman, 2012), and in principle it is possible to stochastically search through such spaces by proposing and rejecting amendments to the current hypothesis. But the problem of everyday thinking suggests that these proposals must be strongly constrained by abstract domain knoweldge, such that most of the proposals that can potentially be considered (a-priori of any data or constraint) will never be considered, and our actual proposals focused efficiently on candidates that have some hope of being useful.

To see the problem more clearly, try to come up with a name for a new pub. Perhaps you never faced such a task, but presumably you can do it with some degree of success. Such a thought process *could* be implemented by a stochastic search through all the possible phrases in the English language, but this would lead to a ridiculous proportion of not only bad proposals, but completely irrelevant proposals. In Figure 1 we consider a fictional space of all possible names for things. Only a tiny portion of that space can even be called 'Pub Space'.

Assuming a thinker (such as yourself) was never asked to come up with a new pub name before, how can they quickly reduce the space of all possible new names to just ones relevant for pubs? Presumably you thought of a better pub name than "Towards a Better Designed Belgian Office", and if someone posed that as a suggestion you could evaluate it as a bad proposal. [1] But "Towards a Better Designed Belgian Office" is a potentially good answer to a different question. Again, the issue is that *before* any particular question, puzzle or problem is posed, we wish to have a large search space that can generate many possible concepts, thoughts and solutions. But *after* a particular task is set, we wish to restrict the space to only the relevant solutions. How can we know ahead of time what counts as a 'relevant' solution, without first proposing it and evaluating it?

We propose a method (Bounded-Space, or BS) for constructing new, relevant concept spaces on-the-fly, illustrated through the pub example in Figure 2. For any particular task the thinker first draws from memory several examples that match the desired concept (Figure 2.1). For example, if asked to come up with a new pub name, the thinker might first draw some known pub names from memory. We assume that relevant examples are available in memory, even if they are relevant in a broad sense. "Broad sense" here means that the examples match the general structure of the task. For example, if the task is coming up with an explanation for why the Roman Empire fell, relevant examples can include causal explanations in general ("State changes in X can be caused by an outside Y"), rather than particular reasons why the Roman Empire fell ("Crises of legitimacy"). Without any relevant examples that come to mind the question itself is a non-starter, e.g. "Can you ganoosh a new Floop?". The initial retrieval of relevant examples might rely on associative memory, but is a problem outside the scope of this paper.

The thinker then uses inverse inference, conditioned on the

[1] Or rather, as "not even bad', just ridiculous.

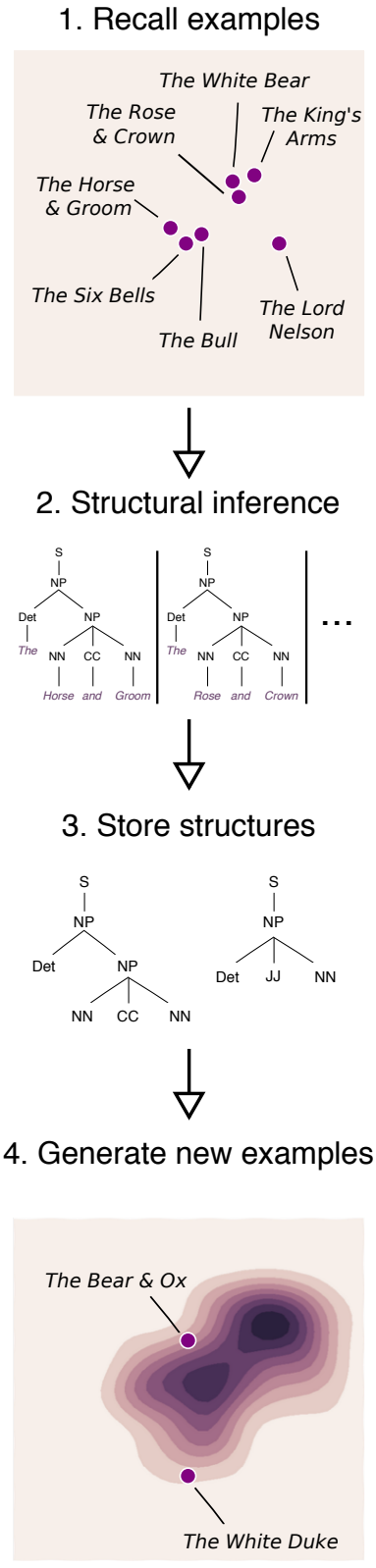


Figure 2: An illustration of the BS model applied to the domain of pub names.

