Pragmatic inference and visual abstraction enable contextual flexibility during visual communication

Judith E. Fan[a,1], **Robert X.D. Hawkins**[a], **Mike Wu**[b], **and Noah D. Goodman**[a,b]

[a]Department of Psychology, Stanford University; [b]Department of Computer Science, Stanford University

This manuscript was compiled on December 11, 2018

Visual modes of communication are ubiquitous in modern life — from maps to data plots to political cartoons. Here we investigate drawing, the most basic form of visual communication. Communicative drawing poses a core challenge for theories of how vision and social cognition interact, requiring a detailed understanding of how sensory information and social context jointly determine what information is relevant to communicate. Participants (N=192) were paired in an online environment to play a sketching-based reference game. On each trial, both participants were shown the same four objects, but in different locations. The sketcher's goal was to draw one of these objects — the target — so that the viewer could select it from the array. There were two types of trials: close, where objects belonged to the same basic-level category (i.e., bird, car, chair, dog) [mwu: this wording sounds like same category is defined as (bird, car, chair, dog) which are actually different categories.], and far, where objects belonged to different categories. We found that people exploited information in common ground with their partner to efficiently communicate about the target: on far trials, sketchers achieved high recognition accuracy while applying fewer strokes, using less ink, and spending less time on their drawings than on close trials. We hypothesized humans succeed in this task by recruiting two core competencies: (1) visual abstraction, the capacity to perceive the correspondence between an object and a drawing of it; and (2) pragmatic inference, the ability to infer what information would help a viewer distinguish the target from distractors. To evaluate this hypothesis, we developed a computational model of the sketcher that embodied both competencies, instantiated as a probabilistic program "wrapped" around a deep convolutional neural network. We found that this model fit human data well and outperformed lesioned variants. Together, this work provides the first algorithmically explicit theory of how perception and social cognition jointly support contextual flexibility in visual communication.

drawing | social cognition | perception | deep learning | probabilistic models

A watershed moment in the history of human cognition and culture was the invention of graphical representation, independently in Europe and Asia, about 30-60 thousand years ago (1, 2). Graphical representation was transformative because it provided a means for people to encode their thoughts in a durable and shareable format (3). From ancient etchings on cave walls to modern digital displays, using graphical representations for visual communication lies at the heart of key human innovations (e.g., mapmaking, data visualization), and forms the foundation for the cultural transmission of knowledge and higher-level reasoning (4, 5). Drawing, in which a person produces marks that form a meaningful image, is a particularly important case study for understanding human visual communication. Drawn images predate symbolic writing systems (6), are pervasive in many cultures (7), and are produced prolifically by children from an early age (8).

Remarkably, we perceive drawings of objects as resembling physical objects in spite of the fact that drawings and objects are profoundly different in composition. What explains their effectiveness for conveying visual concepts? One hypothesis explored in prior work is that the ability to perceive the correspondence between drawings and real-world objects arises from a common, general-purpose neural architecture evolved to handle natural visual inputs (9, 10). In support of this hypothesis, it was recently shown that features learned by deep convolutional neural network models (DCNNs) trained exclusively to recognize objects in photos and had never seen a line drawing succeed in recognizing simple sketches (11, 12). These results are consistent with evidence from other domains, including developmental, cross-cultural, and comparative studies of drawing perception. For example, human infants (13), people living in remote regions without pictorial art traditions and without substantial contact with Western visual media (14), and higher non-human primates (15) are able to recognize line drawings of familiar objects, even without prior experience with drawings. Together, these findings suggest that a thorough understanding of the functional architecture of the visual system, which is largely shared across human cultures, may be sufficient to explain how people are able to perceive semantic content in drawings.

However, such an explanation is clearly incomplete. Even drawings of the same object can be highly variable (16), illustrating the importance of other factors that may influence how a drawing conveys meaning. For example, extended visual communication within a community may lead to the formation of culturally-specific graphical conventions over the course of several generations (17, 18). Recent laboratory studies of visual communication have found that pairs of interacting participants can rapidly learn to produce drawings that are referentially meaningful to their partner in context, even

Significance Statement

Drawing is a versatile tool for communication, spanning detailed renderings and simple sketches. Even the same object can be drawn in many ways, depending on the context. How do people decide how to draw in order to be understood? We collected a large number of drawings in different contexts and found that people adapted their drawings accordingly, producing detailed drawings when necessary, but simpler drawings when sufficient. To explain this contextual flexibility, we developed a computational model combining the capacity to perceive the correspondence between an object and drawing with the ability to infer what information is relevant to the viewer in context. Our results suggest drawing may be so versatile because of humans' joint capacity for visual abstraction and pragmatic inference.[mwu: do people read this on its own? if so, many terms here are defined in the text but not known to me aprior e.g. visual abstraction, pragmatic inference.]

J.E.F and R.X.D.H. designed and conducted human experiments, J.E.F, R.X.D.H, and M.W. analyzed data and performed computational modeling. J.E.F, R.X.D.H, M.W., and N.D.G. formulated models, interpreted results, and wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: jefan@stanford.edu

when these drawings do not strongly resemble any particular real-world referent out of context (19–22). Together, findings in this literature suggest the importance of social context — and over longer timescales, culturally shared conventions — in determining how drawings convey meaning.

Building on this prior work, the current study is guided by the overarching hypothesis that successful visual communication by drawing recruits pragmatic inference (23–26) — the ability to determine which information is not only *valid* to include in a drawing, but also *relevant* in context — in combination with visual abstraction — the set of computations that transform raw visual inputs into semantically meaningful internal representations.

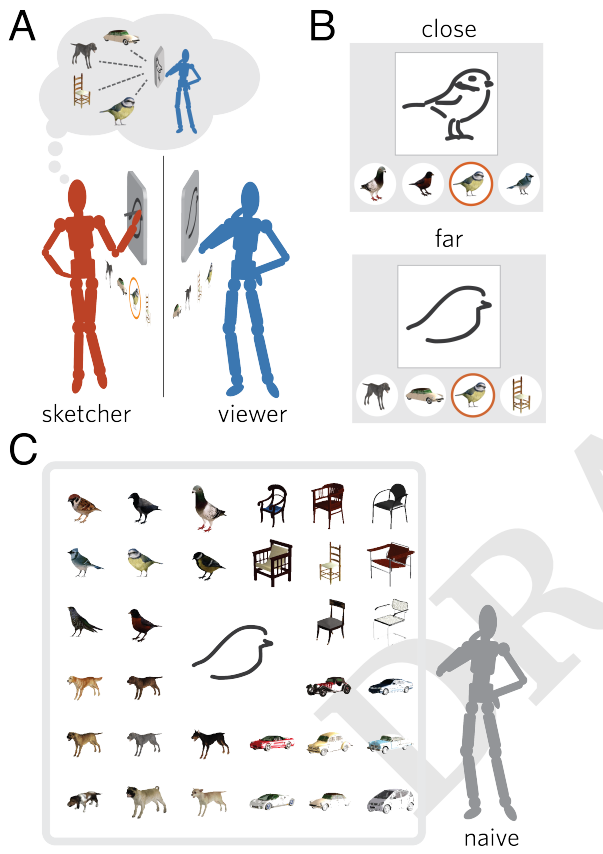

Fig. 1. (A) Communication task. Participants were paired in an online environment to play a sketching-based reference game and assigned the roles of *sketcher* and *viewer*. On each trial, the sketcher's goal was to draw one of these objects so that the viewer could distinguish the target from three distractor objects. (B) Context manipulation. Distractor similarity to target was manipulated across two context conditions: close contexts, where the target and distractors all belonged to the same basic-level category, and far contexts, where the target and distractors belonged to different basic-level categories. (C) Recognition task. Naive participants were presented with a randomly sampled sketch from the communication experiment and an array containing all 32 objects, and were instructed to identify the best-matching object.

To investigate visual communication in a naturalistic yet controlled setting, we employ sketching-based reference games. These reference games involve two players: a *sketcher* who aims to help a *viewer* pick out a target object from an array of distractor objects by representing it in a sketch. This basic arrangement can be traced back to the language games explored by (27) and (28). Such games have proven to be a valuable tool for eliciting pragmatic inferences about *language* use in context, and to make quantitative measurements of the behavioral consequences of these inferences (23, 29–31). Here we generalize this methodology to understand how sketchers account for information in common ground with their viewer in order to produce sketches that are informative (25, 26) yet parsimonious (32).

