

VizNet: Towards A Large-Scale Visualization Learning and Benchmarking Repository

Kevin Hu
MIT Media Lab
kzh@mit.edu

Snehalkumar 'Neil' S. Gaikwad
MIT Media Lab
gaikwad@mit.edu

Michiel Bakker
MIT Media Lab
bakker@mit.edu

Madelon Hulsebos
MIT Media Lab
madelonhulsebos@gmail.com

Emanuel Zraggen
MIT CSAIL
emzg@mit.edu

César Hidalgo
MIT Media Lab
hidalgo@mit.edu

Tim Kraska
MIT CSAIL
kraska@mit.edu

Guoliang Li
Tsinghua University
liguoliang@tsinghua.edu.cn

Arvind Satyanarayan
MIT CSAIL
arvindsatya@mit.edu

Çağatay Demiralp
MIT CSAIL
cagatay@csail.mit.edu

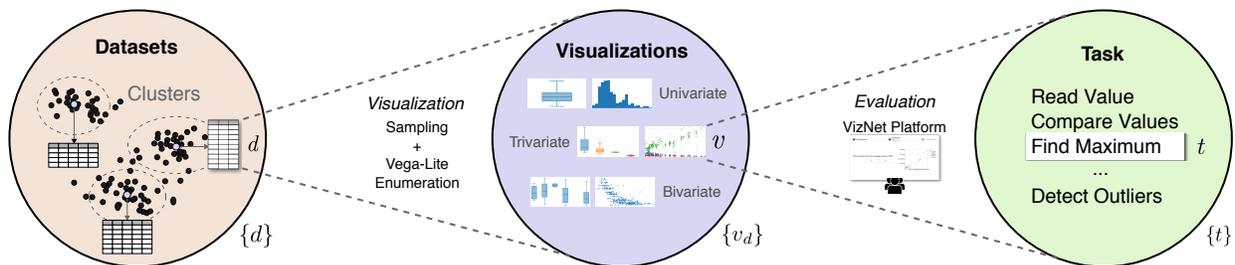


Figure 1: VizNet enables data scientists and visualization researchers to aggregate data, enumerate visual encodings, and crowd-source effectiveness evaluations.

ABSTRACT

Researchers currently rely on ad hoc datasets to train automated visualization tools and evaluate the effectiveness of visualization designs. These exemplars often lack the characteristics of real-world datasets, and their one-off nature makes it difficult to compare different techniques. In this paper, we present VizNet: a large-scale corpus of over 31 million datasets compiled from open data repositories and online visualization galleries. On average, these datasets comprise 17 records over 3 dimensions and across the corpus, we find 51% of the dimensions record categorical data, 44% quantitative, and only 5% temporal. VizNet provides the necessary common baseline for comparing visualization design techniques,

and developing benchmark models and algorithms for automating visual analysis. To demonstrate VizNet's utility as a platform for conducting online crowdsourced experiments at scale, we replicate a prior study assessing the influence of user task and data distribution on visual encoding effectiveness, and extend it by considering an additional task: outlier detection. To contend with running such studies at scale, we demonstrate how a metric of perceptual effectiveness can be learned from experimental results, and show its predictive power across test datasets.

CCS CONCEPTS

• **Human-centered computing** → **Visualization design and evaluation methods; Visualization theory, concepts and paradigms;** • **Computing methodologies** → **Machine learning.**

KEYWORDS

Automated visualization, machine learning, active learning, benchmarking, reproducible research, crowd computing

ACM Reference Format:

Kevin Hu, Snehalkumar 'Neil' S. Gaikwad, Michiel Bakker, Madelon Hulsebos, Emanuel Zraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan, and Çağatay Demiralp. 2019. VizNet: Towards A Large-Scale Visualization Learning and Benchmarking

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-5970-2/19/05...\$15.00

<https://doi.org/10.1145/3290605.3300892>

Repository. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3290605.3300892>

1 INTRODUCTION

A primary concern in visualization is how to effectively encode data values as visual variables. Beginning with Cleveland and McGill’s seminal work [11], researchers have studied this question of *graphical perception* by conducting human subjects experiments. And increasingly, researchers are seeking to operationalize the guidelines such studies produce using handcrafted rule-based systems [49, 73] or learned models [18, 27, 41].

To increase the scale and diversity of the subject pool, modern studies have eschewed traditional laboratory setups in favor of crowdsourcing platforms [24]. But a constraining factor for true ecological validity remains. Collecting, curating, and cleaning data is a laborious and expensive process and, thus, researchers have relied on running studies with ad hoc datasets. Such datasets, sometimes synthetically generated, do not display the same characteristics as data found in the wild. Moreover, as one-off exemplars, their use makes it difficult to compare approaches against a common baseline.

Large-scale databases (such as WordNet [47] and ImageNet [17]) have proven instrumental in pushing the state-of-the-art forward as they provide the data needed to train and test machine learning models, as well as a common baseline for evaluation, experimentation, and benchmarking. Their success has led researchers to call for a similar approach to advance data visualization [3, 20]. However, insufficient attention has been paid to design and engineer a centralized and large-scale repository for evaluating the effectiveness of visual designs.

In response, we introduce VizNet: a corpus of over 31 million datasets (657GB of data) compiled from the web, open data repositories, and online visualization platforms. In characterizing these datasets, we find that they typically consist of 17 records describing 3 dimensions of data. 51% of the dimensions in the corpus record categorical data, 44% quantitative, and only 5% measure temporal information. Such high-level properties, and additional measures such as best statistical fit and entropy, contribute a taxonomy of real-world datasets that can inform assessments of ecological validity of prior studies.

We demonstrate VizNet’s viability as a platform for conducting online crowdsourced experiments at scale by replicating the Kim and Heer (2018) study assessing the effect of task and data distribution on the effectiveness of visual encodings [29], and extend it with an additional task: outlier detection. While largely in line with the original findings,

our results do exhibit several statistically significant differences as a result of our more diverse backing datasets. These differences inform our discussion on how crowdsourced graphical perception studies must adapt to and account for the variation found in organic datasets. VizNet along with data collection and analysis scripts is publicly available at <https://viznet.media.mit.edu>.

Data visualization is an inherently combinatorial design problem: a single dataset can be visualized in a multitude of ways, and a single visualization can be suitable for a range of analytic tasks. As the VizNet corpus grows, assessing the effectiveness of these (*data, visualization, task*) triplets, even using crowdsourcing, will quickly become time- and cost-prohibitive. To contend with this scale, we conclude by formulating effectiveness prediction as a machine learning task over these triplets. We demonstrate a proof-of-concept model that predicts the effectiveness of unseen triplets with non-random performance. Our results suggest that machine learning offers a promising method for efficiently annotating VizNet content. VizNet provides an important opportunity to advance our understanding of graphical perception.

2 RELATED WORK

VizNet is motivated by research in graphical perception, automated visualization based on machine learning, and crowdsourced efforts towards data collection for visualization research. VizNet also draws on the digital experimentation capabilities of large-scale machine learning corpora.