examples, to construct a generative model that can produce these examples. Such a thinker might focus on the form of the examples, noticing for instance that all the examples share a certain abstract causal structure. In our particular case study, which is restricted to new names for pubs and movies, we consider a grammatical analysis of the examples from memory (Figure 2.1), although many other structural constraints are possible in the general case. For example, the thinker would notice that "The Rose and Crown" share the same syntactic structure as "The Horse and Groom", namely "The [Noun] and [Noun]". The thinker then stores the relevant underlying structure without the particular terminals (Figure 2.3). This database of structures is a proxy for a generative model over the space containing the examples.

In order to come up with a new example (Figure 2.4), the thinker chooses a structure in proportion to the number of times it was used in any of the examples from memory. For instance, the thinker might generate from "The [Noun] and [Noun]". When reaching any particular terminal (e.g. [*Noun*]), the speaker either reuses an appropriate part from a stored example (e.g. any of the nouns in the examples) or a semantically similar part ("Ox" rather than "Horse"). In this paper we use high-dimensional word embeddings via Global Vectors (Pennington, Socher, & Manning, 2014) to implement a similarity space. Word-vectors that are close together in this space (in a Euclidean or cosine distance sense) tend to be semantically similar in psychological tasks. At the end of this process, the thinker might for instance come up with "The Bear and Ox" as a plausible pub name.

It is important to stress that this model only implements the proposal stage of a conceptual search process. An additional evaluation step – not modeled here – is necessary to accept or reject a particular proposal. There may be additional constraints that one considers in the evaluation, such as the catchiness of the name. "The Bear and Ox" might be a terrible name for a pub for any number of reasons under additional evaluation, but at least it appears reasonable, as opposed to "A Floral and Tasty Essence Wrapped in Chamomile"[2]. The role of the BS proposal model is to get the thinker within a space where good and bad proposals can be evaluated, rather than having to spend the majority of the time with non-starters.

Under the assumption that an evaluation function exists, it is possible to extend the model so that the thinker draws a number of examples at a time, and only reports one of them. This is natural if the thinker must provide an example or series of examples, and cannot choose to simply refuse to provide an example even if it is evaluated as poor. We consider such an extension by having the model draw $k$ samples (BS-k), and choosing among them in proportion to their quality (provided by the assumed evaluation function), relative to the total quality of the sample. If $k = 1$, we recover the original BS model. The $k$ parameter has the psychological interpretation of 'the number of examples people draw and evaluate internally before reporting a single answer'. In order to compare this model to people, we consider the task of coming up

[2]Which is better as an example response for 'make up a wine review'.

with new names for things, specifically movies of different genres and pubs.

## Experiment

### Participants, materials and methods

Two groups of participants, Producers ($N = 40$, 18 female, median age 29) and Raters ($N = 50$, 16 female, median age 33), were recruited through Amazon's Mechanical Turk service and paid a small monetary sum for their participation.

The Producers were asked to come up with 5 new names for 4 different movie genres (action, horror, comedy and romance give 20 movie names in total per participant), and 5 new names for pub names. Producers entered their responses using a free-form text field. The Raters were asked to rate names for different categories on a 1-5 scale ("Very Bad" to "Very Good"). Names were category specific, meaning a particular question might be "How good is the title *Parade of Bullets* as a name for a action movie?"

In order to construct different names, we used an equal mixture of names from the Production experiment, names from real instances of the category, and names from the BS model. To keep the task manageable for Raters, we used 25 names from each source (Production, Real, BS) per category, creating 375 names in total. Each Rater saw half of these names, such that each name was rated by 25 raters.