Traditionally, a barrier to progress in studying visual communication has been the lack of principled quantitative measures of high-level semantic information in drawings. As such, previous studies employing drawing tasks to study visual communication have typically relied on qualitative assessments of drawings based on provisional criteria specific to each study (22), or quantitative measures of low-level visual features that do not capture semantic information (19), limiting their ability to make detailed, quantitative predictions about visual communication behavior.

The goal of this paper is to provide a computational framework for systematically investigating how social context influences how people convey object identity in drawings. We present an integrated computational model of contextual flexibility in visual communication that combines deep convolutional neural network models of visual perception (11, 33) with a Bayesian probabilistic model of pragmatic reasoning (23) to make detailed predictions about what drawings people will produce across a variety of communicative contexts. We found that our full model fits the data well and outperformed lesioned variants.

Results

Effect of context manipulation on communication task performance. To evaluate the effect of context on visual communication behavior, we developed a paradigm in which participants (N=192) were paired in an online environment to play a sketching-based reference game. On each trial, both participants were shown an array containing the same four real-world objects, but object locations were randomized for each participant so that they could not use object location information to solve the task. Objects belonged to four basic-level categories (i.e., bird, car, chair, dog), and were rendered in the same three-quarter pose, under identical illumatination, and on a gray background, so participants could not use pose, illumination, or background information to distinguish them. On each trial, the sketcher's goal was to draw one of these objects — the target — so that the viewer could pick it out from a set of distractor objects. Across trials, the similarity of the distractors to the target was manipulated, yielding two types of communicative context: close contexts, where the target and distractors all belonged to the same basic-level category, and far contexts, where the target and distractors belonged to different basic-level categories. We predicted that while sketchers would be generally successful at conveying the identity of the target, their sketching behavior would systematically differ between the two contexts. Specifically, we predicted that sketchers would invest more time and ink in producing their sketches in close contexts, but still produce sufficiently informative sketches with less time and ink in far contexts.

Consistent with our prediction, we found that viewers were highly accurate overall at identifying the target from the sketches produced (proportion correct: 93.8%, 95% CI: [92.7%, 94.8%], estimated by bootstrap resampling participants). Moreover, we found that sketchers spent less time (close: 30.3s, far: 13.7s, $p<0.001$), applied fewer strokes (close: 8.03 vs. far: 13.5, 95% CI of difference: [3.75, 7.90], $p<0.001$), and used less ink (proportion of canvas filled; close: 0.054, far: 0.042, 95% CI of difference: [0.01, 0.014], $p<0.001$) to produce their sketches in the far condition than in the close condition. Despite the relative sparsity of sketches in the far condition, viewers were near ceiling at identifying the target on these trials (far: 99.7%, 95% CI: [0.993, 0.999]; close: 87.9%, 95% CI: [0.858, 0.899]), and took

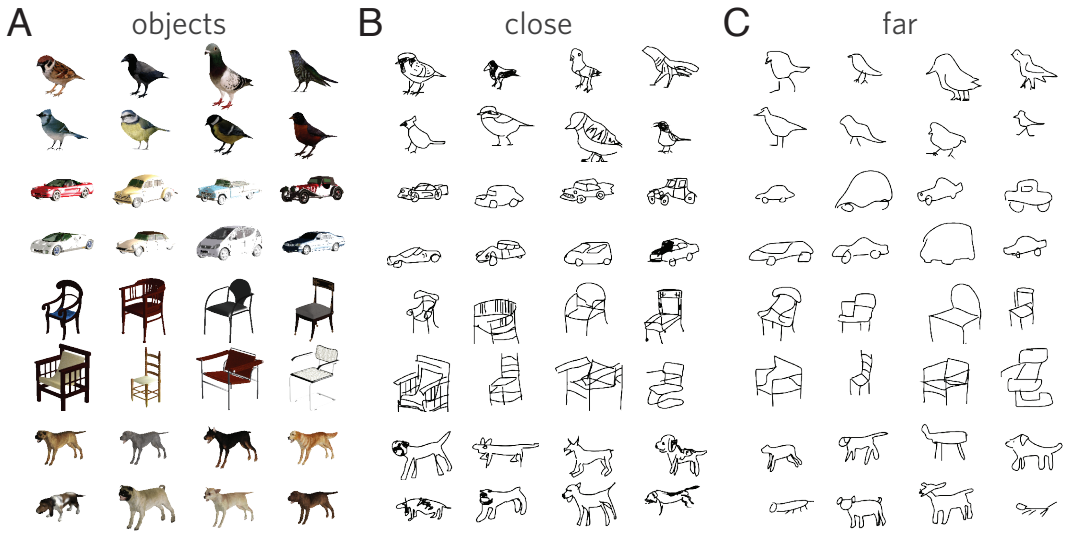

Fig. 2. (A) Object stimuli. (B) Example sketches produced in close condition. (C) Example sketches produced in far condition.

less time to make these decisions than on close trials (far: 6.32 sec vs. close: 8.32 sec, 95% CI: [-2.748, -1.251]).

Effect of context manipulation on sketch recognizability. A natural explanation for the difference between context conditions in how costly the sketches were to produce is that they differed in how informative they were about [mwu: the] identity of the target. Specifically, we hypothesized that the greater time and ink spent on close sketches was associated with a greater degree of perceptual correspondence to the target object, compared with less costly far sketches, which exhibited a weaker correspondence to the target object in absolute terms, while still being communicatively effective in context. To investigate this possibility, another group of naive participants (N=112) was recruited to perform a sketch-object matching task, the data from which were used to estimate the strength of perceptual correspondence between each sketch and every object in the experiment. On each trial of this recognition experiment, participants were presented with a randomly sampled sketch [mwu: sampled from where?] and an array containing all 32 objects, and were instructed to identify the object that best matched each sketch from the array. Consistent with the above account, we found that close sketches were matched with their corresponding object rendering more consistently than far sketches were (close: 54.2%; far: 37.5%; Z=14.1, p<0.001), although both were successfully matched at rates greatly exceeding chance (ps < 0.001).

Computational model of sketch production in context. Our empirical findings suggest that human sketchers spontaneously take information in common ground [mwu: could just be me, but what is common ground? define?] with the viewer into account, producing more informative sketches in close contexts at the cost of additional time and ink. Observing such contextual flexibility argues against the notion that sketch production is constrained exclusively by the appearance of the target of depiction. Rather, these results suggest that information in common ground with the viewer plays a critical role in determining how informative and costly of a sketch people decide to produce during visual communication. And they suggest an analogy to how context influences how people produce linguistic utterances during verbal communication, a key target of current computational theories of pragmatics in language use (30, 31, 34, 35). Informed by this prior work, we propose that human sketchers determine what kind of sketch to produce in context by deploying two main faculties: *visual abstraction*, which refers to the ability to judge how well a sketch evokes a real object, and *pragmatic inference*, which refers to the ability to prefer sketches that are sufficiently detailed to be diagnostic of the target object in context, but no more detailed than necessary. Intuitively, this latter faculty can be decomposed into two aspects: context sensitivity, a preference for sketches that are diagnostic of the target relative to the distractors; and cost sensitivity, a preference for less costly sketches. To test this proposal, we developed a computational model of the sketcher that embodied both visual abstraction and pragmatic inference, and was instantiated as a probabilistic program "wrapped" [mwu: "nested" might be a better word] around a deep convolutional neural network. Constructing such a model allowed us to evaluate the contribution of each component using formal model comparison, as well as quantitatively characterize the model's behavior in novel contexts.

[mwu: great writing!]

Defining communicative utility of sketches. We define a context, O, to be a set of objects containing a target, t, and three distractors, $D = \{d_1, d_2, d_3\}$:

$$O = \{t, D\} \quad [1]$$

We define the sketcher $\mathcal{S}$, to be a decision-theoretic agent that prefers to produce a sketch of the target, s_t, proportional to its communicative utility, U, where the utilities of each sketch have been normalized over the space of producible sketches via the softmax function (Eq. 2):

$$\mathcal{S}(s_t|t, D) \propto \frac{\exp[U(s_t, t, D)]}{\sum_{i=1}^{N} \exp[U(s_i, t, D)]} \quad [2]$$

[mwu: what is N? define. Why is this not an equals sign if we are normalized?]

We introduce three variants of the sketcher, whose decision-making behavior is determined by their respective utility functions. In each variant, this utility consists of two terms which trade off against one another: an informativity term and a cost term.

