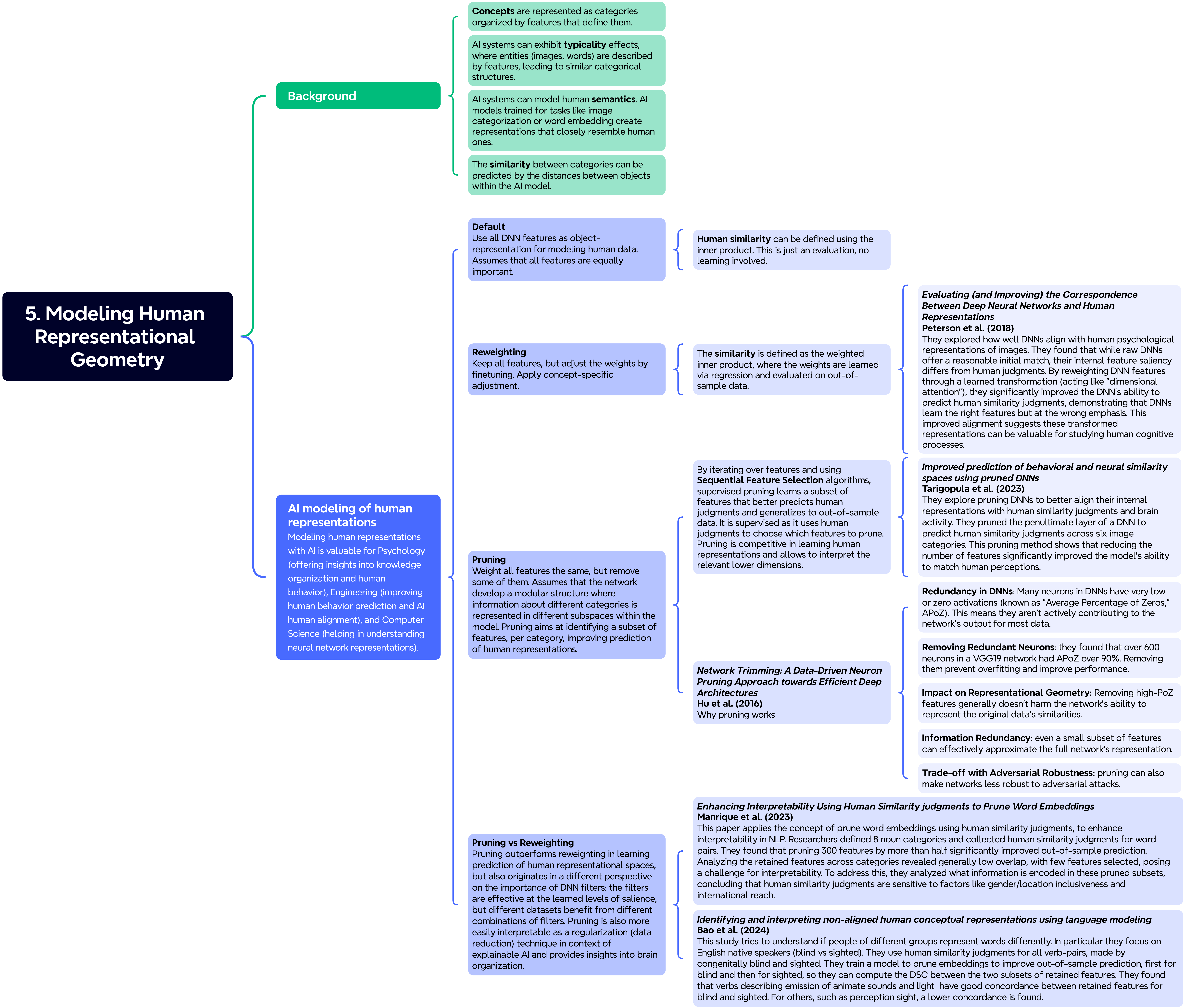

5. Modeling Human Representational Geometry

Background

- **Concepts** are represented as categories organized by features that define them.
- AI systems can exhibit **typicality** effects, where entities (images, words) are described by features, leading to similar categorical structures.
- AI systems can model human **semantics**. AI models trained for tasks like image categorization or word embedding create representations that closely resemble human ones.
- The **similarity** between categories can be predicted by the distances between objects within the AI model.

