AI in Open Worlds: A Progress Report

Tom Dietterich

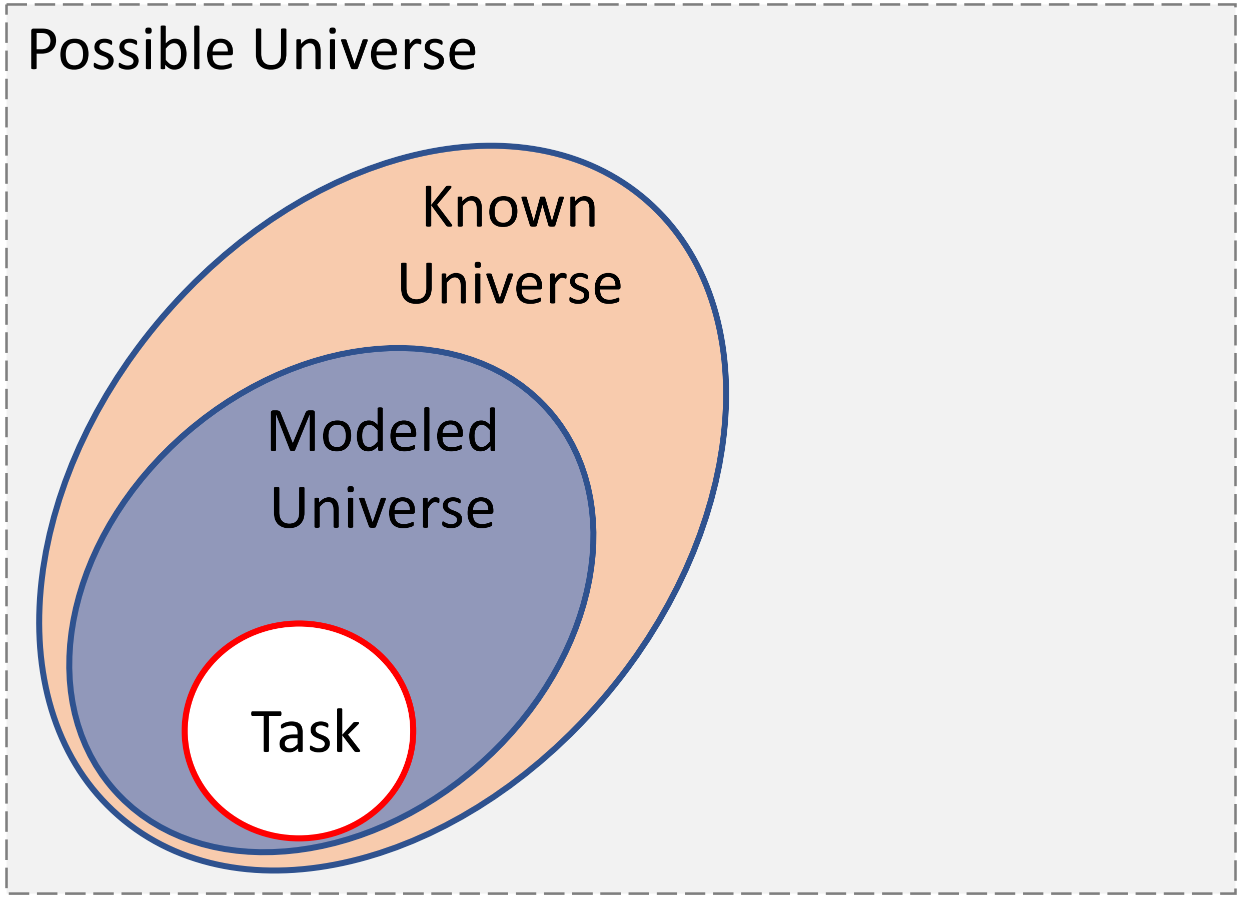
Distinguished Professor Emeritus

Oregon State University

AI in Open Worlds

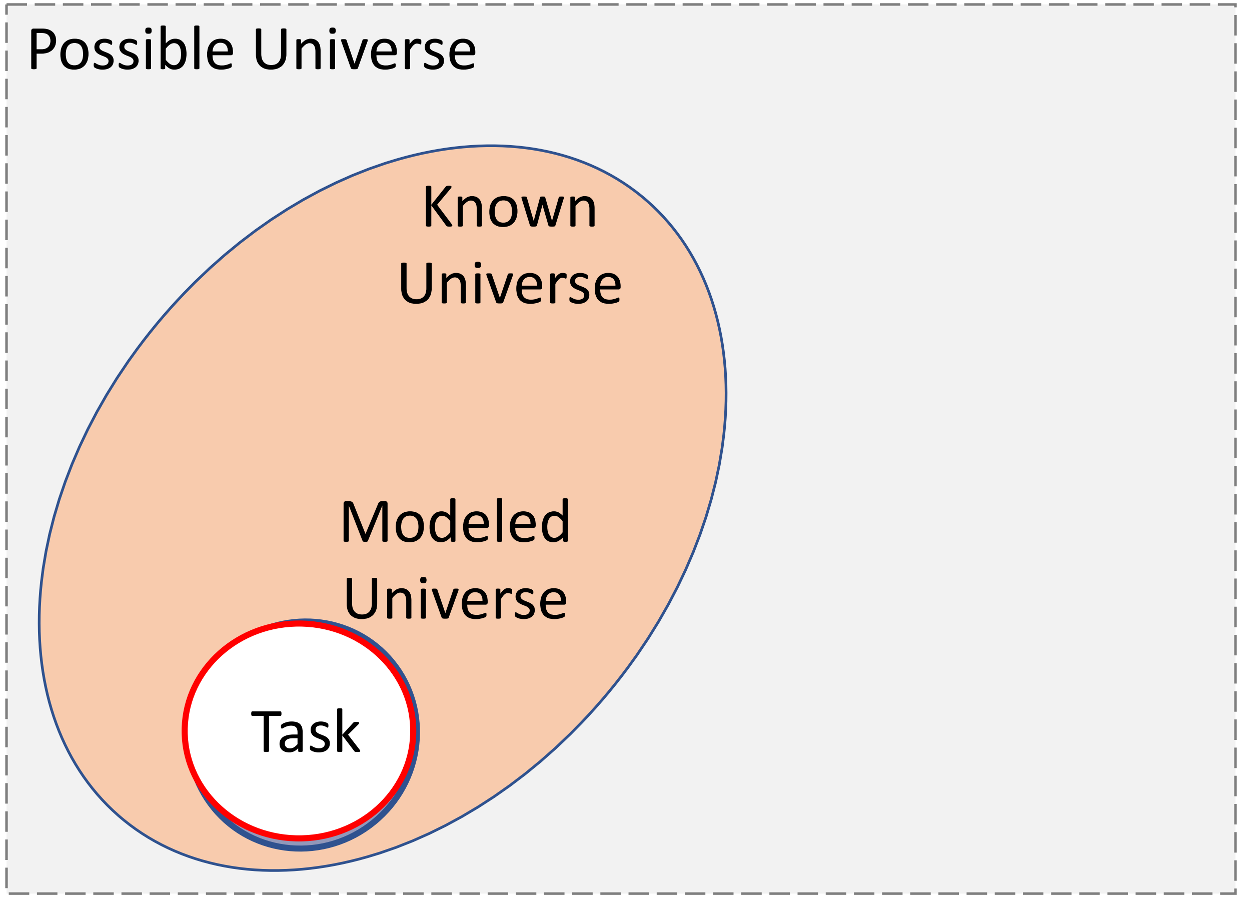
- Eric Horvitz: "Artificial Intelligence in the Open World"
  - (AAAI Presidential Address 2008)
  - The open world is a complex world
  - Requires combining sensing, learning, and reasoning
- Tom Dietterich: "Steps Toward Robust AI"
  - (AAAI Presidential Address 2016)
  - The open world contains unknown failure modes, novel categories and behaviors
  - Methods:
    - Robust optimization for known and unknown model failures
    - Risk-sensitive objectives
    - Anomaly detection
- Today's talk: What we've learned since 2016
  - ML for safety-critical applications
  - Deep anomaly detection
  - Near miss detection
- Future Directions
  - Distribution-Independent Machine Learning