First, we define the utility function for a *pragmatic* sketcher that is sensitive to both context and cost, S_{2C}, which estimates a sketch's

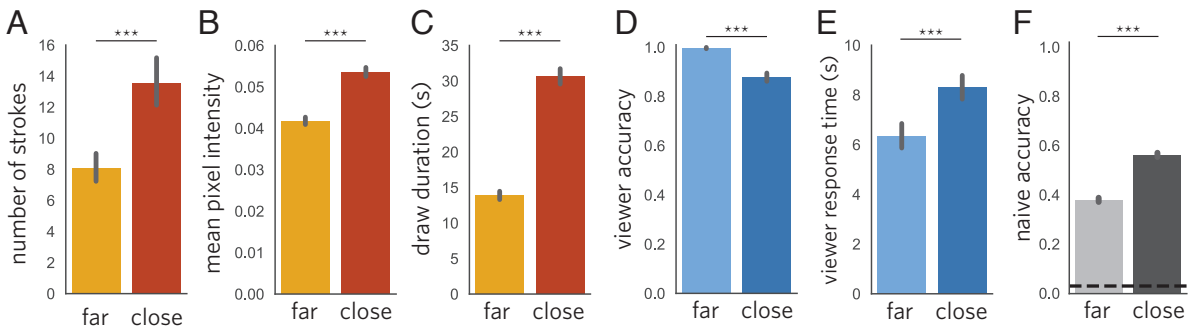

Fig. 3. (A-C) Sketchers used fewer strokes, less ink, and spent less time producing sketches in the far condition, relative to the close condition. (D-E) Viewers were at ceiling accuracy in identifying the target in the far condition, and were highly accurate in the close condition. (F) Naive matcher participants were more accurate for close sketches than far sketches.

informativity in terms of its *diagnosticity* in context. A sketch's diagnosticity is defined by the natural log probability that an imagined viewer, $\mathcal{V}$, would select the target object given the sketch and all objects in context, $\ln \mathcal{V}(t|s,D)$. [mwu: is $\mathcal{V}$ a function of S_{2C}? more rigorous def might be good. Also s is not defined (only s_t is).] A sketch's cost, $C(s)$, is defined to be a monotonic function of the amount of time taken to produce it [mwu: measured in what units?].

$$U_{S_{2C}}(s,t,D) = w_i \cdot \ln \mathcal{V}(t|s,D) - w_c \cdot C(s) \quad [3]$$

where w_i and w_c are independent scaling parameters that are applied to the diagnosticity and cost terms, respectively, and determine how strongly each term contributes to the overall utility of the sketch.

The imagined viewer $\mathcal{V}$, in turn, is assumed to select the target object proportional to the perceptual correspondence between the sketch and the target, $\text{sim}(s,t)$, normalized by the correspondences between the sketch and all four objects in context, again via the softmax function:

$$\mathcal{V}(t|s,D) \propto \frac{\exp\{\alpha \cdot \text{sim(s, t)}\}}{\sum_{i=1}^{4} \exp\{\text{sim}(s,o_i)\}} \quad [4]$$

where α is another scaling parameter determining the assumed optimality of the listener's decision policy: as $\alpha \to \infty$, the imagined listener is more likely to choose the object with highest perceptual correspondence to the sketch. Intuitively, this means that the viewer is more likely to pick the correct object when the sketch corresponds more strongly to the target than to the distractors. [mwu: you need to define o_i (a strict def of s, s_t, d and o might be useful). is α learned or set? if $\mathcal{V}$ is a viewer that only cares about perceptual correspondence, shouldnt α be as big as possible? why vary it? is the sum over 4 bc 4 objects in O?] [mwu: general q: why is S_{2C} context aware but $\mathcal{V}$ is not?]

Second, we define the utility function for a *context-insensitive* sketcher, S_{sim}, which replaces the diagnosticity term (defined in terms of the viewer) with a resemblance term reflecting the absolute degree of perceptual correspondence between a sketch and the target, $\text{sim}(s,o)$, and ignores the distractors. [mwu: mention we will define sim(s,o) later.]

Third, we define the utility function for a *cost-insensitive* sketcher, S_{nocost}, which sets the cost term to zero ($w_c = 0$), and thus captures the aim to produce sketches that would be informative to an imagined viewer in context without regard for how much time they would take to produce.

In our proposed sketcher model, $\mathcal{S}$, [mwu: so there are 4 types of sketchers in total? not 3?] we define the informativity term to be a mixture of the purely diagnosticity-based definition employed by S_{2C} and the purely resemblance-based definition employed by S_{sim}. To do this, we infer a mixture weight parameter, w_d, that interpolates between these two notions of informativity:

$$I(s,t,D) = w_d \cdot \ln \mathcal{V}(t|s,D) + (1 - w_d) \cdot \text{sim}(s,t). \quad [5]$$

Combining the utilities in this way captures the intuition that a communicative sketcher seeks to produce a sketch that both resembles the target object and distinguishes the target from the distractors.

After algebraically simplifying, the full utility is:

$$U(s,o) = w_i \cdot I(s,t,D) - w_c \cdot C(s) \quad [6]$$

[mwu: in total, what are the learnable parameters? (or free parameters)]

Defining perceptual correspondence between sketches and objects. In order for this sketcher to generate quantitative predictions, it needs to be able to compute the perceptual correspondence, $\text{sim}(s,o)$, for any sketch-object pair. We approached this problem in two ways: in our first set of modeling experiments, we employ human recognition behavior in the sketch-object matching task as an empirical approximation to $\text{sim}(s,o)$ for every sketch in our dataset (see Materials and Methods). In our second set of modeling experiments, we employ a visual encoder model adapted to predict $\text{sim}(s,o)$, which was trained in a crossvalidated fashion on empirical estimates of $\text{sim}(s,o)$.

Building on recent advances in computational vision (11, 12), we instantiated the visual encoder as a deep convolutional neural network (DCNN). This choice of model class is motivated by prior work showing that such networks, in addition to being a type of universal function approximator (36), learn higher-layer feature representations that capture high-level perceptual information in drawings (11), capture perceptual judgments of object shape similarity (37), and predict neural population responses in categories across the ventral visual stream (12) when trained on challenging natural object recognition tasks (38). Having an encoding model that operates directly on image inputs is important for a computational theory of visual communication because it allows our full model to generate predictions in novel contexts [mwu: might be worth to elaborate on what novel contexts entails].

Concretely, the visual encoder is a function that accepts a pair of images as input: a sketch, s, and an object rendering, o, and returns a scalar value reflecting the degree of perceptual correspondence between the sketch and object, $\text{sim}(s,o)$, which lies in the range $[0,1]$, where $\text{sim}(s,o) = 0$ reflects no correspondence and $\text{sim}(s,o) = 1$ reflects maximal correspondence (Eq. 7).

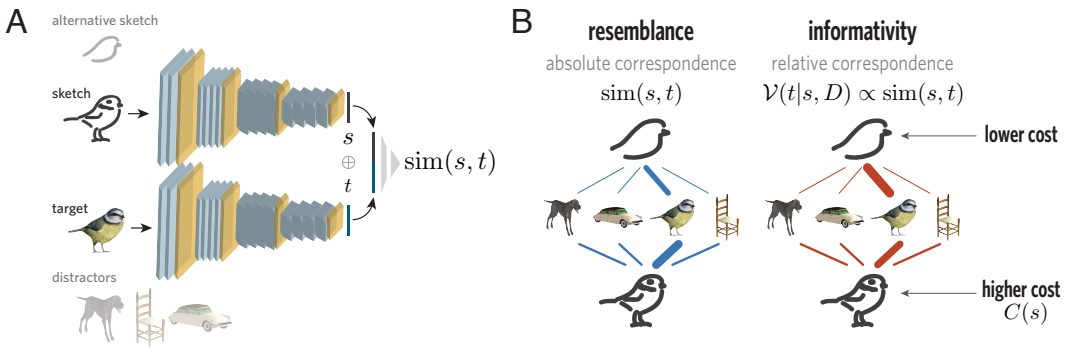

Fig. 4. (A) Architecture of visual encoder model for sketches and object renderings. Consists of pre-trained deep convolutional neural network (VGG-19) and shallow nonlinear "adaptor" neural network that predicts sketch-object correspondence. The encoder takes a sketch and object rendering as input, and outputs a correspondence score reflecting how well the sketch resembles the object. Three adaptor networks trained, using features from a one of the highest, an intermediate, and an early layer of VGG. Each adaptor network trained and evaluated in crossvalidated fashion. (B) Sketch-object correspondence computed for each object in context and each sketch in the test set. To determine the relative informativity of each sketch, the similarity scores between this sketch and all four objects were softmax normalized. The cost of each sketch was assumed to vary with the amount of time taken to produce it.