AI modeling of human representations

Modeling human representations with AI is valuable for Psychology (offering insights into knowledge organization and human behavior), Engineering (improving human behavior prediction and AI human alignment), and Computer Science (helping in understanding neural network representations).

Default
Use all DNN features as object-representation for modeling human data. Assumes that all features are equally important.

- **Human similarity** can be defined using the inner product. This is just an evaluation, no learning involved.

Reweighting
Keep all features, but adjust the weights by finetuning. Apply concept-specific adjustment.

- The **similarity** is defined as the weighted inner product, where the weights are learned via regression and evaluated on out-of-sample data.
  - ***Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations***
    Peterson et al. (2018)
    They explored how well DNNs align with human psychological representations of images. They found that while raw DNNs offer a reasonable initial match, their internal feature saliency differs from human judgments. By reweighting DNN features through a learned transformation (acting like "dimensional attention"), they significantly improved the DNN's ability to predict human similarity judgments, demonstrating that DNNs learn the right features but at the wrong emphasis. This improved alignment suggests these transformed representations can be valuable for studying human cognitive processes.

Pruning
Weight all features the same, but remove some of them. Assumes that the network develop a modular structure where information about different categories is represented in different subspaces within the model. Pruning aims at identifying a subset of features, per category, improving prediction of human representations.

- By iterating over features and using **Sequential Feature Selection** algorithms, supervised pruning learns a subset of features that better predicts human judgments and generalizes to out-of-sample data. It is supervised as it uses human judgments to choose which features to prune. Pruning is competitive in learning human representations and allows to interpret the relevant lower dimensions.
  - ***Improved prediction of behavioral and neural similarity spaces using pruned DNNs***
    Tarigopula et al. (2023)
    They explore pruning DNNs to better align their internal representations with human similarity judgments and brain activity. They pruned the penultimate layer of a DNN to predict human similarity judgments across six image categories. This pruning method shows that reducing the number of features significantly improved the model's ability to match human perceptions.
- ***Network Trimming: A Data-Driven Neuron Pruning Approach towards Efficient Deep Architectures***
  Hu et al. (2016)
  Why pruning works
  - **Redundancy in DNNs**: Many neurons in DNNs have very low or zero activations (known as "Average Percentage of Zeros," APoZ). This means they aren't actively contributing to the network's output for most data.
  - **Removing Redundant Neurons**: they found that over 600 neurons in a VGG19 network had APoZ over 90%. Removing them prevent overfitting and improve performance.
  - **Impact on Representational Geometry:** Removing high-PoZ features generally doesn't harm the network's ability to represent the original data's similarities.
  - **Information Redundancy:** even a small subset of features can effectively approximate the full network's representation.
  - **Trade-off with Adversarial Robustness:** pruning can also make networks less robust to adversarial attacks.

Pruning vs Reweighting
Pruning outperforms reweighting in learning prediction of human representational spaces, but also originates in a different perspective on the importance of DNN filters: the filters are effective at the learned levels of salience, but different datasets benefit from different combinations of filters. Pruning is also more easily interpretable as a regularization (data reduction) technique in context of explainable AI and provides insights into brain organization.

- ***Enhancing Interpretability Using Human Similarity judgments to Prune Word Embeddings***
  Manrique et al. (2023)
  This paper applies the concept of prune word embeddings using human similarity judgments, to enhance interpretability in NLP. Researchers defined 8 noun categories and collected human similarity judgments for word pairs. They found that pruning 300 features by more than half significantly improved out-of-sample prediction. Analyzing the retained features across categories revealed generally low overlap, with few features selected, posing a challenge for interpretability. To address this, they analyzed what information is encoded in these pruned subsets, concluding that human similarity judgments are sensitive to factors like gender/location inclusiveness and international reach.
- ***Identifying and interpreting non-aligned human conceptual representations using language modeling***
  Bao et al. (2024)
  This study tries to understand if people of different groups represent words differently. In particular they focus on English native speakers (blind vs sighted). They use human similarity judgments for all verb-pairs, made by congenitally blind and sighted. They train a model to prune embeddings to improve out-of-sample prediction, first for blind and then for sighted, so they can compute the DSC between the two subsets of retained features. They found that verbs describing emission of animate sounds and light have good concordance between retained features for blind and sighted. For others, such as perception sight, a lower concordance is found.