The Open World



- Possible Universe: (presumably unbounded) space of additional possibilities
- Known Universe: Space that is known to the designer
- Modeled Universe: Space that is representable by the system's ontology/features
- Task: Problem space needed to perform the task

Narrow AI Systems



- Modeled Universe = Task Problem Space
- Representation can only capture the task problem space
- Reasoning is only performed over the task problem space

Closed-World Design

- Closed task description
  - Closed set of diseases and symptoms
  - Fixed goal language (e.g., PDDL over fixed ontology)
- Optimize a problem solver
  - Machine learning approach
    - Collect data
    - Train a classifier or a decision making policy
  - Planning/reasoning approach
    - Customize a general inference engine
    - Optimize heuristics to guide the reasoning

Example:
Automated Counting of Freshwater Macroinvertebrates

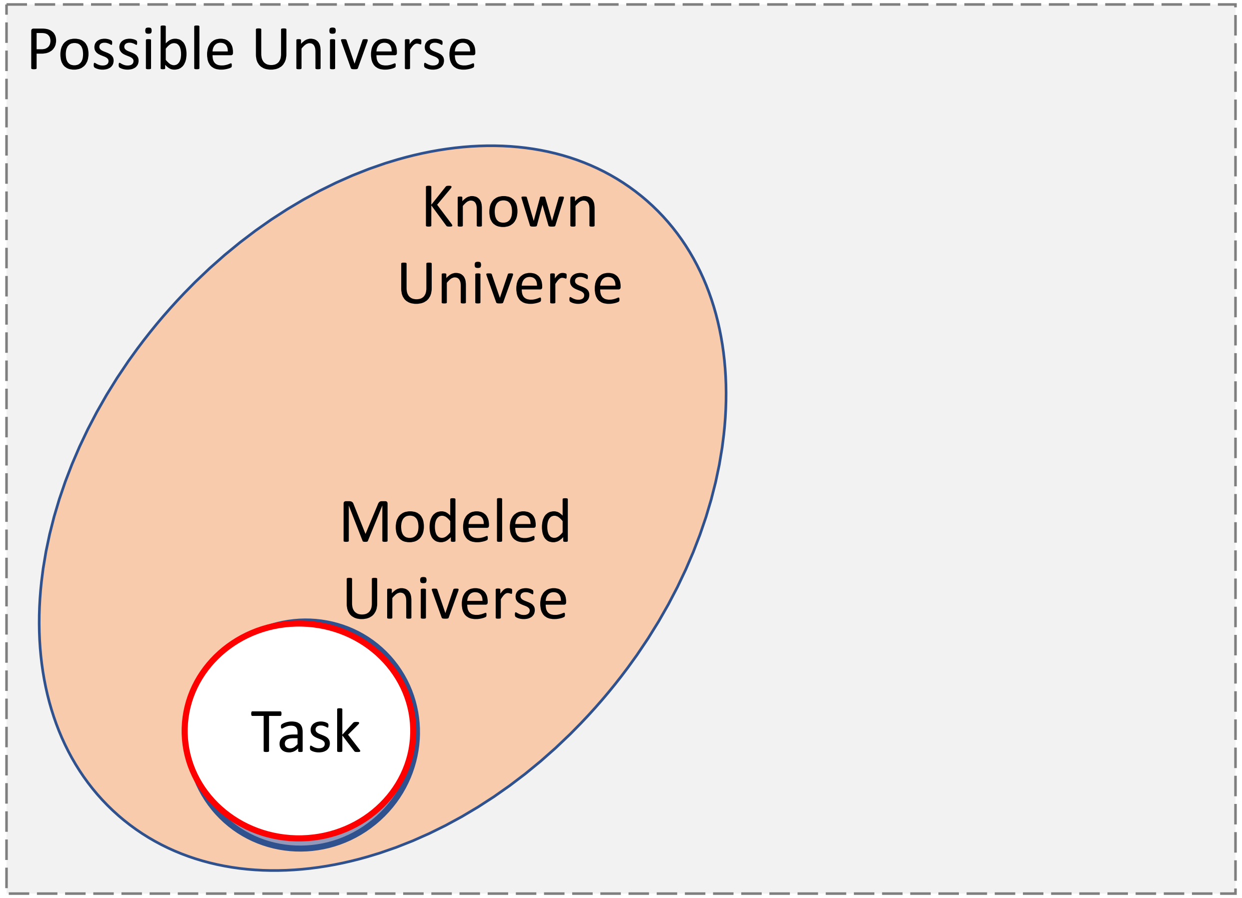
- Goal: Assess the health of freshwater streams
- Method:
  - Collect specimens via kicknet
  - Photograph in the lab
  - Classify to genus and species
  - Histogram of species tells us what pollutants have been in the water