$$\mathrm{sim}(s,o) = A(B(s,o)) \qquad [7]$$

This encoder consists of two functional components: a base visual encoder network, B, and an adaptor network, A. We employ VGG-19 pretrained to recognize objects from the Imagenet database as our base visual encoder, whose parameters remain frozen (33). We then augment the pretrained feature representation in a higher layer of VGG-19 (i.e., the first fully-connected layer) with a shallow adaptor network, which is trained to predict the perceptual correspondence between specific sketch-object pairs. We train an adaptor network because while prior work has shown that the representation of object *categories* converges for sketches and photos at higher layers in DCNN models trained only on photos (11), additional supervision can substantially improve the accuracy of predictions involving comparisons between sketches and photos at the *instance* level (39). [mwu: this might not be important for this paper, but actually pretty interesting that we need nonlinearities in the adaptation, meaning that the spaces are not aligned in an affine sense... these two sentences about (1) we can generalize to sketches out of the box, and (2) we need additional transformations seem contradictory. How do we explain this?] To explore the degree to which the visual abstraction afforded by the complex transformations applied by successive layers of VGG-19 are required to capture human behavior during visual communication, we hypothesized that higher layers of these networks would provide a more transferable visual feature basis for modeling human-like visual abstraction in this task, relative to the lower-level image statistics represented in earlier layers.

Within this modeling framework, we conducted a series of targeted lesion studies to test each aspect of our hypothesis about the core cognitive faculties employed by humans to decide which sketch to produce in context. First, we evaluated the contribution of pragmatic inference by removing context sensitivity and cost sensitivity from the sketcher agent; these experiments relied upon empirical estimates of $\mathrm{sim}(s,o)$. Second, we evaluated the contribution of visual abstraction by comparing how well visual features adapted from different layers of a deep convolutional neural network [mwu: just say VGG-19?] predicted human visual communication behavior.

Evaluating contribution of pragmatic inference. We hypothesized that a *pragmatic* computational sketcher model that is sensitive to both context and cost would provide a strong fit to human sketch production behavior, as well as outperform lesioned alternatives lacking either component. To test this hypothesis, we evaluated three model variants characterized by their distinct definitions of communicative utility: a pragmatic one sensitive to both context and cost (S), one *not* sensitive to context but sensitive to cost (S_{sim}), and one that *is* sensitive to context but *not* sensitive to cost (S_{nocost}). In this set of model evaluations, we employed empirical estimates of the correspondence between sketches and objects, $\mathrm{sim}(s,o)$, to compute sketch diagnosticity in context. In this and subsequent model evaluations, we defined a sketch's cost, $C(s)$, to be a monotonic function of the amount of time taken to produce it.

Our goal was to ascertain how well each model could produce informative sketches and appropriately modulate its behavior according to the context condition, and not necessarily to reproduce exactly the same sketch as a particular participant had on a specific trial. As such, we aggregated all sketches of the same object produced in the same context condition when estimating their key functional properties (i.e., perceptual correspondence, cost), so each sketch was represented by the mean in its object-context category. Model predictions were generated to the same level of granularity, in the form of a probability distribution over 64 types of sketches, representing each combination of object and context condition (e.g., 'basset sketch produced in a close context').

To generate these predictions, first we employed Bayesian data analysis to infer a posterior distribution over the four latent scaling parameters in the model, applied to the informativity (w_i), cost (w_c), diagnosticity (w_d), and correspondence (α) terms. Next, we presented each model with exactly the same set of contexts that were presented to human sketchers in the communication experiment, and evaluated the posterior predictive probabilities that each model assigned to sketches in every object-context category, marginalizing over the posterior distribution over latent parameters. We conduct these evaluations on each of the test sets from the same five crossvalidation folds that were used to train and test the visual encoder, to permit direct comparison of these two sets of modeling results.

We found that the pragmatic sketcher, S, provided a much stronger overall fit to human behavior than the context-insensitive variant, S_{sim} (median Bayes Factor = 16.1; see Table 1), and the cost-insensitive variant, S_{nocost} (BF = 9.54).

To explore the behavioral patterns that may explain these differences in overall performance, we examined three aspects of each model's behavior: (a) *sketch retrieval*: its ability to assign a high absolute rank to the target sketch category in context, out of the 64 object-context alternatives; (b) *context congruity*: its ability to consistently prefer the context-congruent version of the target object

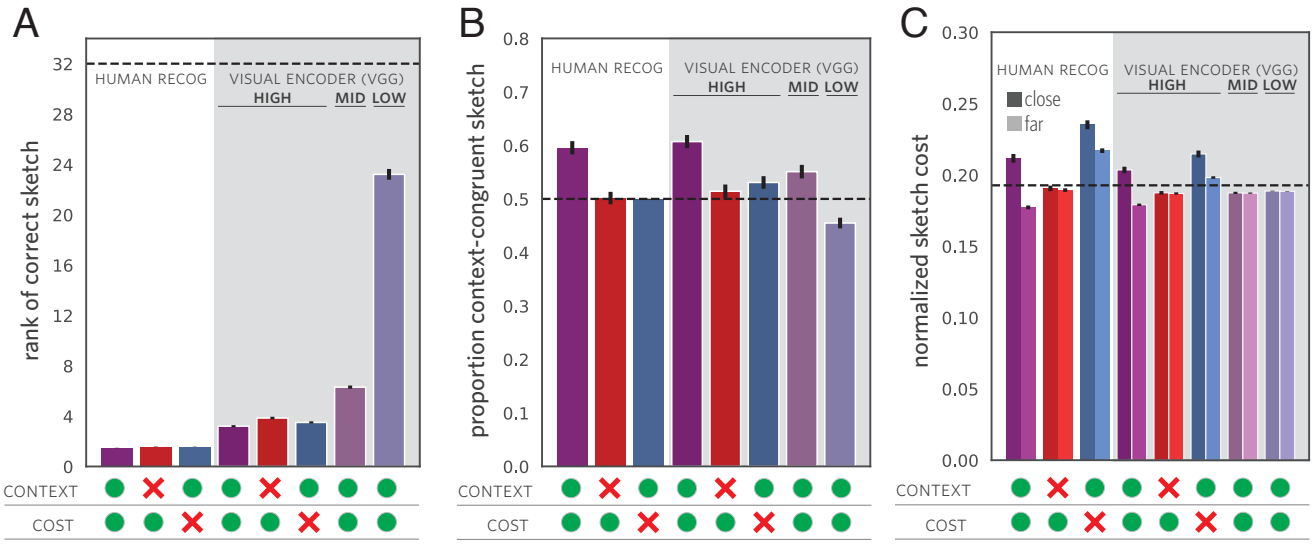

Fig. 5. Sketch production behavior by model variant. The table below each panel indicates whether the corresponding model variant plotted above is sensitive to context or cost. A green disc indicates context/cost sensitivity; a red 'X' indicates the lack of context/cost sensitivity. Results in lefthand region of each panel (white background) reflect model predictions when using empirical estimates of $sim(s,o)$ based on human sketch recognition behavior. Results in the righthand region (gray background) reflect model predictions when using variants of the DCNN visual encoder, the adaptor components of which were trained in a five-fold crossvalidated manner using human sketch recognition behavior. All results reflect average model behavior on test data only across identical train-test splits. Error bars represent 1 s.e. for this average estimate, found by applying inverse-variance weighting on individual confidence intervals from each train-test split. A: Rank of target sketch in list of 64 object-context categories, ordered by the probability assigned by each model. Dashed line reflects expected target rank under uniform guessing. Distribution of target rank scores across models suggest that high-quality estimates of $sim(s,o)$ are critical for strong performance. B: Proportion of trials on which each model assigned a higher rank to the context-congruent sketch of the correct object than the context-incongruent version of the correct object. Dashed line reflects indifference between the two versions of the sketch. Only models above this line show consistent and appropriate modulation of sketch producton by context. C: Relative time cost of sketches produced by each model, after applying min-max normalization to raw draw duration measurements. Predicted sketch cost on each trial computed by marginalizing over probabilities assigned to each sketch category. Darker bars reflect behavior in the close condition; lighter bars the far condition. Dashed line indicates the average cost of sketches in the full dataset; bars below this line reflect a preference for sketches that are less costly than average, bars above this line for sketches that are costlier than average. Only models that span this dashed line match the pattern of contextual modulation of sketch cost displayed by human sketchers.

over the context-incongruent version (i.e., on a close trial, preferring a close sketch to a far sketch); and (c) *cost modulation*: how consistently it produced costlier sketches than average in the close condition, and less costly sketches than average in the far condition, mirroring human behavior.