Graphical Perception

Visual encoding of data is central to information visualization. Earlier work has studied how different choices of visual encodings such as position, size, color and shape influence *graphical perception* [12], the decoding of data presented in graphs. Through human subjects experiments, researchers have investigated the effects of visual encoding on the ability to read and make judgments about data represented in visualizations [12, 25, 32, 37, 64–66, 70]. Consequently, prior research has provided rankings of visual variables by user performance for nominal, ordinal, and numerical data [12, 37, 42, 43, 63]. Researchers have also studied how design parameters beyond visual encoding variables such as aspect ratio [9, 23, 67], size [10, 26, 34], chart variation [33, 69], and axis labeling [68] impact the effectiveness of visualizations. Previous studies have evaluated how user task, data types and distributions influence the effectiveness of charts [56] and visual encoding variables [29].

Graphical perception experiments in current practice are typically conducted on single datasets with small size and variety, lacking the characteristics of real-world data. Studies based on ad hoc datasets may provide useful results but are inherently partial and difficult to generalize, reproduce and

compare against. VizNet provides a corpus of real-world tables from diverse domains to make it easier for researchers to run visualization design evaluation studies at scale. VizNet is sufficiently rich both in size and variety to satisfy the data needs of a substantial number of experimental designs, facilitating the comparison of and reasoning about results from different experiments on a common baseline.

Data Collection for Visualization Research

Although researchers recognize the need for data collection and generation to facilitate evaluation across a broad range of real datasets [59, 61], little effort has been made to create centralized corpora for data visualization research. Beagle [4] has been used to scrape over 41,000 visualizations from the web. Similarly, the MassVis [5] database was compiled by scraping over 5,000 visualizations from the web and partially annotating them. Lee et al. [62] recently extracted and classified 4.8 million figures from articles on PubMed Central. However, these datasets do not include the raw data represented by the visualizations, limiting their utility for generalized and reproducible visualization research.

Automated Visualization using Machine Learning

Data-driven models based on responses elicited through human subjects experiments are common in the psychophysics and data visualization literature. For example, low-level perceptual models such as the Weber-Fechner Law, Stevens' Power Law, the CIELAB color space, and perceptual kernels [15] all fit various models to empirical user data, informing low-level visual encoding design. Earlier researchers propose using such models to generate and evaluate visualizations (e.g., [15, 16, 60]).

In a natural extension to these earlier ideas, researchers have recently introduced machine learning-based systems for automated visualization design. Data2Vis [18] uses a neural machine translation approach to create a sequence-to-sequence model that maps JSON-encoded datasets to Vega-lite visualization specifications. Draco-Learn [49] learns trade-offs between constraints in Draco. DeepEye [41] combines rule-based visualization generation with models trained to classify a visualization as "good" or "bad" and rank lists of visualizations. VizML [27] uses neural networks to predict visualization design choices from a corpus of one million dataset-visualization pairs harvested from a popular online visualization tool. Results from this recent work are promising but also point at the need for large-scale real-world training data with sufficient diversity [57]. VizNet addresses this research gap and provides 31 million real-world datasets from everyday domains and can be used for training machine learning models to drive visualization systems.

Machine Learning Corpora

Recent developments of large-scale data repositories have been instrumental in fostering machine learning research. Access to rich, voluminous data is crucial for developing successful machine learning models and comparing different approaches on a common baseline. To this end, researchers have created centralized data repositories for training, testing, and benchmarking models across many tasks. Publicly available repositories such as ImageNet [17], SUN [74], COCO [39], etc. are one of the main drivers behind the rapid advances in deep learning. VizNet is informed and inspired by the digital experimentation capabilities of large-scale data repositories in machine learning research.

3 DATA

VizNet incorporates four large-scale corpora, assembled from the web, online visualization tools, and open data portals.

Corpora

The first category of corpora includes data tables harvested from the web. In particular, we use horizontal relational tables from the WebTables 2015 corpus [6], which extracts structured tables from the Common Crawl. In these tables, entities are represented in rows and attributes in columns.

The second type of corpus includes tabular data uploaded by users of two popular online data visualization and analysis systems. Plotly [53] is a software company that develops visualization tools and libraries. Once created, Plotly charts can be posted to the Plotly Community Feed [54]. Using the Plotly API, we collected approximately 2.5 years of public visualizations from the feed, starting from 2015-07-17 and ending at 2018-01-06. The second system, ManyEyes [71] allowed users to create and publish visualizations through a web interface. It was available from 2007–2015, and was used by tens of thousands of users [50].

The third type of corpus includes public data from the Open Data Portal Watch [48, 51], which catalogs and monitors 262 open data portals such as `data.noaa.gov` from CKAN, `finances.worldbank.org` from Socrata, and `open-data.brussels.be` from OpenDataSoft. The majority of these portals are hosted by governments, and collect civic and social data.

VizNet aggregates these corpora into a centralized repository. However, the majority of datasets are from WebTables. Therefore, in the following sections, we describe each corpus individually with 250K randomly sampled datasets, to avoid oversampling the WebTable corpus. We combine these datasets into a balanced sample of one million datasets, which we refer to as the **VizNet 1M corpus**.

Characterization

Summary statistics and underlying distributions of each of the five corpora are shown in Figure 2. The data type of a column is classified as either categorical, quantitative, or temporal, which we abbreviate as C, Q and T, respectively. This data type is detected using a heuristic-based approach that incorporates column name and value information. For quantitative columns, we use the Kolmogorov-Smirnov test [45] to examine the goodness-of-fit of six distributions: the normal, log-normal, exponential, power law, uniform and chi-squared distributions. We reject the null hypothesis of a distribution fit if the p-value of the associated test is lower than the level $\alpha = 0.05$. If all distributions are rejected at α , we consider the distribution to be undefined. If multiple distributions are not rejected, we consider the “best” fit to be that with the highest p-value. We also report the skewness and percent of outliers, defined as data points that fall more than $1.5 \times IQR$ below the first quartile or above the third quartile, where IQR is the interquartile range. The statistical distribution of categorical columns within each corpus is characterized using the normalized entropy.

4 EXPERIMENT DESIGN

To evaluate the utility of VizNet as a resource for data scientists and visualization researchers, we conducted an experiment where we first replicated the Kim and Heer (2018) prior study [29] using real-world datasets from the VizNet corpus to assess the influence of user task and data distribution on visual encoding effectiveness. These datasets were sampled to match constraints from the prior study and ensure that participants only saw valid data. We then extended this experiment by including an additional task on outlier detection. Finally, we trained a machine learning model that learns the perceptual effectiveness of different visual designs and evaluated its predictive power across test datasets.

Replication of Kim and Heer (2018)

Kim and Heer (2018), “*Assessing Effects of Task and Data Distribution on the Effectiveness of Visual Encodings*,” conducted a crowdsourced experiment measuring subject performance (*i.e.* error rate and response time) across data distributions (**D**), visualization designs (**V**), and task types (**T**). The 24 data distributions characterize trivariate data involving one categorical and two quantitative fields ($C=1, Q=2$) sampled from 2016 daily weather measurements [46] according to univariate entropies of the quantitative fields, cardinalities, and number of records per category.