The Production names were selected by randomly choosing 5 Producers for each category (we chose this method, rather than randomly selecting from all the production data, in order to asses whether Producers come up with better or worse names as their guesses progress). The real names were selected by randomly choosing among all Wikipedia entries for that category (for movies) and from a list of popular pub names (for pubs). The model names were selected by first choosing examples at random from the Wikipedia entries for movie genres, and a list of popular pub names. There were on the order of 1,000 examples for each movie genre, and 260 pub names.[3] These examples were syntactically parsed using a variant of the NLTK package in Python (Bird, Klein, & Loper, 2009), and a library of syntax trees was built for each genre. Syntax trees were then chosen in proportion to how often they appear in the examples, and terminals in each tree were chosen from the appropriate part of speech as it appears in the examples, or randomly replaced with a semantically similar part of speech using GloVe (Pennington et al., 2014) with probability $P_{\text{replacement}} = 0.2$. Once examples are available, generating a new BS example takes less than a second.

### Results

Participants rated both the Production names (those made up by people) and the Real names (those taken from Wikipedia) similarly: the mean ratings were 3.12 for Production and 3.07 for Real. People rated the BS model names lower on average (mean rating was 2.57, difference is significant at $p < 0.001$). The BS-k model with $k = 5$ achieves a mean rating of 2.99, which is still statistically different from the average rating for Production and Real names, but the difference is now much smaller (we expand on why we chose $k = 5$ below).

Figure 3 shows the comparison in more detail, as a distribution over the ratings from 1 to 5. A $\chi^2$ test shows the distributions for the Production names and the Real names are not distinguishable, while the BS and BS-5 models are highly distinguishable from both and from each other ($p < 0.0001$). Table 1 illustrates examples of high quality, low quality and average names for different genres.

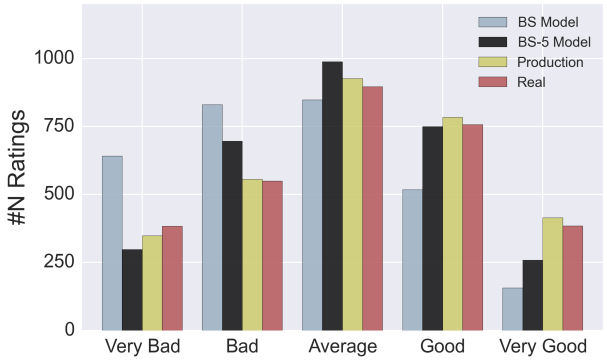


Figure 3: The distribution of ratings for different names, by the source of the name.

The ordinary BS model is not as successful at producing new examples as people, but it is not far. As mentioned, the idea of the BS model is to construct a reasonable space to sample from, as a first step towards in a propose-evaluate cycle. It is likely that people are also considering bad names but not reporting all of them, and this evaluation step is not captured by the BS model. The BS-k extention of the model uses the rating of the model names as a proxy for an evaluation step, considering $k$ samples at a time and choosing one in proportion to its ratings relative to the total rating in the $k$ samples. If $k = 1$, we recover the original model. But as $k$ grows larger, we come closer to the rating distribution of the Production and Real names (Figure 4), where 'closer' means a lower Kullback-Leibler divergence ($KL$), between the distribution of quality scores for the names produced by the BS-k model and the Production or Real names. As Figure 4 indicates, considering only around 3 to 7 samples at a time can make a marked improvement to the BS model. Larger values of $k$ provide diminishing returns, and so for the analyses here we considered $k = 5$. Note that the limiting behavior of increasing $k$ is *not* to produce higher and higher rated names, but rather to converge on a steady-state distribution reflecting a two-stage generate-and-evaluate loop for new names. It is striking how quickly this distribution converges to that of the

[3] A more psychologically plausible version of example recall would use likelihood sampling over the space of movies, and would require asking a different group of subjects to recall actual movies in response to genre prompts.

names made up by people (or the Real names), suggesting that the BS-k model represents at least a plausible first guess for how people come to produce the new names they do.