We found that in general, sketch retrieval performance was high for all three model variants (target rank 95% CI: pragmatic = $[1.43, 1.50]$, context-insensitive = $[1.54, 1.60]$, cost-insensitive = $[1.55, 1.60]$) (Fig. 5A, left).

However, only the pragmatic sketcher was able to reliably produce the sketch appropriate for the context condition more frequently than would be predicted by chance (95% CI proportion: $[0.571, 0.620]$; Fig. 5B, left); neither the context-insensitive nor the cost-insensitive variants displayed this context congruity (95% CI: context-insensitive = $[0.478, 0.525]$, cost-insensitive = $[0.498, 0.501]$). We observed that the lack of context congruity in the lesioned variants was attributable to an overall bias to produce close sketches in all contexts, which are highly informative in absolute terms, and thus higher in communicative utility if the distractors or sketch cost is ignored. [mwu: this is cool!]

Moreover, only the pragmatic sketcher produced costlier sketches than average in the close condition (95% CI normalized cost: $[0.205, 0.218]$ vs. grand mean cost = 0.196; Fig. 5C, left), and less costly sketches than average in the far condition (95%CI: $[0.175, 0.180]$). The context-insensitive variant is inherently unable to modulate the cost of the sketches it produces by context condition, and thus was no more or less likely to select a costlier, more diagnostic sketch on a close trial (95% CI: $[0.187, 0.194]$) than a far trial (95% CI: $[0.187, 0.192]$), and preferred slightly less costly sketches overall. While the cost-insensitive variant did exhibit cost modulation by context, because it ignores their cost, it preferred costlier sketches overall in both close (95%CI: $[0.229, 0.241]$) and far contexts (95%CI: $[0.214, 0.220]$).

Together, these results suggest that both context and cost sensitivity are critical for capturing key aspects of contextual flexibility in human visual communication.

Evaluating contribution of visual abstraction. Having established the importance of pragmatic inference, we next sought to evaluate the contribution of visual abstraction. To accomplish this, we replaced the empirical estimates of perceptual correspondence used by the model with predictions generated by three variants of a visual encoder, each of which adapted features from different layers of VGG: *high-level*, *mid-level*, and *low-level* (see Materials and Methods) .

Critically, all adaptor networks contained approximately the same number of learnable parameters, and were trained on the same data using exactly the same optimization procedure for an equal number of epochs. All model evaluations involving the base encoder and trained adaptor network were performed in a fivefold crossvalidated manner, with the full communication task dataset split into training, validation, and test sets in a 80%, 10%, and 10% ratio. We followed the same procedure as above to generate model predictions for the test set of each crossvalidation fold of our dataset, [mwu: reword: such that disjoint sets of sketches and communicative contexts were used to train each visual encoder and evaluate them.]

Consistent with our hypothesis, we found that a pragmatic sketcher model employing high-level features provided a substantially better fit to the data than one using mid-level features (high vs. mid median BF: 94.8), which in turn vastly outperformed a

model using low-level features from an early layer (mid vs. low BF: 257). These results suggest that making fuller use of the depth of VGG to compute the perceptual correspondence between a sketch and object yields a stronger basis for predicting and explaining human visual communication behavior.

Critically, using high-level features supported strong performance on sketch retrieval (95% CI target rank: $[3.03, 3.37]$, Fig. 5A), compared to mid-level features (target rank: $[6.05, 6.56]$) and low-level features (target rank: $[22.4, 24.1]$). These results show that without a high-performing visual encoder, the model is much less likely to produce sketches of the correct object, a basic prerequisite for successful visual communication even in the absence of contextual variability.

Moreover, the pragmatic sketcher model using high-level features also displayed context congruity (95% CI: $[0.583, 0.632]$, Fig. 5B), comparable in degree to the best-performing pragmatic model that lacked an encoder, showing that our full sketcher model was able to successfully reproduce this key signature of contextual flexibility for novel communicative contexts and sketches. The variant using mid-level features also displayed context congruity to a weaker extent (95% CI: $[0.526, 0.576]$), suggesting that an intermediate level of visual abstraction could achieve an intermediate degree of context congruity. By contrast, the variant using low-level features failed to prefer the context-congruent sketch category (95% CI: $[0.435, 0.475]$), providing a lower bound on the level of visual abstraction required in the underlying vision model to support flexible visual communication behavior.

Again, only the pragmatic sketcher model using high-level features displayed the same pattern of cost modulation as the best-performing pragmatic model lacking an encoder (95%CI: close = $[0.199, 0.208]$, far = $[0.178, 0.181]$, Fig. 5C), while both of the other variants using mid-level and low-level features failed to do so (95%CI: mid-level: close = $[0.186, 0.189]$, far = $[0.186, 0.188]$; low-level: close = $[0.188, 0.189]$, far = $[0.188, 0.189]$).

Having identified the best-performing visual encoder as the one that adapted features from a higher layer of VGG, we then performed the same context and cost sensitivity lesion experiments as before in order to evaluate the contribution of pragmatic inference in our full model. Again, we found that the pragmatic sketcher provided a stronger overall fit to human behavior than the context-insensitive variant (median Bayes Factor = 28.1; see Table 1), and a modestly better fit than the cost-insensitive variant (BF = 1.98). [mwu: do we have intuition as to why this one is only modestly better?] Critically, we found that removing context and cost sensitivity diminished the ability of this model to produce the context-congruent sketch of the correct object (context-insensitive 95% CI: $[0.489, 0.539]$; cost-insensitive 95% CI: $[0.507, 0.554]$; Fig. 5B), and appropriately modulate the cost of the sketches it produced (context-insensitive 95% CI: close = $[0.185, 0.190]$, far = $[0.185, 0.189]$; cost-insensitive 95% CI: close = $[0.210, 0.219]$, far = $[0.196, 0.200]$; Fig. 5C). By contrast, these lesions led to only modest decrements in overall sketch retrieval performance (95% CI target rank: context-insensitive = $[3.65, 4.05]$, cost-insensitive = $[3.33, 3.67]$; Fig. 5A), suggesting that the visual encoder itself is a major determinant of the ability to produce sketches of the correct object, whether or not the context-congruent version. These results converge with those of the lesion experiments conducted on the pragmatic sketcher model without a visual encoder, and together provide strong evidence for the importance of both visual abstraction and pragmatic inference for explaining contextual flexibility in human visual communication.

	human recog		visual encoder			
split	**context**	**cost**	**context**	**cost**	**mid**	**low**
1	18.0	11.9	44.5	2.70	105	282
2	8.46	9.89	20.9	-0.33	92.5	242
3	19.2	8.95	31.9	1.98	94.8	257
4	13.4	9.54	8.35	-0.67	93.4	248
5	16.1	7.92	28.1	5.99	114	269
median	**16.1**	**9.54**	**28.1**	**1.98**	**94.8**	**257**

Table 1. Bayes Factors for model comparisons between full and lesioned model variants (columns) for each crossvalidation fold (rows). Results under 'human recog' heading (first two columns) reflect model predictions when using empirical estimates of perceptual correspondence based on human sketch recognition behavior. Results under 'visual encoder' heading (final four columns) reflect model predictions when using variants of the DCNN visual encoder, the adaptor components of which were trained in a five-fold crossvalidated manner using human sketch recognition behavior. Other column labels indicate type of lesion: 'context' refers to comparison between full model and context-insensitive variant; 'cost' refers to comparison between full model and cost-insensitive variant; 'mid' refers to comparison between full model using the high adaptor vs. mid adaptor; 'low' refers comparison between full model using high adaptor vs. low adaptor. [mwu: do we ever define BayesFactor? maybe in appendix.. i dont actually know what this table is showing]

Discussion

The present study examined how communicative context influences visual communication behavior in a sketching-based reference game. We explored the hypothesis that people spontaneously account for information in common ground with their communication partner to produce drawings that are diagnostic of the target relative to the alternatives, while not being too costly to produce. We found that people spontaneously modulate how much time they invest in their drawings according to how similar the distractors are to the target, spending more time to produce more informative drawings when the alternatives were highly similar, but getting away with spending less time and producing less informative drawings when the alternatives were highly distinct. Observing such contextual flexibility suggests that while visual abstraction — the capacity to perceive the correspondence between an object and a drawing of it — is important for explaining why drawings of objects look the way they do, that visual production is not constrained exclusively by the perceptual properties of an object. Rather, our findings exposed an additional role for pragmatic inference — the ability to infer what information would be *relevant* to communicate, and not merely true. To test this hypothesis, we developed a computational model that embodied both pragmatic inference and visual abstraction, and found that it predicted human communication behavior well, and outperformed variants of the model lacking either component. Together, this paper provides a first algorithmically explicit theory of how perceptual and social cognition support contextual flexibility during visual communication.