The authors employed a mixed design using a within-subjects treatment for visual encodings and between-subjects treatments for tasks and data characteristics. They analyzed

responses from 1,920 participants on Amazon’s Mechanical Turk (MTurk), who individually completed 96 questions and 12 engagement checks, and calculated the absolute and ranked performance of different ($\mathbf{D} \times \mathbf{V} \times \mathbf{T}$) conditions, as well as the interaction effects between different data characteristics, visual channels, and task types. These results extended existing models of encoding effectiveness, such as APT [43], and provided valuable insights for automatic visualization design systems.

Datasets

For this experiment, we sampled VizNet datasets according to a procedure that matched constraints from Kim and Heer (2018) and ensured that participants only saw valid data without missing values. This procedure was developed after an initial pilot study with a subset of the corpus in which all datasets were manually verified.

To begin, we identified all datasets with more than one categorical field and two quantitative fields ($C \geq 1$ and $Q \geq 2$). Then, we sampled all possible three column subsets with exactly one categorical and two quantitative fields ($C=1, Q=2$). Following this sampling, we filtered out datasets using a number of constraints. First, we rejected datasets containing any null values. Second, we required that the column names of all datasets must contain between 1 and 50 ASCII-encoded characters. Third, we limited the cardinality (*e.g.* the number of unique groups) of the categorical columns between 3 and 30. Fourth, we restricted the group names between 3 and 30 characters, at least one of which is alphanumeric. Lastly, we required that each of the groups must contain 3 to 30 values. We chose these values to be consistent with the upper and lower constraints of Kim and Heer (2018).

Our sampling procedure resulted in 2,941 valid datasets from the Open Data Corpus (100,626 possible combinations), 6,090 valid datasets from Many Eyes (354,206 combinations), 1,368 from Plotly (347,387 combinations), and 82,150 from a subset of the Webtables corpus (1,512,966 combinations). From this set of candidates, we randomly selected 200 candidates per visualization specification \times task condition. We use **V** to denote the number of visualization specifications and **T** to denote the number of tasks, which leads to 60 such conditions ($\mathbf{V} \times \mathbf{T} = 12 \times 5 = 60$). The 200 number of datasets sampled from the VizNet corpus is consistent with the 192 datasets sampled in Kim and Heer (2018). As a result, this sampling resulted in $200 \times 12 = 2,400$ datasets per task, 2,400 datasets per corpus, and $9,600 = 2,400 \times 4$ total datasets.

Visual Encodings

We selected the twelve visual encoding specifications chosen in Kim and Heer (2018). These encodings are specified using the Vega-Lite grammar [58], which specifies plots using a geometric mark type (*e.g.* bar, line, point) and a mapping

Corpus	Source	Size		Dimensions		Types			Statistical	
		# Data	Gb	Cols	Rows	C (%)	Q (%)	T (%)	Distribution	Entropy
WebTables 2015	Web	90.26M	137	4	5	57.58	35.56	6.86	norm, log-norm, power	0.94
Plotly	Tool	1M	140	3	50	17.29	75.47	7.24	log-norm, norm, power	0.68
Many Eyes	Tool	311K	14	2	19	51.58	46.04	2.48	norm, log-norm, expon	0.81
ODPW	Repository	269K	366	2	70	76.55	21.20	2.24	norm, log-norm, power	0.50
VizNet 1M	Combined	1M	405	3	17	50.71	44.58	4.71	norm, log-norm, power	0.79

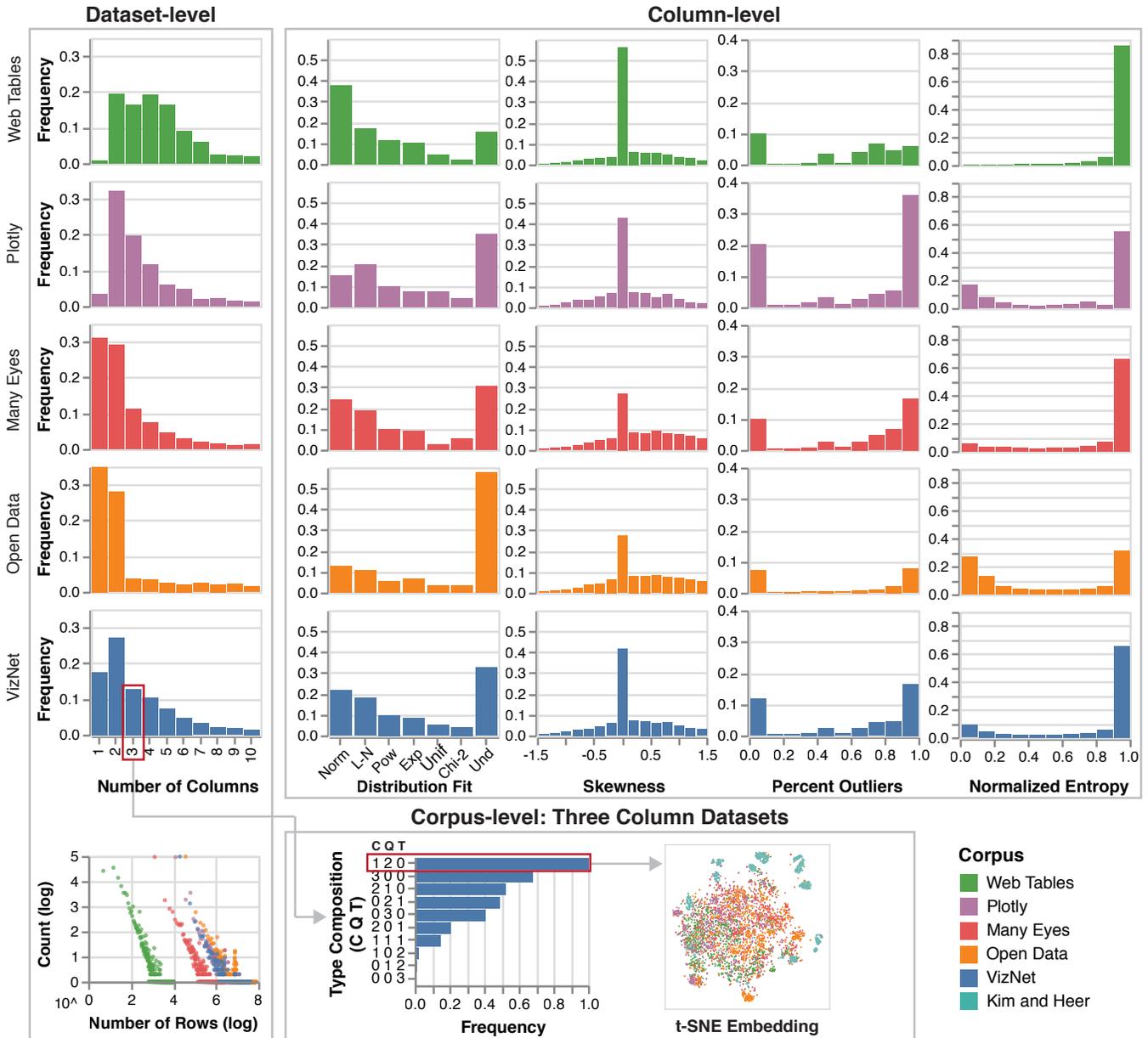


Figure 2: Summary statistics (top) and distributions (bottom) of the four source corpora and the VizNet 1M corpus. In the top table, we report the median number of rows and columns. The Distribution column includes the top three most frequent column distributions. Distributions are abbreviated as Norm = normal, L-N = log-normal, Pow = power law, Exp = exponential, Unif = uniform, and Und = undefined. The bottom part of the figure contains distributions describing columns, datasets, and the entire corpus. The bars outlined in red represent three column datasets and the subset which contain one categorical and two quantitative fields. The clustering of three column (C=1, Q=2) datasets is shown in more detail in Figure 5.