www.epa.gov

Data Collection and Training

- Entomologists collected 100 specimens each from 54 taxa
- Trained a computer vision classifier
  - accuracy $\approx 90\%$
  - Larios, N., Soran, B., Shapiro, L., Martínez-Muños, G., Lin, J., Dietterich, T. G. (2010). **Haar Random Forest Features and SVM Spatial Matching Kernel for Stonefly Species Identification.** *IEEE International Conference on Pattern Recognition (ICPR-2010)*.
  - Lin, J., Larios, N., Lytle, D., Moldenke, A., Paasch, R., Shapiro, L., Todorovic, S., Dietterich, T. (2011). **Fine-Grained Recognition for Arthropod Field Surveys: Three Image Collections**. *First Workshop on Fine-Grained Visual Categorization (CVPR-2011)*
  - Lytle, D. A., Martínez-Muñoz, G., Zhang, W., Larios, N., Shapiro, L., Paasch, R., Moldenke, A., Mortensen, E. A., Todorovic, S., Dietterich, T. G. (2010). **Automated processing and identification of benthic invertebrate samples.** *Journal of the North American Benthological Society*, 29(3), 867-874.

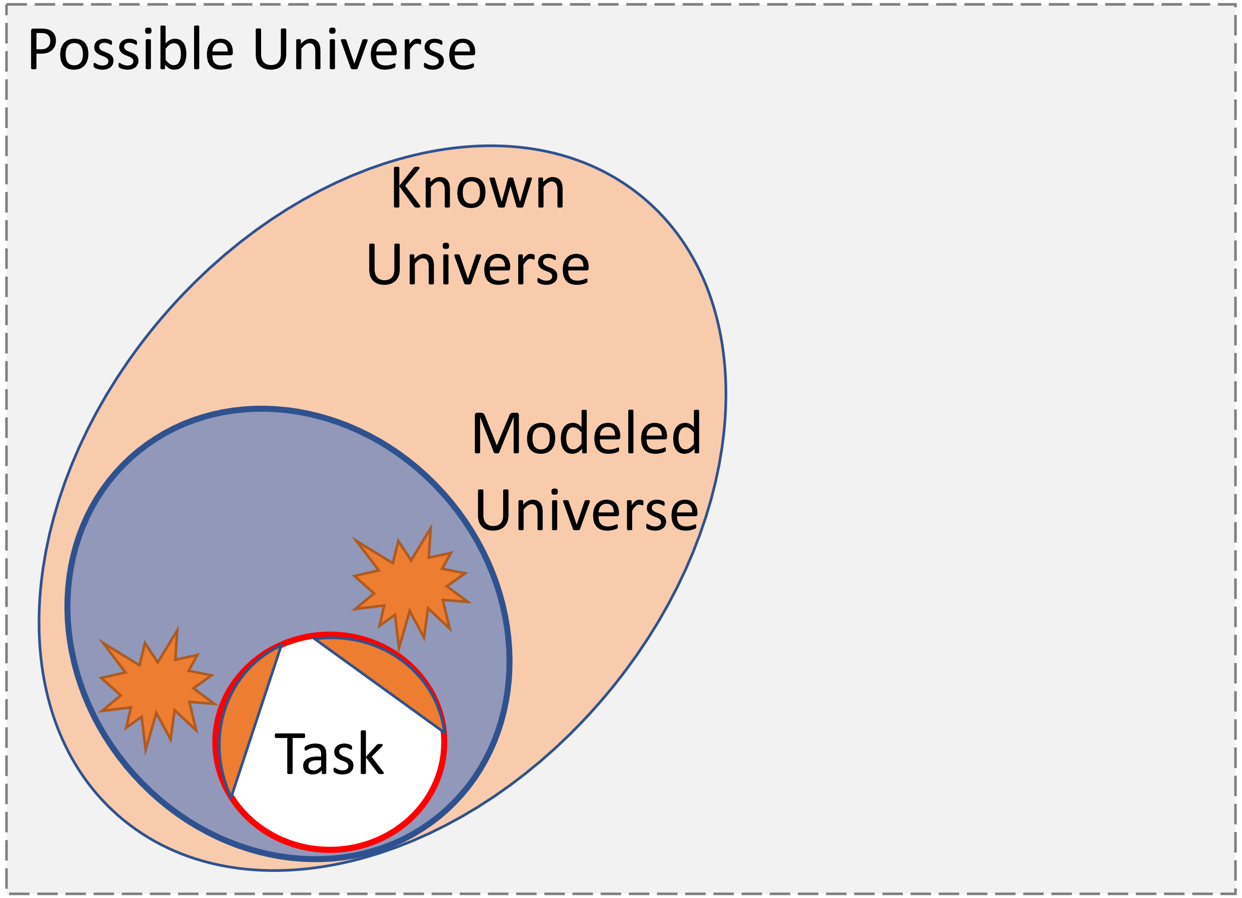
Result: Narrow AI System



- Supervised deep learning only learns features sufficient to separate the known classes
- Modeled Universe = Task Problem Space

Closed-World Safety-Critical Design

- Define task and Operational Design Domain (ODD)
- Enumerate known hazards
  - Hazard = a region of state space likely to lead to a harm
- Introduce margins of safety around each known hazard
- Design optimal control policy that respects margins of safety