This work generalizes the Rational Speech Act (RSA) modeling framework, originally developed to explain contextual effects in verbal communication (30, 31, 34, 35), to the domain of visual communication. RSA models take inspiration from the insights of Paul Grice (26), and incorporate ideas from decision theory, probabilistic models of cognition, bounded rationality, and linguistics, to understand how substantial variance in natural language use can be

explained by general principles of social cognition. They have been shown to capture key patterns of natural language use (30), achieve good quantitative fits with experimental data (29), and enhance the ability of artificial agents to produce informative language in reference game tasks (40, 41). In successfully adapting this modeling approach to the visual domain, our findings provide novel evidence for the possibility that similar cognitive mechanisms may underlie pragmatic behavior in both verbal and nonverbal communication modalities, a notion implicitly endorsed by prior work that has used nonverbal modalities (e.g., sketching, gesture) to investigate functional constraints on communication shared with language (19–21, 42–45). Moreover, they provide a principled strategy for understanding how variability across depictions of the same object can be derived from the task goals of the sketcher — in this case, to coordinate with a viewer on the same object of reference in context. In particular, future work exploring the joint consideration of communicative goals and perceptual representation during visual communication may help to provide mechanistic insight into the emergence of graphical conventions among communicators who build up common ground across repeated interactions.

There are several limitations of our model that would be fruitful to address in future work. First, obtaining a visual encoder that could produce accurate predictions of perceptual correspondence between sketch-object pairs required substantial supervision. While heavy supervision is not uncommon when developing neural network models of sketch representation (39, 46, 47), future work should investigate architectures that require weaker supervision to estimate image-level correspondences between sketches and natural photographs. One promising approach may be to exploit the hierarchical and compositional structure of natural objects (i.e., parts, subparts, and their relations), as they are expressed in both natural images and sketches of objects (48, 49). Second, our model produces a decision over which *type* of sketch to produce in context, rather than producing a *particular* sketch. This is of course different from the action selection problem human participants face — they must decide not only what stroke to make, but where to place them, how many, and in what order. While there have been recent and promising advances in modeling sketch production as a sequence of such actions (50–52), these approaches have not yet been shown to successfully emulate how people sketch real objects, much less how this behavior is modulated by communicative context. Future work should develop sketch production models that both operate on natural visual inputs and more closely approximates the actual dimensionality of the action space inherent to the task. Meeting these challenges is not only important for developing more human-like artificial intelligence, but may also shed new light on the nature of human visual abstraction, and how online perception and long-term conceptual knowledge guide decision making during complex, natural behaviors.

In the long term, elucidating the computational basis of visual communication may help to elucidate the sources of cultural variation in pictorial style, the origins of modern graphical techniques for conveying patterns in data, and lead to enhanced interactive visualization tools for education and research.

Code availability. The code for the analyses presented in this article is publicly available in a Github repository at: https://github.com/judithfan/visual_communication_in_context.

Data availability. The data presented in this article are publicly available in a figshare repository.

Materials and Methods

Communication experiment: Manipulation of context in sketch-based reference game. A total of 192 unique participants, who were recruited via Amazon Mechanical Turk (AMT) and grouped into pairs, completed the experiment. They were provided a base compensation of $2.70 for participation and earned a $0.03 bonus for each correct trial. In this and subsequent behavioral experiments, participants provided informed consent in accordance with the Stanford IRB. Stimuli were thirty two 3D mesh models of objects belonging to 4 categories (i.e., birds, chairs, cars, dogs), containing eight objects each. Each object was rendered in color on a gray background at three-quarter perspective, 10° viewing angle (i.e., slightly above), and fixed distance. [mwu: is there any significance to these choices? why did we choose these?] Sketchers drew using black ink on digital canvas (pen width = 5 pixels; 300 x 300 pixels) embedded in a web browser window using Paper.js (http://paperjs.org/). Participants drew using the mouse cursor, and were not able to delete previous strokes. Each stroke of which was rendered on the viewer's screen immediately upon the completion of each stroke. There were no restrictions on how long participants could take to make their drawings. After clicking a submit button, the viewer guessed the identity of the drawn object by clicking one of the four objects in the array. Otherwise, the viewer had no other means of communicating with the sketcher. Both participants received immediate task-related feedback: the sketcher learned which object the viewer had clicked, and the viewer learned the identity of the target. Both participants earned bonus points for each correct response. For each pair [mwu: im lost; is this describing how pairs were chosen or what the viewer sees? what is a quartet?], objects were randomly allocated to eight quartets: Four of these quartets contained objects from the same category ("close"); the other four of these quartets contained objects from different categories ("far" condition). Each quartet was presented four times, such that each object in the quartet served as the target exactly once. The assignment of objects to quartet and condition was randomized across pairs.

Recognition experiment: Measuring perceptual similarity between sketches and objects. A total of 112 participants were recruited via Amazon Mechanical Turk (AMT). They were provided a base compensation of $1.00 for their participation, and earned an additional $0.01 bonus for each correct response. On each trial, participants were presented with a randomly selected sketch collected in the communication experiment, surrounded by a grid containing the 32 objects from that experiment. Their goal was to select the object in the grid that best matched the sketch. Participants received task feedback in the form of a bonus earned for each correct trial. Participants were instructed to prioritize accuracy over speed. We applied a conservative outlier removal procedure based on response latency, whereby trials that were either too fast to have supported careful consideration of the sketch and menu of objects (RT<1000*ms*), or too slow and suggestive of an attentional lapse (RT>30*s*), were filtered from the dataset. The removal of these outlier trials (8.01%) did not have a substantial impact on the pattern of recognition behavior. In order to mitigate the possibility that participants could adjust their matching strategy according to any particular sketcher's style, each session was populated with 64 sketches sampled randomly from different reference games. To obtain robust estimates of sketch-object perceptual correspondences, each sketch was presented approximately 10 times across different sessions.

Computational modeling.

Sketch data preprocessing. To train and evaluate our sketcher model, we first filter the sketch dataset to retain only sketches that were correctly identified by the viewer during the communication task (6.2% incorrect) and were compliant with task instructions by not including 'drawn' text annotations (4.4% non-compliant). This filtered sketch dataset was then split into training, validation, and test sets in a 80%, 10%, and 10% ratio, and this split was performed in a 5-fold cross-validated manner. Splits were based on context, defined as the set containing a specific target object and three distractor objects, such that no context appeared both in the training and test splits of any cross-validation fold. Specifically, we ensured that: (1) the number of sketches from each category (i.e. car) and (2) the proportion of sketches from close and far trials were equated across splits. This was done to control for biases in model peformance due to imbalances in the training or test set.

Deriving empirical estimates of perceptual correspondence between sketches and objects. In the recognition experiment, most sketches were not matched exclusively to a single object, but to several. These sketches can thus

be thought of having some degree of perceptual correspondence to the several objects it was matched to at least once. For a single sketch, we estimate the perceptual correspondence between that sketch and any object as the proportion of recognition task trials on which it was matched to that object. For a set of sketches of a specific object produced in a specific context condition, we estimate the *aggregated* sketch-object correspondence to be the proportion of recognition task trials on which any sketch from this set was matched to that object. Because our goal was to understand how well each model could produce produce informative sketches according to the context condition, and not necessarily to reproduce exactly the same sketch as a particular participant had on a specific trial, we use this aggregate correspondence measure in all of our modeling experiments. As a result, sketch-object correspondence scores lie in the range $[0,1]$, and sum to 1 for sketches in the same object-context category. Because all sketches from the same object-context category share the same correspondence to each object, there are a total of 32 sketch categories x 32 objects x 2 contexts = 2048 empirical perceptual correspondence scores.

Deriving empirical estimates of sketch costs. We reasoned that drawing time would be a natural proxy for the cost incurred by workers on Amazon Mechanical Turk, who increase their total compensation by completing tasks in a timely manner. However, as there were no absolute constraints on the amount of time that could be spent on each trial, there was considerable variability across different participants in terms of how much time they spent producing their sketches. To control for this variability across participants and to ensure robust estimates, we first removed outliers (draw times exceeding 5 s.d. from the mean), then z-score normalized drawing times across all remaining trials within a participant, and finally averaged these normalized draw times across sketches within the same object-context category as above, yielding 32 objects x 2 contexts = 64 empirical cost estimates in total.