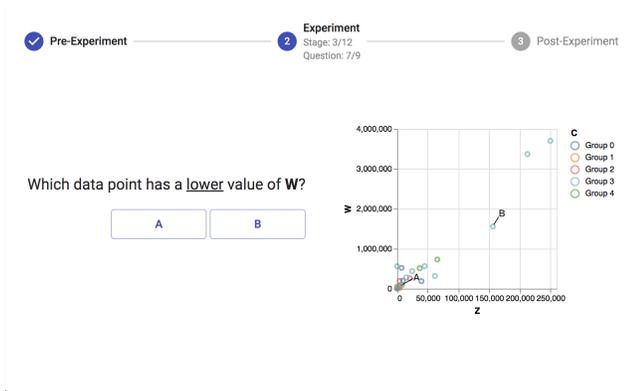


Figure 3: VizNet user interface for the *Compare Values* task experiment.

from data fields to visual encoding channels (e.g. x , y , *color*, *shape*, and *size*). In particular, Kim and Heer (2018) used twelve visualization designs, all of which are scatterplots (a *point* mark) with different mappings between data and encoding channels.

We used the *Tableau-10* scheme for color encoding categorical fields with cardinality less than 10, and *Tableau-20* for categorical fields with cardinality greater than or equal to 20. For positional encodings, in contrast to Kim and Heer (2018), we used a heuristic to determine whether an axis should start at zero. If the range of a variable Q is less than 10% of maximum value $0.1 \times |\max(Q)|$, then we default to Vega-lite axis ranges. Based on a pilot study, we found that this heuristic was necessary to ensure that no questions were prohibitively difficult.

Tasks

Following Kim and Heer (2018), we considered 4 visualization tasks informed by the Amar et al. (2005) [2] taxonomy of low-level analytic activities. Two of those tasks were *value tasks*: *Read Value* and *Compare Values* asked users to read and compare individual values. The other two tasks were *summary tasks*: *Find Maximum* and *Compare Averages* required the identification or comparison of aggregate properties. Each of these tasks was formulated as a binary question (two-alternative forced choice questions). We generated the two alternatives using the procedure described in the prior study.

Procedure

Identical to Kim and Heer (2018), we also employed a mixed design incorporating a within-subjects treatment for visual encodings and a between-subjects treatment for tasks. Each participant answered 9 questions (1 attention check and 8

real) for each of the 12 visual encodings, presented in a random order. Every participant was assigned to a specific task. Unlike Kim and Heer (2018), we did not incorporate dataset conditions. Each dataset was selected randomly from the pool of 200 datasets per $V \times T$ condition. In order to ensure reliable human judgment, we followed the process from Kim and Heer (2018) and incorporated 12 evenly distributed gold standard tasks. The gold standard tasks presented a user with a real dataset encoded in the present visual encoding condition, and asked what information is presented in the visual channel that encodes the first quantitative column (Q_1).

Participants

Crowdsourcing platforms such as MTurk are widely used to recruit participants and conduct online experiments at scale [30, 44]. We recruited in total 1,342 MTurk workers who were located in the U.S. and had $\geq 95\%$ HIT approval rating.

During the analysis, we included the following criteria to ensure the quality of human judgment: we selected subjects who accurately answered 100% of the gold standard questions, had an experimental error rate of less than 60%, and can effectively distinguish colors. We had set the gold standard response exclusion threshold to 100% (i.e., discarding responses if even 1 out of these 12 questions was answered incorrectly). We have verified that a more lenient 80% exclusion threshold does not significantly change the results. Kim and Heer (2018) does not report a dropout rate, making it difficult to assess whether and by how much our dropout rate differs. We included two Ishihara color blindness plate tests [28] along with two pre-screen questions to ensure the participants can effectively distinguish colors. A total of 96.47% reported no vision deficiency and were allowed to participate in the experiment. This resulted in a total of 624 participants' data for in the analysis.

Of the 624 participants, 43.75% were male, 55.44% female, and 0.48% non-binary. 6.38% of the participants had no degree, whereas others had bachelor's (43.10%), master's (14.90%), Ph.D. (3.04%), associate (14.58%) degrees as well as a high school diploma (17.46%). Each participant received 1.00 USD in compensation, which we calculated using the average times of a pilot study and the same hourly wage of Kim and Heer (2018).

5 RESULTS

In this section, we describe the results of our experiment, compare them with the results of Kim and Heer (2018) [29], and demonstrate a machine learning-based approach to predicting effectiveness from (*data*, *visualization*, *task*) triplets.

Comparing Subject Performance

We first compared subject performance with the quantitative results of Kim and Heer (2018) by considering aggregate error rates and log response times per visualization specification and task condition ($V \times T = 12 \times 4$). Following this, we calculated mean error rates with 95% bootstrapped confidence intervals, performed by sampling participants with replacement. To analyze the difference of mean error rates and response times we conducted permutation tests with 10^4 permutations. We test significance at a significance level of $\alpha = 0.05$ with Bonferroni correction for our $m = 48$ hypotheses. The results for the error rate and log response times are shown in Figure 4.

The absolute error rates of our replication tend to agree with those of Kim and Heer (2018) for the *Read Value* task, and to a lesser extent for the *Compare Values* task. The rankings of different visual encodings are also similar. However, for the the *summary tasks* (*Find Maximum* and *Compare Averages*), our observed error rates depart from those of Kim and Heer (2018). Though more data points are needed to draw meaningful conclusions, these results suggest that real-world data affects error rates for more complex tasks.

In contrast, the absolute response times in our study seem to be systematically longer for all tasks except the *Compare Values* task. However, the relative rankings of different encoding are consistent with those of Kim and Heer (2018).

Extending with an Outlier Detection Task

As suggested by Kim and Heer (2018), investigating additional task types is a promising direction of future research. In particular, tasks with more subjective definitions, such as *Cluster* and *Find Anomalies* were not included in Kim and Heer (2018). Nevertheless, as outlier detection is one of the most important data analysis tasks in practice, it warrants further empirical study. We extended the prior work by considering this latter task of identifying “which data cases in a set S of data cases have unexpected/exceptional values.”

We generated 2,400 datasets using the sampling methodology described in the previous section. First, we presented users with a definition of outliers as “observations that lie outside the overall pattern of distribution.” Then, using the same experiment design, we assessed answers to the question “Are there outliers in Q_1 ?” “Yes” and “No” are provided as response options. Outliers were determined using the median absolute deviation (MAD)-based approach described in [38], which is robust to varying sample sizes, compared to other simple approaches.

