# *What's behind the box? Measuring scene context with Shannon's guessing game on indoor scenes.*

*Greene, Oliva, Wolfe, and Torralba*

## Abstract

Natural scenes are lawful, predictable entities: objects do not float unsupported, spoons are more often found with forks than printers, and it makes little sense to search for toilets in dining rooms. Although visual context has often been manipulated in object and scene recognition studies, it has not yet been formally measured. Information theory specifies how much information is required to encode objects in a scene, assuming no contextual knowledge. We can then measure, in bits per object, the information benefit provided by human observers' contextual knowledge. We used a database of 100 indoor scenes, containing 352 unique objects labeled using the LabelMe tool. If all objects were equally probable in a scene, 8.46 bits per object would be required (log2(352)). Taking object frequency into account (i.e., chairs are more common than basketballs), would only reduce this number to 7.22 bits per object. To measure the information required by humans to represent objects in scenes, we adapted the guessing game proposed by Shannon (1951). Between 5-80% of the objects in each scene were occluded by opaque bounding boxes. Observers guessed the identity of each occluded object until the object was correctly named. More than 60% of objects were correctly guessed on the first try, because context massively constrains the identity of a hidden object (What might you guess was hidden next to a plate on a dinner table?). Fully 93% of objects were correctly guessed within 10 tries. Overall, we found that observers could represent the database with just 1.86 bits per object when 5-10% of objects were masked. Just 2.00 bits per object were needed even when the majority of objects were masked. This technique can be used to measure the redundancy provided by aspects of context such as scene category, object density and object consistency.