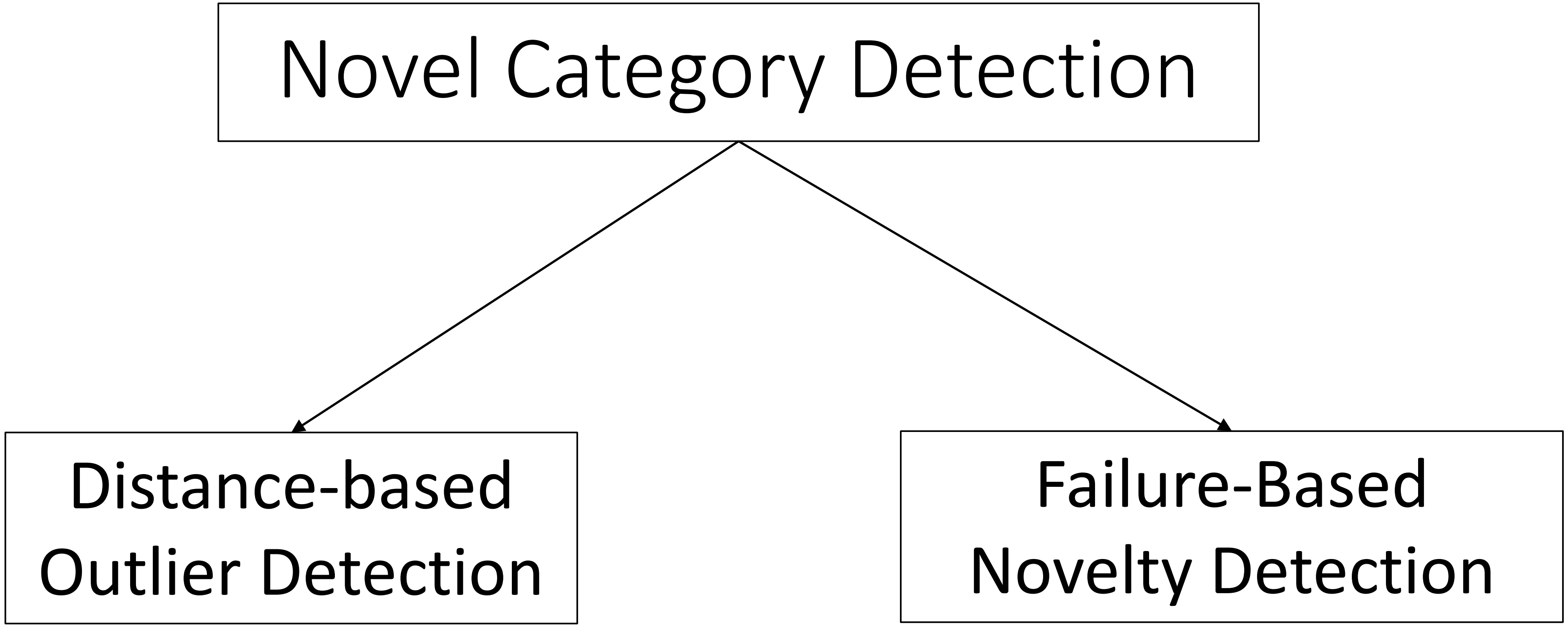
Key Challenge of Open Worlds: Novelty

- New diseases (e.g., COVID in chest x-rays)
- New objects (e.g., OneWheel for automated cars)
- New items (movies, books, songs, restaurants, etc.) for recommender systems
- Novel hazards
- Change in system dynamics
  - Flat tire
  - Loss of power steering

Source: onewheel.com

Insect Identification: There are $\approx 76{,}000$ species of freshwater insects worldwide

- 1,200 species in US
- Field samples may contain other things
  - leaves
  - trash
- Simple estimate of equal error rate for novel classes vs. the 54 classes was 20% (in 2011)
  - classifier is not usable without addressing the novel class problem
- "Novel Category Problem" or "Open Set Problem"


Novel Category Detection
Distance-based
Outlier Detection
Failure-Based
Novelty Detection