We found that the error rates for the outlier detection task are higher compared to the other tasks (see Figure 4). This may be due to an inadequate measure of ground truth, inconsistent definitions, or lack of prior training. It is important

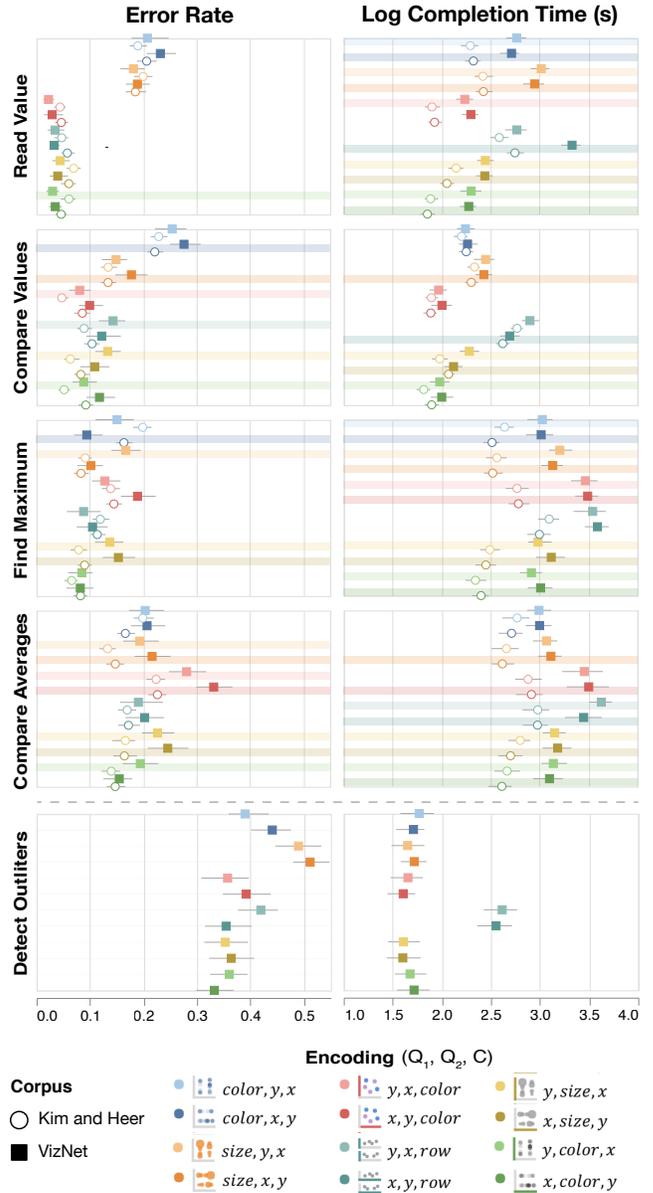


Figure 4: Bootstrapped means and 95% confidence intervals for error rates (left) and log response times (right) across tasks and visual encodings for Kim and Heer (2018) original data, and our replication on VizNet. We reuse the original color encoding of Kim and Heer (2018). Shading indicates a statistically significant difference.

to note that the specification rankings resemble that of the *Read Value* task: *color* and *size* trail behind other encodings channels. Conversely, the log response times are significantly shorter than for other tasks, for all except the faceted charts with *row* encodings.

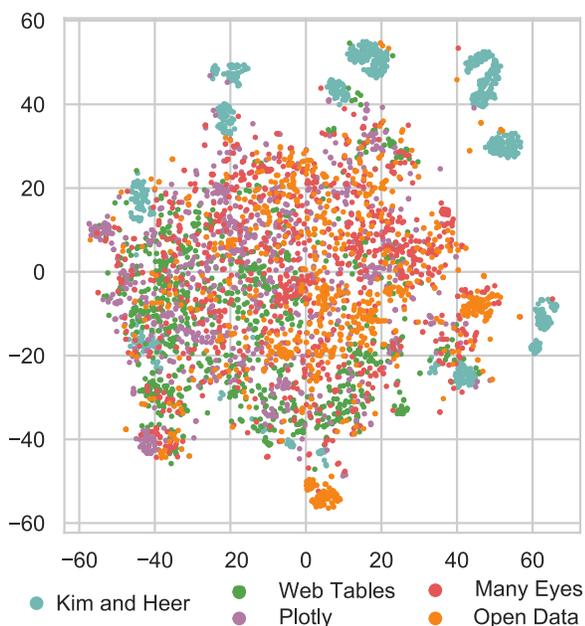


Figure 5: Two-dimensional t-SNE projection of datasets with one categorical and two quantitative columns, evenly sampled from Kim and Heer (2018) and the four corpora within VizNet, with a perplexity of 75.

Learning a Model to Predict Effectiveness

To characterize a dataset, we extracted 167 features: 60 per quantitative field Q , 11 for the categorical field C , 15 for the Q - Q pair, 6 for the two C - Q pairs, and 9 which consider all three fields. These features characterized summary statistics (e.g. coefficient of variance and kurtosis), statistical distributions (e.g. entropy and statistical fits), pairwise relationships (e.g. correlations and one-way ANOVA p-values), clusteredness and spatial autocorrelation.

We first decoded diversity within our space of datasets using these features. Using principal components analysis, we computed 32 principal components which collectively explain over 85% of the variance within our dataset. Then, we generated a two-dimensional t-SNE projection of these principal components, as shown in Figure 5. It is important to note that the datasets used in Kim and Heer (2018) [29] are highly clustered and separate from the datasets used within our replication. This observation is robust for different numbers of principal components and values of perplexity (5-200).

To predict log completion time we use gradient boosted regression trees, a model with strong “off-the-shelf” performance. Training on 80% sample of the data, we were able to predict log completion times in a 20% hold-out test set with a 5-fold cross-validated R^2 of 0.47, which strongly outperforms

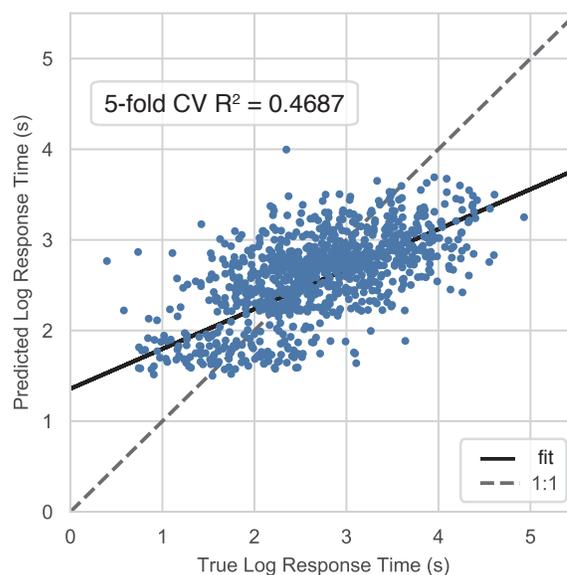


Figure 6: Observed log response times (in seconds) vs. those predicted by a gradient boosted regression tree. The dotted diagonal line denotes a perfect relationship between observation and prediction.

baseline models such as K-nearest neighbors and simple linear regression. A scatter plot of observed vs. predicted values for the top performing model is shown in Figure 6. Learning curves in Figure 7 indicate that, despite the large number of features, our model does not overfit on the training set, and that there are still gains from increasing the number of training samples.