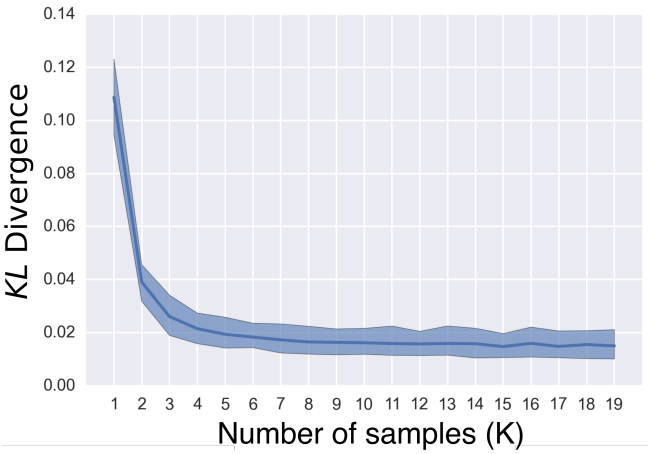


Figure 4: Kullback-Leibler divergence (*KL*) between the distribution of ratings over model names and Production names, for different values of *k* in the BS-k model. Error bars show 95% confidence intervals.

## Discussion

There is something magical about everyday thought, and something odd about how people can respond quickly and reasonably to new questions they never heard with answers no one else has heard. A child could take years to come up with an intuitive biology that unites trees and animals into one concept (Carey, 2009), but a 5-year-old can answer 'What makes the wind?' at the speed of thought. And even if their answer is wrong it might be amusing ("The trees waving their arms make the wind"), rather than absurd ("The wind blows because the moon is bigger than a bag of Doritos").

Such magical stuff deserves hard-nosed experimental scrutiny (Magid et al., 2015), and a better computational answer than "People search randomly through all possible concepts and explanations and evaluate each candidate by how well it explains the data" (Ullman et al., 2012). Here we considered one particular proposal for constructing a reasonable hypothesis space on the fly by using a structural analysis of examples that match the task at hand. This BS model presupposes that thinkers have a way of carrying out such a structural analysis (in our particular case, we assumed a thinker can use grammar to recognize the structural similarity between instances, using "Saving Private Ryan" and "Chasing Amy" to construct a general "VERBing PROPER-NAME" movie schema).

The BS model serves as just one part in a propose-evaluate search, where the evaluation step is difficult to capture and can involve a large amount of world knowledge. Still, for the limited task considered it appears to produce reasonable-sounding new examples for a given category. Within the language domain, the model can potentially be extended beyond short titles to include such things as wine reviews ("The finishing notes are not supported by the light body"). Outside language, a similar approach to idea generation can take advantage of other structures used to organize thoughts. For example, a causal-explanation BS model could analyze the underlying Bayes-net structure of examples from memory, and use those to propose new explanations from a common structure (say, $X \wedge Y -> Z$, the Roman Empire fell because of a combination of tribal invasions and Jewish thought[4]). Applying this approach to generate good hypotheses for causal learning and intuitive theory formation is, we hope, a promising next step – perhaps wrong, but at least not ridiculous, as an account of where learners' new concepts come from.

**Acknowledgments** We are grateful to Laura Schulz for stimulating discussions. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

## References

Baillargeon, R. (2008). Innate ideas revisited: For a principle of persistence in infants' physical reasoning. *Perspectives on Psychological Science*, *3*, 2–13.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. " O'Reilly Media, Inc.".

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological review*, *118*(1), 110.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*, 1085.

Magid, R. W., Sheskin, M., & Schulz, L. E. (2015). Imagination and the generation of new ideas. *Cognitive Development*, *34*, 99–110.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *EMNLP*, *12*, 1532–1543.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*, 199–217. doi: 10.1016/j.cognition.2011.11.005

Schulz, L. (2012a). Finding new facts; thinking new thoughts. *Rational constructivism in cognitive development. Advances in child development and behavior*, *43*, 269–294.

