2. Overview of ML approaches to modeling cognitive neuroscience data

Analyzing biological and artificial neural networks: challenges with opportunities for synergy? Barrett et al. (2019)

There are **analogies** between deep neural networks and neuroscience. Both fields need to:
- understand how neural networks transform representations of stimuli to implement complex computations;
- describe and analyze very high dimensional data.

Analogies appear in **four areas.**

Receptive fields
Neurons in visual cortex start by focusing on tiny parts of what we see. As information goes deeper into the brain, these cells expand their focus to see larger areas and recognize more complex things. **Repetition priming** means that brain remembers seeing something, making it easier to recognize again, even if it looks a little different.
AI researchers think DNN neurons specialize. They study this by seeing what images activate these neurons most and how their receptive fields change in deeper layers, particularly how these larger fields become more complex.

Retinotopy map
The occipital cortex (visual cortex) processes images, coding information based on eccentricity (distance from the center of vision) and radial degree (angle). This coding can be mapped using fMRI during simple experiments. For example, by having participants focus on a shrinking/expanding or rotating shape, researchers can identify neurons that fire at specific times, indicating their sensitivity to certain eccentricities.

Ablation
By mapping lesions to symptoms, we can understand which brain areas are important for given tasks. Pruning and ablation cause performance deficits. However, thanks to **neuroplasticity**, the human brain can achieve again the same performance.
The ablation analysis is applicable to **DNNs**. We can silence neurons and observe how this impacts the network output. Networks trained for generalization are more robust to ablation than those trained on memorizing labels.

Dimensionality reduction
The brain processes information in a widespread way, using many neurons. Dimensionality reduction is used to simplify understanding these complex brain codes.
There are three main reasons why this is useful: Redundancy (different neurons do similar jobs), Distributed coding (information is spread out among many neurons), Correlated activity (neurons are active in similar patterns).
DNNs can compress complex image data (e.g., from 4096 to 512 dimensions) by over 80% without losing much information. This shows that only a few key "dimensions" or features are needed to tell images apart.

Representational Geometry
To understand how brains and AI systems process information, we study their representational geometries (how they represent different things).

Matrix Factorization measures
Compares how two different systems (like two AI networks) represent objects using **object-by-feature matrices**. Techniques like Canonical Correlation Analysis help find common patterns or correlations between these representations.

Representational Similarities Analysis
Instead of breaking down matrices, RSA compares how objects are grouped or clustered together in different systems. It use **object-by-object distance matrices**, showing how similar or different objects are perceived.

Linear Regression
Helps to understand if AI models are learning similar ways of representing the world as humans.

What does the mind learn? A comparison of human and machine learning representations Spicer and Sanborn (2019)

This paper reviews the modern machine learning techniques and their use in models of human mental representations, detailing three notable **branches**.

Spatial Methods
Placing items in a multidimensional space and using their location to draw conclusions about categorization.

Prototype approaches
Assume that learning is based on similarity to the center of a category (mean), which is stored after training.

Exemplar approaches
Calculate similarity as a ratio between the similarity of an item to all items within a class and the similarity of that item to all other items.

Clustering
organizes items into groups, with quality often quantified by the distance between items within and between clusters.

Logical Methods
Concepts are based on a definition that is applied to the features of the object. One viable solution is searching for rules that maximize discrimination between stimuli.

Artificial Neural Networks ANNs
Don't make assumptions about the representations involved, but offer an implementation method.

Thoughts
Instead of solely asking which model is most accurate, we should prioritize how a model helps us understand the **underlying representations**. We need to consider not only accuracy but also **confusion** (how errors occur).