Kim and Heer (2018) reports the trade-off between response time and error rate. To capture this trade-off, we created a combined metric from the log response times and error rate metrics by partitioning the log response times into 20% quantiles, and the error rates into five bins of equal width, for a total of 25 pairs. Then, we characterized each (d, v, t) triplet with the associated (response time + error rate) pair, and resampled minority classes using the Synthetic Minority Over-sampling Technique (SMOTE) [7]. Training a gradient boosted classification tree on the balanced training set resulted in a Top-3 prediction accuracy of 52.48%.

Limitations

Although we have successfully demonstrated the effectiveness of VizNet, it is important to acknowledge limitations. Replication and reproducibility are essential to advance research [52]. In the experiment, we attempted to replicate Kim and Heer (2018) as closely as possible. However, due to practical constraints, we introduced clarifying modifications to the question text and interface design. Due to variance between crowd workers, we were not able to recruit the

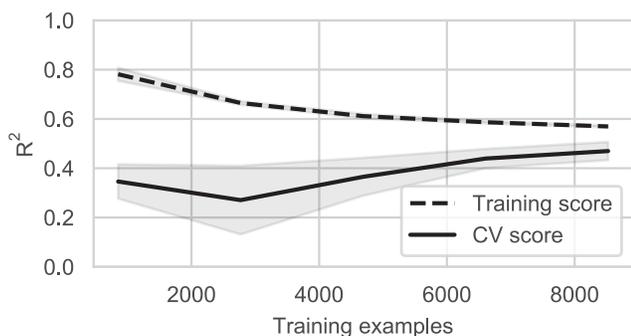


Figure 7: Training R^2 and 5-fold cross-validation R^2 as the number of training examples increases.

same participants; nor do we control for question difficulty, which is calibrated in Kim and Heer (2018). Most of all, we did not exactly replicate the original conditions of the synthetic datasets, which would have limited the amount of real-world VizNet datasets available for sampling. Notwithstanding these limitations, our work provides an important direction to understand the opportunities and challenges faced in replicating prior work in human-computer interaction and visualization research.

With respect to extending the experiment to include an additional task, we note that outlier detection, unlike the other tasks, does not have a defined ground truth. Though we used a robust outlier detection method, there may be a limitation to any purely quantitative method that does not rely on human consensus. The lack of an objective notion of outliers and absence of a clear definition thereof in the questions, reinforces the inconsistency between ground truth and crowdsourced labels presumably partially explaining the consistently high error rate. In the context of the machine learning model, while human judgments can play an important role in help predicting perceptual effectiveness, crowdsourced training data can be noisy. The current experiment was unable to analyze lower bound requirements of quality data, but VizNet’s diverse dataset offers such opportunity for future research.

6 DISCUSSION

There are several important areas where VizNet makes important contributions. VizNet provides a noteworthy contribution to advance our knowledge of effective graphical perception by enabling scientific community access to rich datasets for visualization learning, experimentation, replication, and benchmarking. VizNet offers both the full corpus and the sampled corpus of one million datasets (VizNet 1M). It further described the dimensions, types, and statistical properties of these datasets. The voluminous collection of VizNet complements synthetically generated data. Moreover,

the properties of the VizNet corpus can inform assessments of the ecological validity of other corpora from domains beyond VizNet.

Implications of enabling the VizNet interface for the scientific community. We envision that in the long run, adoption of a common corpus and benchmarks by the visualization community will facilitate the sharing and comparing of results at scale. We have made VizNet publicly available at <https://viznet.media.mit.edu>. A taxonomy in VizNet is formed by splitting our corpus first on the number of columns of a datasets, and then on the composition of column types. Therefore, we should design interactions to help users query, filter, sample datasets within this taxonomy (e.g. give me all datasets with one categorical, two quantitative, and one temporal field). Moreover, this informs the need for supporting keyword search to allow filtering by domain, in addition to filtering on other dataset properties (e.g. give me highly correlated datasets with exactly two quantitative fields).

Implications of VizNet for replication and experimentation. We replicate Kim and Heer (2018) to demonstrate the utility of using VizNet. Our results with real-world data are largely consistent with their findings. As a result of our more diverse backing datasets, however, there are statistically significant differences in error rates for the complex tasks. We also note that task completion times with real data are consistently longer for all but one task. These discrepancies suggest that graphical perception studies must account for the variation found in real datasets. Kim and Heer (2018) acknowledge this direction of future work by describing the need for investigating “all [data] distributions of potential interest.” The process of harvesting these diverse distributions would be facilitated by using VizNet. We further extend the original experiment by considering an additional “detect outliers” task, an important but subjective visual analysis task that is difficult to assess using synthetic data.

Implications of VizNet for learning a metric of perceptual effectiveness. While Kim and Heer (2018) employed a mixed effects model to analyze their results, we proposed to conceive the harvested data as a collection of (*data, visualization, task*) triplets, each of which is associated with effectiveness measures. Using machine learning models, we predicted the completion time with an R^2 value of 0.47. Acknowledging the trade-off between completion time and error rate, we constructed a combined metric and achieved a top-3 prediction accuracy of 52.48%. Despite the noise and skew of crowdsourced labels, and a relatively small sample size, these results out-perform both random chance and baseline classifiers. In doing so, they illustrate the potential for learning a metric of perceptual effectiveness from experimental results.

7 FUTURE WORK

We plan to extend VizNet along three major directions: (1) incorporate and characterize more datasets, (2) harness the wisdom of the crowd, and (3) develop active learning algorithms for optimal experiment design.

Incorporate and characterize more datasets. VizNet currently centralizes four corpora of data from the web, open data portals, and online visualization galleries. We plan to expand the VizNet corpus with the 410,554 Microsoft Excel workbook files (1,181,530 sheets) [8] extracted from the ClueWeb09 web crawl¹. Furthermore, Morton et. al. [50] report 73,000 Tableau workbooks and 107,500 datasets from Tableau Public, which could be integrated into VizNet. Lastly, we plan to incorporate 10,663 datasets from Kaggle², 1,161 datasets included alongside the R statistical environment³, and to leverage the Google Dataset Search⁴ to source more open datasets.

In the future work, we plan to characterize the semantic content within column and group names using natural language processing techniques such as language detection, named entity recognition, and word embeddings. Moreover, as we describe the features of datasets within the VizNet corpus, we can characterize the bias between corpora in terms of dimensions, type composition, and statistical properties of columns. This will enable us to systematically study the extent to which these corpora differ. The existence of such bias between corpora is clear from the previous data section § 3. A clearer understanding of between-corpus bias could inform future techniques for sampling from the VizNet corpus.

