Examples of recognition problems as large multi-class classification.

Multi-class classification is often applied to problems with large and complex label sets.
These large multi-class problems are encountered when there exists a large domain of objects or categories to be recognized from data.
Leading examples include detecting the speaker from his voice patterns \citep{togneri2011overview},
identifying the author from her written text \citep{stamatatos2014overview}, or labeling the object category from its image \citep{duygulu2002object,deng2010does,oquab2014learning}.

One practical challenge in developing recognition systems for large label sets is collecting high quality training data.

When developing a specific system, it is more affordable to collect data on a small set of classes first, even if the long-term goal is for the system to generalize to large sets.
Even when the data is easy to collect, the development stage can be accelerated by training first on small sets of classes.
Indeed, comparisons on the generalizability of small-set performance to larger-set performance can found in the literature (\cite{oquab2014learning}, \cite{griffin2007caltech}).
A natural question, then, is how changing the number of possible labels affects the classification accuracy?

Technically, consider a pair of classification problems on finite label subsets

\|$mathcal{S}_k| < |\mathcal{S}_K|, \mathcal{S}_i \subset \mathcal{Y}$, where in the $k$-th task, one constructs the classification rule $h^{(k)}:\mathcal{X} \to \mathcal{S}_k$. Supposing that $(X, Y)$ have a joint distribution, define the generalization accuracy for the $k$-th problem as
\begin{equation}\label{eq:ga_k}
\text{GA}_k = \Pr[h^{(k)}(X) = Y|Y \in \mathcal{S}_k].
\end{equation}
The problem of \emph{performance extrapolation} is the following: using data
from only $\mathcal{S}_k$, can one predict the accuracy
for the larger label set $\mathcal{S}_K$?

A natural use case for this prediction would be in the development of a facial recognition system.

The system was developed in lab A on their database of $k_1$ individuals. The client would then deploy it on a new, perhaps larger, set of $k_2$ individuals. Performance extrapolation could allow the lab to tell the client how well their algorithm will perform.

Extrapolation should be possible when the smaller and larger classifications are two representations of the same recognition problem.

In many recognition problems, the set of categories is to some degree a random or arbitrary selection out of a much larger, perhaps infinite, set of potential categories. Yet in any specific experiment they are a fixed finite set.

For example, categories in the classical cal-tech 256 image recognition dataset \citep{griffin2007caltech} were assembled by aggregating keywords proposed by students and collecting matching images.

The arbitrary nature of the label set is even more apparent in biometric applications (face recognition, authorship, fingerprint identification) where the labels correspond to human individuals (\cite{togneri2011overview}, \cite{stamatatos2014overview}).

In these cases, the number of the labels used to define a concrete dataset is therefore an experimental choice rather than a property of the domain.

Nevertheless, such datasets are still viewed as representing the larger problem of recognition within the given domain, and we think success on such a dataset is reflective of performance on similar problems within the domain, even when the label sets may differ.

We propose to view the labels sets as randomly sampled from some population.

Not only does the assumption of randomness capture the ambiguity of label sets used in recognition problems, but it also provides a powerful formalism for answering the question of how to extrapolate.

While many sampling mechanisms are possible, in this paper we assume that both $\mathcal{S}_K$ and $\mathcal{S}_k$ are iid samples
from a population (or distribution) of labels.

The condition of i.i.d. sampling of labels ensures that the separation of labels in a random set $\mathcal{S}_K$ can be inferred by
looking at the empirical separation in $\mathcal{S}_k$, and therefore that some estimate of the achievable accuracy on \$mathcal{S}_K$ can be obtained.

We define average generalization accuracy and consider how to infer it

Our analysis considers a restricted set of classifiers,

These two assumptions (iid and marginal) allow us to give a nice characterization\

Our main contributions:

Theory to understand extrapolation

    - There is one function D which links performance on all the label set sizes
    - We also develop an estimation method based on this theory
- Outline of paper
    - Related work (rest of this section)
    - Framework (section 2) and toy example
    - Theory to understand extrapolation (section 3)
    - There is one function D which links performance on all the label set sizes
    - We also develop an estimation method based on this theory (section 4)
    - And we demonstrate the method on a facial recognition example based on OpenFace (section 5)

- Put in related work
    - Example 2: A neuroscientist is interested in how well the brain activity in various regions of the brain can discriminate between different classes of stimuli. \cite{Kay2008a} obtained fMRI brain scans which record how a single subject's visual cortex responds to natural images. They wanted to know how well the brain signals could discriminate between different images. For a set of 1750 photographs, they constructed a classifier which achieved over 0.75 accuracy of classification. Based on exponential extrapolation, they estimate that it would take on the order of $10^{9.5}$ classes before the accuracy of the model drops below 0.10! A theory of performance extrapolation could be useful for the purpose of making such extrapolations in a more principled way.