Schulz, L. (2012b). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*, 382–389.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*, 455–480.

[4] http://courses.washington.edu/rome250/gallery/ROME%20250/210%20Reasons.htm

| | Rating | BS Model | Production | Real |
|---|---|---|---|---|
| ACTION | High | The Raging V (3.7)<br>Phoenix First (3.4) | War Hawks (4.1)<br>Retribution (3.9) | Bloodfist II (3.9)<br>Marked for Death (3.8) |
| | Average | Drop (3.0)<br>Conspiracy Mars (2.7) | Stunt Man (3.1)<br>Run Faster (3.1) | Blue Steel (3.2)<br>Bullet in the Head (3.1) |
| | Low | Of Art (1.9)<br>Teenage Chinese (1.5) | The State of Iraq (2.6)<br>Eat It (2.1) | I Come in Peace (2.3)<br>Curry and Pepper (1.5) |
| | | | | |
| HORROR | High | The Dawn (3.5)<br>The Prophecy on One (3.4) | The House of Death (4.2)<br>Cabin of the Dead (4.0) | The Dunwich Horror (4.1)<br>Murders in the Rue Morgue (4.0) |
| | Average | Sharks (2.7)<br>Seed 5 (2.4) | Silence (3.1)<br>Blood Draws (3.0) | The Headless Eyes (3.3)<br>The Wizard of Gore (3.0) |
| | Low | Space Vegas (1.9)<br>Family (1.5) | Paid Maidens 3000 (2.1)<br>Cat Napping (1.7) | Count Yorga, Vampire (2.7)<br>Mephisto Waltz (2.2) |
| | | | | |
| COMEDY | High | Love Punch Drunk (3.0)<br>All American Rita (2.8) | Hot Mess (3.5)<br>Tame Your Own Shrew (3.4) | Above the Limit (3.0)<br>Mr. Flip (3.0) |
| | Average | Puddles of Max (2.6)<br>America Dave (2.6) | Come on Man (2.8)<br>Match This (2.7) | The Enchanted Drawing (2.6)<br>Those Awful Hats (2.6) |
| | Low | Princess Year (2.2)<br>West (2.1) | Jane 2 (2.1)<br>Ha Ha Ha (1.8) | New Pillow Fight (2.3)<br>Clowns Spinning Hats (2.0) |
| | | | | |
| ROMANCE | High | Love in a Separation (3.5)<br>Private Woman (3.4) | When You Least Expect It (3.7)<br>Daydreaming in New York (3.7) | At First Sight (4.0)<br>Bed of Roses (3.7) |
| | Average | The Pearl Rollercoaster (2.7)<br>Walls of Sky (2.5) | Take it Slow (3.3)<br>Cool Happiness in New York (3.0) | Bitter Moon (3.2)<br>Ballistic Kiss (2.9) |
| | Low | A Speckled McKee (2.2)<br>Death (1.4) | Cheeky (2.5)<br>Red Beans and Rice (1.9) | 1871 (2.3)<br>The 5th Monkey (2.0) |
| | | | | |
| PUBS | High | The Hound's Head (3.5)<br>The Royal Hood (3.1) | The Old Barrel (3.8)<br>The Rusty Spur (3.8) | The Rose & Crown (3.8)<br>The White Lion (3.7) |
| | Average | The Green Bay (2.6)<br>The Garter (2.4) | The Dark Forest (3.0)<br>The Dog's Ear (3.0) | The Castle (3.2)<br>The Cross Keys (3.1) |
| | Low | The Cow (2.2)<br>The Pear (1.9) | The Cat's Meow (2.6)<br>The Paper Cut (2.3) | The White Hart (2.5)<br>The New (1.8) |
| | | | | |

Table 1: Examples of different names, organized by source and average rating, with their rating indicated in parenthesis. The names were assigned as Low, Average and High based on their relative position compared to the median rating for that source and genre. 90 names are shown out of the total 375.