Harness the wisdom of the crowd. Domain specific crowdsourcing platforms such as FoldIt, EteRNA, Galaxy-Zoo, and Game with Purpose, have incentivized citizen scientists to discover new forms of proteins [14], RNAs [36], galaxies [40], and artificial intelligence algorithms [72]. We envision VizNet will enable citizen scientists and visualization researchers to execute graphical perception experiments at scale. In recent years, crowdsourcing has been pivotal in the creation of large-scale machine learning corpora. Daemo [21], a self-governed crowdsourcing marketplace, was instrumental in the creation of the Stanford Question Answering Dataset (SQuAD) [55], whereas MTurk was used to curate the ImageNet dataset[17].

The effectiveness of the crowdsourcing has also been exemplified in our experiment while collecting the human judgments for the critical evaluation of visual designs. It is interesting to note that some of the crowd workers enjoyed the intellectual aspect of the experiment, as illustrated by

¹<http://lemurproject.org/clueweb09.php>

²<https://www.kaggle.com/datasets>

³<https://github.com/vincentarelbundock/Rdatasets>

⁴<https://toolbox.google.com/datasetsearch>

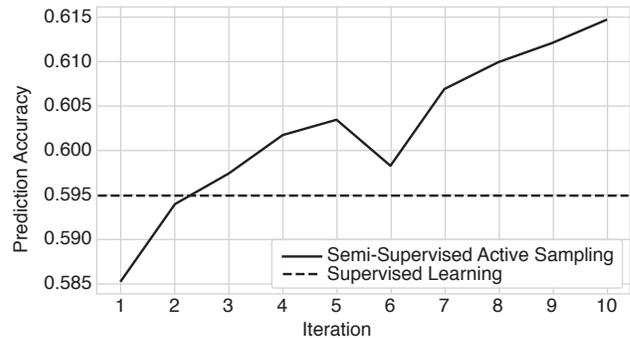


Figure 8: Performance curves obtained by semi-supervised active learning and supervised learning over 10 iterations.

post experiment responses: (1) ‘I found this survey entertaining, it makes you think and use your head’ (2) ‘It is a very interesting survey to carry out since it promotes the capacity of analysis I congratulate you for that’. A natural progression to harness crowdsourcing mechanisms for VizNet includes extension of literature on task design [31], crowd work quality improvements [19, 35], and incentive design [22, 72].

Develop active learning for optimal experiment design. Although gathering human-judgment labels for each triplet is costly, it is possible to learn the effectiveness from labeled triplets to predict labels for unseen ones (see section § 5). In order to further illustrate this strategy we conducted a small experiment on the same data as in section § 5 where the completion times are categorized into low, medium and high. To propagate labels we employed self-learning [1], so we added the model predictions with high certainty to the labelled set. The predictions with low certainty were replaced with crowdsourced labels following the uncertainty algorithm [13]. Figure 8 shows how this strategy improves the accuracy on a test set after a number of iterations against the baseline of training on all labeled samples (supervised learning). In the future, we plan to harness active learning to assess the quality of human judgment.

8 CONCLUSION

Large-scale data collection efforts for facilitating research are common across sciences and engineering, from genomics to machine learning. Their success in accelerating the impact of research in respective fields is a testament to the importance of easy access to large-scale realistic data as well as benchmarking and performing research on shared databases. As the field of data visualization research grows from its infancy, we expect the need for and utility of large-scale data and visualization repositories to significantly grow as well. VizNet is a step forward in addressing this need.

REFERENCES

- [1] Ashok K. Agrawala. 1970. Learning with a probabilistic teacher. (1970).
- [2] Robert Amar, James Eagan, and John Stasko. 2005. Low-level components of analytic activity in information visualization. In *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*. arXiv:15334406
- [3] Leilani Battle, Remco Chang, Jeffrey Heer, and Michael Stonebraker. 2017. Position statement: The case for a visualization performance benchmark. In *2017 IEEE Workshop on Data Systems for Interactive Analysis (DSIA)*. 1–5.
- [4] Leilani Battle, Peitong Duan, Zachery Miranda, Dana Mukusheva, Remco Chang, and Michael Stonebraker. 2018. Beagle: Automated Extraction and Interpretation of Visualizations from the Web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 594, 8 pages.
- [5] Michelle A. Borkin, Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315.
- [6] Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. WebTables: Exploring the Power of Tables on the Web. *Proc. VLDB Endow.* 1, 1 (Aug. 2008), 538–549.
- [7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321–357.
- [8] Zhe Chen, Sasha Dadiomov, Richard Wesley, Gang Xiao, Daniel Cory, Michael Cafarella, and Jock Mackinlay. 2017. Spreadsheet Property Detection With Rule-assisted Active Learning. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 999–1008.
- [9] William S. Cleveland. 1993. *Visualizing Data*. Hobart Press.
- [10] William S. Cleveland, Persi Diaconis, and Robert McGill. 1982. *Variables on scatterplots look more highly correlated when the scales are increased*. Technical Report. DTIC Document.
- [11] William S. Cleveland and Robert McGill. 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *J. Amer. Statist. Assoc.* 79, 387 (1984), 531–554.
- [12] William S. Cleveland and Robert McGill. 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *J. Amer. Statist. Assoc.* 79, 387 (1984), 531–554.
- [13] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research* 4 (1996), 129–145.
- [14] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756.
- [15] Çağatay Demiralp, Michael Bernstein, and Jeffrey Heer. 2014. Learning Perceptual Kernels for Visualization Design. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2014).
- [16] Çağatay Demiralp, Carlos Scheidegger, Gordon Kindlmann, David Laidlaw, and Jeffrey Heer. 2014. Visual Embedding: A Model for Visualization. *Computer Graphics and Applications* (2014).
- [17] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. 2009. Imagenet: A large-scale hierarchical image database. In *In CVPR*.
- [18] Victor Dibia and Çağatay Demiralp. 2018. Data2Vis: Automatic Generation of Data Visualizations Using Sequence to Sequence Recurrent Neural Networks. *CoRR abs/1804.03126* (2018). arXiv:1804.03126
- [19] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1013–1022.
- [20] Philipp Eichmann, Carsten Binnig, Tim Kraska, and Emanuel Zraggen. 2018. IDEBench: A Benchmark for Interactive Data Exploration. *ArXiv e-prints* (April 2018). arXiv:cs.DB/1804.02593
- [21] Snehal Neil Gaikwad, Durim Morina, Rohit Nistala, Megha Agarwal, Alison Cossette, Radhika Bhanu, Saiph Savage, Vishwajeet Narwal, Karan Rajpal, Jeff Regino, et al. 2015. Daemo: A self-governed crowdsourcing marketplace. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 101–102.
- [22] Snehal Neil S Gaikwad, Durim Morina, Adam Ginzberg, Catherine Mullings, Shirish Goyal, Dilrukshi Gamage, Christopher Diemert, Mathias Burton, Sharon Zhou, Mark Whiting, et al. 2016. Boomerang: Rebounding the consequences of reputation feedback on crowdsourcing platforms. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 625–637.
- [23] Jeffrey Heer and Maneesh Agrawala. 2006. Multi-Scale Banking to 45 Degrees. *IEEE Trans. Visualization & Comp. Graphics* 12 (2006), 701–708.
- [24] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 203–212.
- [25] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *ACM Human Factors in Computing Systems (CHI)*.
- [26] Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. 2009. Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations. In *ACM Human Factors in Computing Systems (CHI)*.
- [27] Kevin Zeng Hu, Michiel A. Bakker, Stephen Li, Tim Kraska, and César Hidalgo. 2018. VizML: A Machine Learning Approach to Visualization Recommendation. *ArXiv e-prints* (Aug. 2018). arXiv:cs.HC/1808.04819
- [28] Shinobu Ishihara. 1960. *Tests for colour-blindness*. Kanehara Shuppan Company.
- [29] Younghoon Kim and Jeffrey Heer. 2018. Assessing Effects of Task and Data Distribution on the Effectiveness of Visual Encodings. *Computer Graphics Forum (Proc. EuroVis)* (2018).
- [30] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.
- [31] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.
- [32] Nicholas Kong, Jeffrey Heer, and Maneesh Agrawala. 2010. Perceptual Guidelines for Creating Rectangular Treemaps. *IEEE Trans. Visualization & Comp. Graphics* 16, 6 (2010), 990–998.
- [33] Robert Kosara and Drew Skau. 2016. Judgment Error in Pie Chart Variations. In *Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization*. Wiley Online Library, 91–95.
- [34] Heidi Lam, Tamara Munzner, and Robert Kincaid. 2007. Overview Use in Multiple Visual Information Resolution Interfaces. *IEEE Trans. Visualization & Comp. Graphics* 13, 6 (2007), 1278–1285.
- [35] John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, Vol. 2126.
- [36] Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille, et al. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* 111, 6 (2014), 2122–2127.