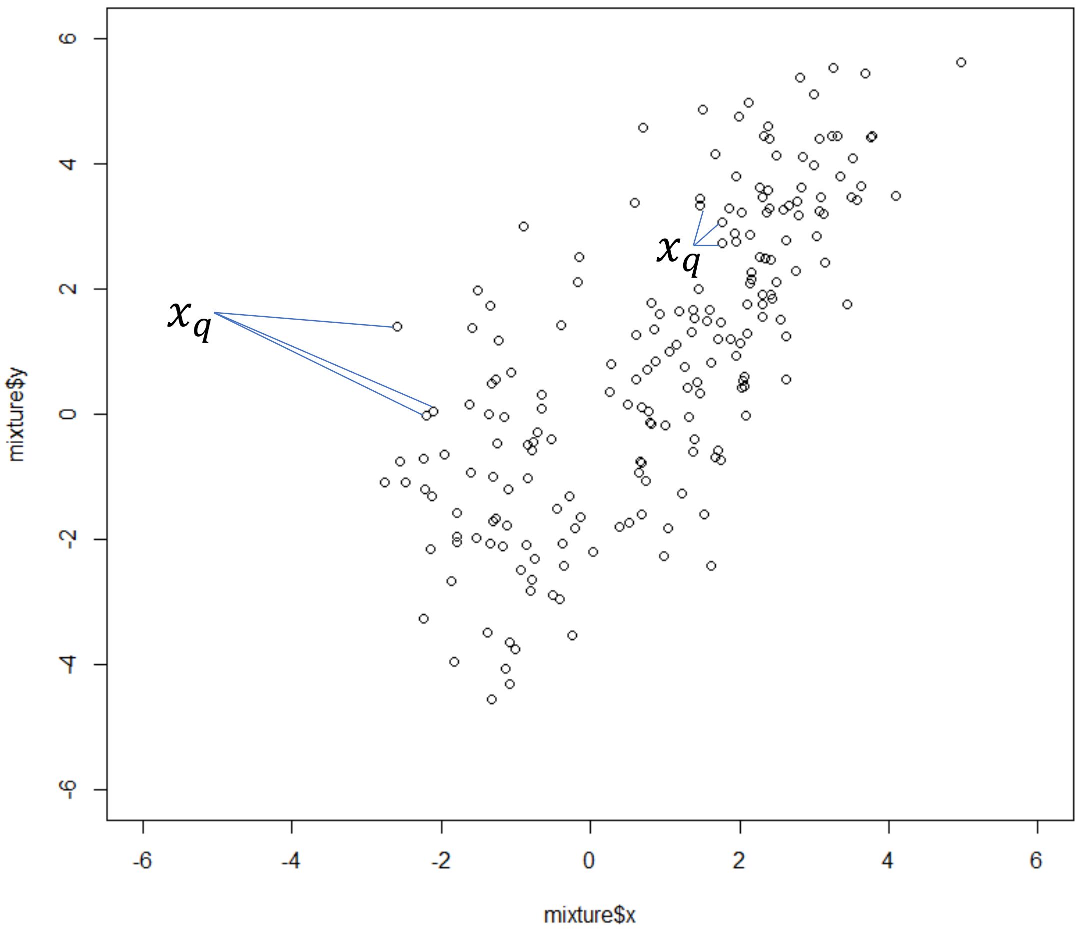
Distance-Based Outlier Detection

- Define a distance $d(x_i, x_j)$
- $A(x_q) = \min_{x \in D} d(x_q, x)$

- Can be made more robust by looking at the average distance to the k-nearest points
  - "k-nn anomaly detection"