Visual encoder architecture. The visual encoder is a function that accepts a pair of images, a sketch and an object rendering (both 224 x 224 RGB images), as input and returns a scalar value in the range $[0,1]$, reflecting the degree of perceptual correspondence between the sketch and object.

[mwu: this paragraph is repetitive with main text] The encoder consists of two components: a base visual encoder and an adaptor network. We employ VGG-19 (33) as our base visual encoder architecture, which had been pretrained to categorize objects on the ImageNet database, and whose parameters remain frozen (38). We augment VGG-19 with a shallow fully-connected *adaptor* network that is trained to adapt the generic visual feature representation computed by VGG-19 to predict the perceptual correspondence between individual sketches and objects. Here only the parameters of this adaptor network are trained and we do not finetune the base visual encoder. We compare three adaptor networks that intercept VGG-19 image representations at different layers: the first max pooling layer (early), the tenth convolutional layer (mid), and the first fully connected layer (high). To facilitate comparison between adaptor networks, we ensured that each of the three contain a comparable number of trainable parameters (number of learnable parameters for high: 1048839; mid: 1049115; low: 1048833) with identical training hyperparameters (i.e., learning rate, batch size, etc.). To discriminate which layer provides the best starting feature basis for predicting sketch-object correspondence, these adaptor networks were also deliberately constrained to be shallow, i.e., consisting only of two linear layers with an intervening point-wise nonlinearity.

When applying the high-level visual encoder, a sketch and object are first passed through VGG and a feature vector in $\mathbb{R}^{4096}$ for each image is extracted from a higher layer (i.e., the first fully-connected layer, also known as *fc6*). These two vectors are then concatenated to form a single vector in $\mathbb{R}^{8192}$, to be passed into the high adaptor network. The high adaptor is composed of a linear layer that maps from $\mathbb{R}^{8192} \rightarrow \mathbb{R}^{128}$, followed by a "Swish" nonlinearity * and dropout, then another linear layer that maps from $\mathbb{R}^{128} \rightarrow \mathbb{R}^{1}$. [mwu: maybe mention that we added dropout bc it helps not overfit, leading to improved generalization. Btw, "high", "middle", "low" are not the same notation we used above. we should be consistent.]

When applying the mid-level visual encoder, two mid-level feature representations are intercepted from an intermediate layer (i.e., the 10th convolutional layer, *conv_4_2*). Features from the 10th convolutional layer of VGG are three-dimensional: 512 by 28 by 28 pixels. We first "flatten" the input into a one dimensional vector in $\mathbb{R}^{512}$ using a weighted linear combination over the spatial dimensions $\sum_{i=1}^{28}\sum_{j=1}^{28} w_{ij} * x_{ij}$, where x_{ij} indexes a spatial location in the image representation at this layer. The weight parameters $\{w_{ij} | 1 \leq i,j \leq 28\}$ are learned jointly with the parameters of mid adaptor, and are separately learned for the sketch and 3D object modalities. [mwu: might want to mention that this process is soft-attention over the spatial features + citation?] The two vectors in $\mathbb{R}^{512}$ are then concatenated to form a single vector in $\mathbb{R}^{1024}$, to be passed to the mid adaptor network. The mid adaptor consists of a linear layer that maps from $\mathbb{R}^{1024} \rightarrow \mathbb{R}^{1021}$, followed by a Swish nonlinearity, dropout, then a linear layer from $\mathbb{R}^{1021} \rightarrow \mathbb{R}^{1}$. This hidden layer size (i.e., 1021 units) was chosen to ensure that the total number of learnable parameters was as similar to the high adaptor as possible.

The architecture of the low adaptor is almost identical to that of the mid adaptor, except VGG features are intercepted at the first max pooling layer (i.e., *pool1*). Features in this layer are 64 by 112 by 112 pixels. As above, a weighted sum of model activations over the 112 by 112 spatial features is applied, yielding one vector in $\mathbb{R}^{64}$ each for the sketch and object, which are then concatenated. The low adaptor maps from $\mathbb{R}^{128} \rightarrow \mathbb{R}^{7875}$, followed by the same Swish nonlinearity, dropout, and another linear layer that maps from $\mathbb{R}^{7875} \rightarrow \mathbb{R}^{1}$. As above, the hidden layer size (i.e., 7875 units) was chosen so that this adaptor contained a comparable number of learnable parameters.

[mwu: text in the last two paragraphs needs tightening.]

Visual encoder training. We trained each adaptor (i.e., high, mid, low) to predict, for each sketch, a 32-dimensional vector that captures the *pattern* of perceptual correspondences between that sketch and all 32 objects. Each encoder accepts a sketch-object pair as input and returns a real number as output (in range $(-\infty,+\infty)$), reflecting their perceptual correspondence. We iterate over all objects in the stimulus set I to generate the predicted 32-vector for each sketch, and then apply softmax normalization, yielding a vector that sums to 1. We define the loss function, $\mathcal{L}$, to be the cross entropy loss between the predicted distribution, q and the empirically estimated perceptual correspondence vector, p (which also sums to 1):

$$\mathcal{L} = \sum_{x \in I} p(x) \log q(x) \qquad [8]$$

This loss function explicitly encourages the adaptor to learn not only to predict the strength of the correspondence between a sketch and the object it was intended to depict (measured by correct matches during recognition), but also to predict its correspondence to all of the other objects (measured by the pattern of confusions during recognition).

We use the Adam optimization algorithm (54) (learning rate = 1e-4) over minibatches of size 10 for 100 epochs, where an epoch is a full pass through the training set. After training each adaptor for 100 epochs, we freeze the model with the best performance on a validation set. †

Generating encoder-based estimates of perceptual correspondence between sketches and objects. To generate sketch-object correspondence scores for sketches in each test split, we first pass each sketch-object pair into a visual encoder, yielding a single image-level correspondence score lying in the range $(-\infty,+\infty)$. To map these raw image-level scores to the appropriate range for a correspondence score ($[0,1]$), we first z-score them ($f(x) = \frac{x-\bar{x}}{s}$), then apply the logistic function ($f(x) = \frac{1}{1+e^{-x}}$). These normalized image-level correspondence scores are then averaged across all sketches belonging to the same object-context category, yielding 32 objects x 32 sketches x 2 contexts = 2048 model-based perceptual correspondence scores for each visual encoder variant (i.e., high, mid, low).

Model comparison. In order to test the contribution of each component of our sketcher model, we conducted a series of lesion experiments and formal model comparisons. To quantify the evidence for one model over another, we computed Bayes Factors: the ratio of likelihoods for each model [mwu: ah... this is the def i was looking for.], integrating over all their respective parameters under the prior:

$$BF = \frac{\int P(D|M_1,\theta_1)P(\theta_1)}{\int P(D|M_2,\theta_2)P(\theta_2)}$$

Unlike classical likelihood ratio tests, which use the maximum likelihood, the Bayes Factor naturally penalizes models for their complexity (55, 56). We placed uninformative uniform priors over all five parameters required to

*Swish is a recently discovered nonlinearity that outperforms the common rectified linear nonlinearity (ReLU) in deep models on a suite of tasks (53). Regardless, the ability of the model to fit the data appeared to be robust to the choice of nonlinearity (alternative nonlinearities were TanH and ReLU) and size of hidden layer, within the range $(2^7, 2^9)$).

†As a property of the input domain, the gradients with respect to adaptor parameters are very small (1.51e-4 $\pm$ 2.61e-4), inevitably resulting in poor learning (we can reproduce this effect from several intializations). We find that naively increasing the learning rate led to unstable optimization, but that multiplying the loss by a large constant C leads to a much smoother learning trajectories and good test generalization. Critically, increasing the learning rate and multiplying the loss by a constant are not equivalent for second moment gradient methods. In practice, $C = 1e4$.

specify our models: a discrete choice over alternative approaches to computing perceptual correspondance:

$$m \sim \text{Unif}\{\text{“empirical”}, \text{“high”}, \text{“mid”}, \text{“low”}\}$$

and over the continuous latent parameters,

$$w_i, w_c, w_d, \alpha \sim \text{Unif}(0, 50).$$

To compute the likelihood function $P(D|M, \theta)$ for a speaker model M under parameters θ, we perform exact inference for our sketcher model using (nested) enumeration and sum over all test set datapoints within a crossvalidation fold.