- [37] Stephan Lewandowsky and Ian Spence. 1989. Discriminating strata in scatterplots. *Journal of American Statistical Association* 84, 407 (1989), 682–688.
- [38] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. 49 (07 2013), 764–766.
- [39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014).
- [40] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 3 (2008), 1179–1189.
- [41] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. 2018. DeepEye: Towards Automatic Data Visualization. *The 34th IEEE International Conference on Data Engineering (ICDE)* (2018).
- [42] Alan M. MacEachren. 1995. *How Maps Work: Representation, Visualization, and Design*. Guilford Press.
- [43] Jock D. Mackinlay. 1986. Automating the Design of Graphical Presentations of Relational Information. *ACM Trans. Graph.* 5, 2 (1986), 110–141.
- [44] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.
- [45] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.
- [46] Matthew J. Menne, Imke Durre, Russell S. Vose, Byron E. Gleason, and Tamara G. Houston. 2012. An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology* 29, 7 (2012), 897–910.
- [47] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
- [48] Johann Mitlohner, Sebastian Neumaier, Jurgen Umbrich, and Axel Polleres. 2016. Characteristics of Open Data CSV Files. In *2016 2nd International Conference on Open and Big Data (OBD)*. 72–79.
- [49] Dominik Moritz, Chenglong Wang, Greg L. Nelson, Halden Lin, Adam M. Smith, Bill Howe, and Jeffrey Heer. 2018. Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2018).
- [50] Kristi Morton, Magdalena Balazinska, Dan Grossman, Robert Kosara, and Jock Mackinlay. 2014. Public data and visualizations: How are many eyes and tableau public used for collaborative analytics? *SIGMOD Record* 43, 2 (6 2014), 17–22.
- [51] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. 2016. Automated Quality Assessment of Metadata across Open Data Portals. *Journal of Data and Information Quality* (2016). arXiv:arXiv:1502.07526v1
- [52] Roger D Peng. 2011. Reproducible research in computational science. *Science* 334, 6060 (2011), 1226–1227.
- [53] Plotly. 2018. Plotly. <https://plotly>.
- [54] Plotly. 2018. Plotly Community Feed. <https://plot.ly/feed>.
- [55] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2383–2392. <https://doi.org/10.18653/v1/D16-1264>
- [56] Bahador Saket, Alex Endert, and Çağatay Demiralp. 2018. Task-Based Effectiveness of Basic Visualizations. *IEEE Transactions on Visualization and Computer Graphics* (2018), 1–1.
- [57] Bahador Saket, Dominik Moritz, Halden Lin, Victor Dibia, Çağatay Demiralp, and Jeffrey Heer. 2018. Beyond Heuristics: Learning Visualization Design. arXiv:arXiv:1807.06641
- [58] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan. 2017), 341–350.
- [59] Christoph Schulz, Arlind Nocaj, Mennatallah El-Assady, Steffen Frey, Marcel Hlawatsch, Michael Hund, Grzegorz Karch, Rudolf Netzler, Christin Schätzle, Miriam Butt, et al. 2016. Generative data models for validation and evaluation of visualization techniques. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*. ACM, 112–124.
- [60] Michael Sedlmair and Michaël Aupetit. 2015. Data-driven Evaluation of Visual Quality Measures. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 201–210.
- [61] Michael Sedlmair, Andrada Tatu, Tamara Munzner, and Melanie Tory. 2012. A taxonomy of visual cluster separation factors. In *Computer Graphics Forum*, Vol. 31. 1335–1344.
- [62] Po shen Lee, Jevin D. West, and Bill Howe. 2018. Viziometrics: Analyzing Visual Information in the Scientific Literature. *IEEE Transactions on Big Data* 4, 1 (2018), 117–129.
- [63] Barbara G. Shortridge. 1982. Stimulus processing models from psychology: can we use them in cartography? *The American Cartographer* 9 (1982), 155–167.
- [64] David Simkin and Reid Hastie. 1987. An Information-Processing Analysis of Graph Perception. *Journal of American Statistical Association* 82, 398 (1987), 454–465.
- [65] Drew Skau and Robert Kosara. 2016. Arcs, Angles, or Areas: Individual Data Encodings in Pie and Donut Charts. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 121–130.
- [66] Ian Spence and Stephan Lewandowsky. 1991. Displaying Proportions and Percentages. *Applied Cognitive Psychology* 5 (1991), 61–77.
- [67] Justin Talbot, John Gerth, and Pat Hanrahan. 2011. Arc Length-based Aspect Ratio Selection. *IEEE Trans. Visualization & Comp. Graphics* (2011).
- [68] Justin Talbot, Sharon Lin, and Pat Hanrahan. 2010. An Extension of Wilkinson’s Algorithm for Positioning Tick Labels on Axes. *IEEE Trans. Visualization & Comp. Graphics* (2010).
- [69] Justin Talbot, Vidya Setlur, and Anushka Anand. 2014. Four Experiments on the Perception of Bar Charts. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec 2014), 2152–2160. <https://doi.org/10.1109/TVCG.2014.2346320>
- [70] Lothar Tremmel. 1995. The Visual Separability of Plotting Symbols in Scatterplots. *Journal of Computational and Graphical Statistics* 4, 2 (1995), 101–112.
- [71] Fernanda B. Viégas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. 2007. ManyEyes: A Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov. 2007), 1121–1128.
- [72] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.
- [73] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2016. Towards A General-Purpose Query Language for Visualization Recommendation. In *ACM SIGMOD Human-in-the-Loop Data Analysis (HILDA)*.
- [74] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. 2016. SUN Database: Exploring a Large Collection of Scene Categories. *Int. J. Comput. Vision* (2016).