4. Modeling conceptual organization

- ***Deep Neural Networks Predict Category Typicality Ratings for Images***
  Lake et al. (2015)
  They evaluate deep convolutional networks trained for classification on their ability to predict category typicality (human typicality ratings), and try to understand whether deep learning systems can serve as potential cognitive models.
  - **Problem**
    The core problem explored is how typicality influences various cognitive tasks like categorization speed, learning, and inductive inference. The researchers wanted to see if CNNs, known for learning categorization, also develop internal representations that reflect human notions of category typicality.
  - **Methodologies**
    The study involved two main parts.
    - **Behavioral Experiment**
      Human participants rated images from 8 categories based on how well each image fit their idea of the category.
    - **Computational Experiment**
      Three different CNN architectures were used. The researchers devised two ways to estimate typicality from the CNNs' internal workings
      - **Raw Typicality**
        This measured how strongly a specific category was activated by an image.
      - **Contrast Typicality**
        This assessed how much more active the correct category was compared to others for a given image. The CNNs were trained to classify images, and then their typicality scores were correlated with the human ratings. For hidden layers, typicality was redefined as the cosine distance between an image's activation vector and a "category prototype" (average activation for that category).
  - **Results**
    - Both raw and contrast typicality scores from CNNs similarly predicted human ratings.
    - Some CNNs significantly correlated with human typicality, suggesting they learn graded categories aligned with human perception.
    - Deeper convolutional layers showed better prediction of human typicality, indicating that representations closer to the output layer better reflect human judgments.

- ***Predicting Human Brain Activity Associated with the Meanings of Nouns***
  Mitchell et al. (2008)
  Words-as-features as models of cognition. Can feature models explain behavioral and brain responses?
  The underlying idea is to understand how/where word meaning is stored in the brain, under the assumption that our brain represents words as features.
  - **Problem**
    The main question was whether feature models can explain brain activity related to word meanings. This tackles core neuroscience questions about whether neural activity differs for different concepts, if concept representation is localized or distributed, and how consistent it is across people.
  - **Methodologies**
    The researchers proposed a generative encoding model to predict brain activity from word meaning
    - **Defining Word Meaning**
      They described 60 nouns using 25 semantic features, which were defined by how often a noun co-occurs with specific verbs in a large text collection.
    - **Predicting Brain Activity**
      For each voxel they learned a relationship between its activity and these 25 semantic features across the 60 nouns (**single voxel analysis**). This allowed to predict the fMRI activation (brain activity) for any given word by multiplying its semantic features by these learned weights for each voxel.
    - **Testing Generalization**
      To show the model could generalize, they trained it on 58 words and tested its ability to predict activity for the remaining 2, checking if the predicted activity matched the actual activity better for the correct word than for the incorrect one.
  - **Results**
    - The model achieved 79% accuracy in predicting word activity, suggesting that word meanings are indeed represented as features in the brain.
    - They could map which brain areas were most important for specific semantic features.
    - Even though individual differences exist, there are words that maximally activate specific brain regions.
    - Using randomly selected features to define word meaning resulted in significantly worse.

- ***Matching categorical object representations in inferior temporal cortex of man and monkey***
  Kriegeskorte et al. (2008)
  Studying representations via similarity spaces. How to model similarity spaces expressed in human behavior and brain responses.
  - **Problem**
    The core problem is to understand and compare how various systems (human brain, human behavior, computational models) represent categorical object information. The goal is to see if their "similarity spaces" (how similar or different they perceive pairs of stimuli to be) align.
  - **Methodologies**
    Representational Similarity Analysis (**RSA**) is a method for comparing how different "things" (like human behavior, brain activity, or computational models) represent information RSA is a versatile technique used to tackle this type of problems.
    - **Creating Representational Dissimilarity Matrices (RDMs)**
      For any set of stimuli (e.g., images of objects), an RDM is built. Each cell in the RDM represents how "dissimilar" two stimuli are.
      - **Human Behavior RDMs:** These are created by directly asking people for similarity judgments or by observing how they sort objects.
      - **Brain Data RDMs:** These are built by measuring brain activity (e.g., using fMRI) for different stimuli.
      - **Computational Model RDMs:** These are generated by feeding stimuli into a model (like a CNN) and measuring the "dissimilarity" between their internal representations (features).
    - **Comparing RDMs (Second-Order Similarity)**
      Compares RDMs from different sources (e.g., human brain vs. human behavior). If two RDMs are very similar, it means the underlying systems represent the stimuli in a similar way.
    - **Modality Independence**
      RSA is powerful because it can compare completely different types of data.
    - **Analyzing Hidden Information**
      RSA can be used to see how representations change over time or across different brain regions. It also helps to understand if information is carried by patterns across multiple brain regions (**multivariate analysis**) rather than just in individual regions (**univariate analysis**).
  - **Results**
    - A multivariate approach, considering patterns across multiple voxels, can uncover more information than looking at individual voxels alone.
    - While RSA effectively compares "similarity spaces," it loses information about the specific underlying features that create that similarity, which is its main limitation.