Specifically, we compute the exact likelihood at every point on a discrete grid of parameters. This is of particular interest for nested model comparisons, e.g. comparing our full model to a context-insensitive variant. Rather than computing the full marginalized likelihood for both models, we can use the Savage-Dickey method (57) to simply compare the posterior probability against the prior at the nested point of interest (e.g. $w_c = 0$) for the full model.

To evaluate the contribution of pragmatic inference, we begin by comparing the pragmatic sketcher model using empirically estimated perceptual correspondences to nested “cost-insensitive” ($w_c = 0$) and “context-insensitive” ($w_d = 0$) variants. To evaluate the contribution of visual abstraction, we then proceed to compare the three visual encoder variants that adapt features from different layers of VGG-19, marginalizing over all other parameters. Finally, we perform the same context and cost lesion experiments on the full model, employing best-performing visual encoder (i.e., “high”).

Evaluating model predictions. We implemented our models and conducted inference in the probabilistic programming language WebPPL (58). We use MCMC to draw 1000 samples from the joint posterior with a lag of 0, discarding 3000 burn-in samples. We constructed posterior predictive distributions by computing each measure of interest (i.e., target rank, context congruity, sketch cost) over the test data set, for every MCMC sample. To estimate standard errors on predictions across models, we employed a multi-stage bootstrapping procedure to account for three nested sources of variation: variation across trials within a test split, variation across the parameter posterior within a test split, and variation across test splits. Specifically, for each model variant and for each test split we bootstrap resampled trials with replacement from the test dataset 1000 times to estimate the mean and standard error on each measure of interest, marginalizing over MCMC samples from the parameter posterior. We applied inverse-variance weighting to aggregate these estimates of the mean and standard error across test splits, such that test splits with lower variance contribute more than do splits with higher variance, which yielded a single overall estimate of the mean and standard error on each measure of interest, for each model variant. We estimated the half-widths of the 95% confidence interval for each measure of interest under the the assumption of normality for the sampling distribution of the mean.

ACKNOWLEDGMENTS. Thanks to Dan Yamins and the Stanford CoCo Lab for helpful comments and discussion.

1. Hoffmann D, et al. (2018) U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. *Science* 359(6378):912–915.
2. Aubert M, et al. (2014) Pleistocene cave art from Sulawesi, Indonesia. *Nature* 514(7521):223–227.
3. Donald M (1991) *Origins of the modern mind: Three stages in the evolution of culture and cognition.* (Harvard University Press).
4. Tomasello M (2009) *The cultural origins of human cognition.* (Harvard university press).
5. Card M (1999) *Readings in information visualization: using vision to think.* (Morgan Kaufmann).
6. Clottes J (2008) *Cave Art.* (Phaidon London).
7. Gombrich E (1989) *The story of art.* (Phaidon Press, Ltd.).
8. Kellogg R (1969) *Analyzing children's art.* (National Press Books Palo Alto, CA).
9. Sayim B (2011) What line drawings reveal about the visual brain. pp. 1–4.
10. Gibson JJ (2014) *The ecological approach to visual perception: classic edition.* (Psychology Press).
11. Fan JE, Yamins DLK, Turk-Browne NB (2018) Common object representations for visual production and recognition. *Cognitive Science* 0(0).
12. Yamins DL, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* 111(23):8619–8624.
13. Hochberg J, Brooks V (1962) Pictorial recognition as an unlearned ability: A study of one child's performance. *the american Journal of Psychology* pp. 624–628.
14. Kennedy JM, Ross AS (1975) Outline picture perception by the songe of papua. *Perception* 4(4):391–406.
15. Tanaka M (2007) Recognition of pictorial representations by chimpanzees (pan troglodytes). *Animal cognition* 10(2):169–179.
16. Jongejan J, Rowley H, Kawashima T, Kim J, Fox-Gieg N (2017) ‘google quickdraw’.
17. Toku M (2001) Cross-cultural analysis of artistic development: Drawing by japanese and us children. *Visual Arts Research* 27(1):46–59.
18. Boltz WG (1994) *The origin and early development of the Chinese writing system.* (Eisenbrauns) Vol. 78.
19. Garrod S, Fay N, Lee J, Oberlander J, MacLeod T (2007) Foundations of representation: where might graphical symbol systems come from? *Cognitive science* 31(6):961–987.
20. Fay N, Garrod S, Roberts L, Swoboda N (2010) The interactive evolution of human communication systems. *Cognitive Science* 34(3):351–386.
21. Galantucci B (2005) An experimental study of the emergence of human communication systems. *Cognitive Science* 29(5):737–767.
22. Healey P, Swoboda N, Umata I, King J (2007) Graphical language games: Interactional constraints on representational form. *Cognitive Science.*
23. Goodman ND, Frank MC (2016) Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20(11):818–829.
24. Clark H (1996) Using language. new york, ny, us.
25. Wilson D, Sperber D (1986) *Relevance: Communication and cognition.* (Mass.).
26. Grice HP, Cole P, Morgan JL (1975) Syntax and semantics.
27. Wittgenstein L (1953) *Philosophical investigations.* (Macmillan).
28. Lewis D (1969) *Convention: A philosophical study.* (Harvard University Press).
29. Kao J, Bergen L, Goodman N (2014) Formalizing the pragmatics of metaphor understanding in *Proceedings of the annual meeting of the Cognitive Science Society.* Vol. 36.
30. Goodman ND, Stuhlmüller A (2013) Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science* 5(1):173–184.
31. Frank MC, Goodman ND (2012) Predicting pragmatic reasoning in language games. *Science* 336(6084):998–998.
32. Zipf GK (1936) *The psycho-biology of language: An introduction to dynamic philology.* (Routledge).
33. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*
34. Franke M, Jäger G (2016) Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft* 35(1):3–44.
35. Bergen L, Levy R, Goodman N (2016) Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9.
36. Hornik K (1991) Approximation capabilities of multilayer feedforward networks. *Neural networks* 4(2):251–257.
37. Kubilius J, Bracci S, de Beeck HPO (2016) Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology* 12(4):e1004896.
38. Deng J, et al. (2009) Imagenet: A large-scale hierarchical image database in *Computer Vision and Pattern Recognition, 2009.* (IEEE), pp. 248–255.
39. Sangkloy P, Burnell N, Ham C, Hays J (2016) The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)* 35(4):119.
40. Monroe W, Hawkins RX, Goodman ND, Potts C (2017) Colors in context: A pragmatic neural model for grounded language understanding. *arXiv preprint arXiv:1703.10186.*
41. Cohn-Gordon R, Goodman ND, Potts C (2018) Pragmatically informative image captioning with character-level inference in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers).* pp. 439–443.
42. Goldin-Meadow S, Feldman H (1977) The development of language-like communication without a language model. *Science* 197(4301):401–403.
43. Theisen CA, Oberlander J, Kirby S (2010) Systematicity and arbitrariness in novel communication systems. *Interaction Studies* 11(1):14–32.
44. Garrod S, Fay N, Rogers S, Walker B, Swoboda N (2010) Can iterated learning explain the emergence of graphical symbols? *Interaction Studies* 11(1):33–50.
45. Verhoef T, Kirby S, De Boer B (2014) Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics* 43:57–68.
46. Yu Q, et al. (2017) Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision* 122(3):411–425.
47. Song J, Yu Q, Song YZ, Xiang T, Hospedales TM (2017) Deep spatial-semantic attention for fine-grained sketch-based image retrieval. in *ICCV.* pp. 5552–5561.
48. Battaglia P, Pascanu R, Lai M, Rezende DJ, , et al. (2016) Interaction networks for learning about objects, relations and physics in *Advances in Neural Information Processing Systems.* pp. 4502–4510.
49. Mrowca D, et al. (2018) Flexible neural representation for physics prediction in *Advances in Neural Information Processing Systems.*
50. Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–1338.
51. Ha D, Eck D (2017) A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477.*
52. Ganin Y, Kulkarni T, Babuschkin I, Eslami S, Vinyals O (2018) Synthesizing programs for images using reinforced adversarial learning. *arXiv preprint arXiv:1804.01118.*
53. Ramachandran P, Zoph B, Le QV (2018) Searching for activation functions.
54. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*
55. Wagenmakers EJ, et al. (2018) Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review* 25(1):35–57.
56. Jefferys WH, Berger JO (1992) Ockham's razor and bayesian analysis. *American Scientist* 80(1):64–72.
57. Wagenmakers EJ, Lodewyckx T, Kuriyal H, Grasman R (2010) Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive psychology* 60(3):158–189.
58. Goodman ND, Stuhlmüller A (2014) The design and implementation of probabilistic programming languages.