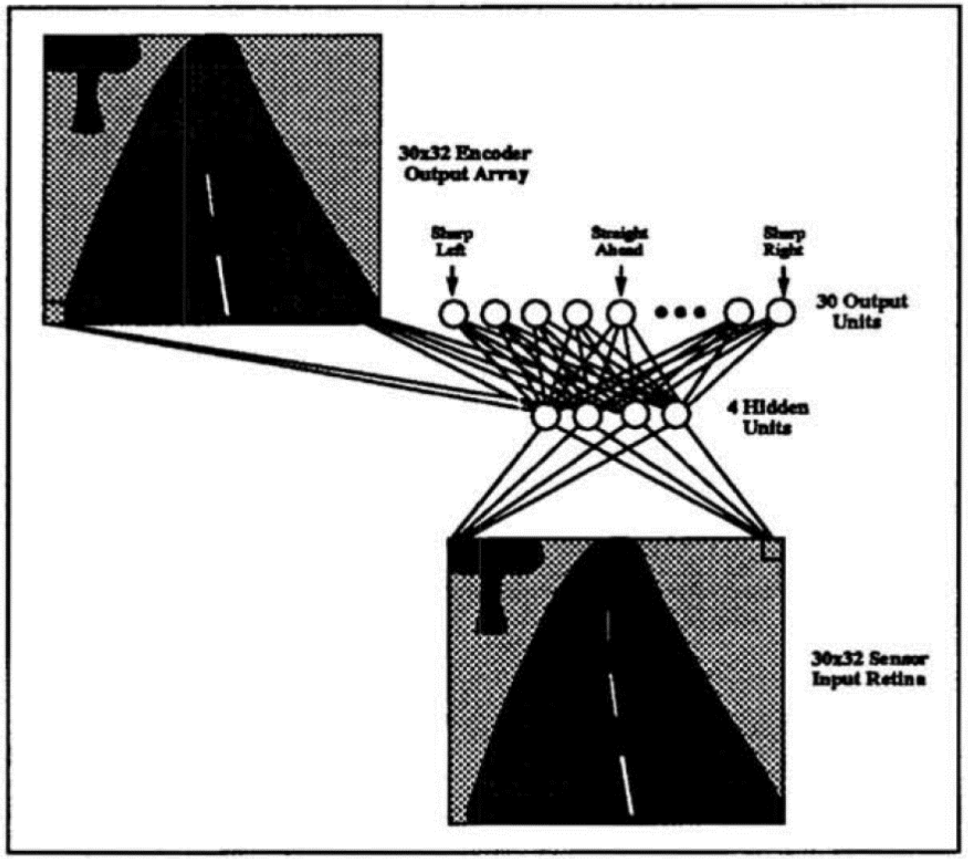
- Can be normalized by dividing by the distance of each neighbor to *their* k-nearest neighbors
  - "Local Outlier Factor (LOF)" Breunig, et al., 2000

- Efficient approximation: "Isolation Forest" Liu, et al., 2012



Reconstruction Failure

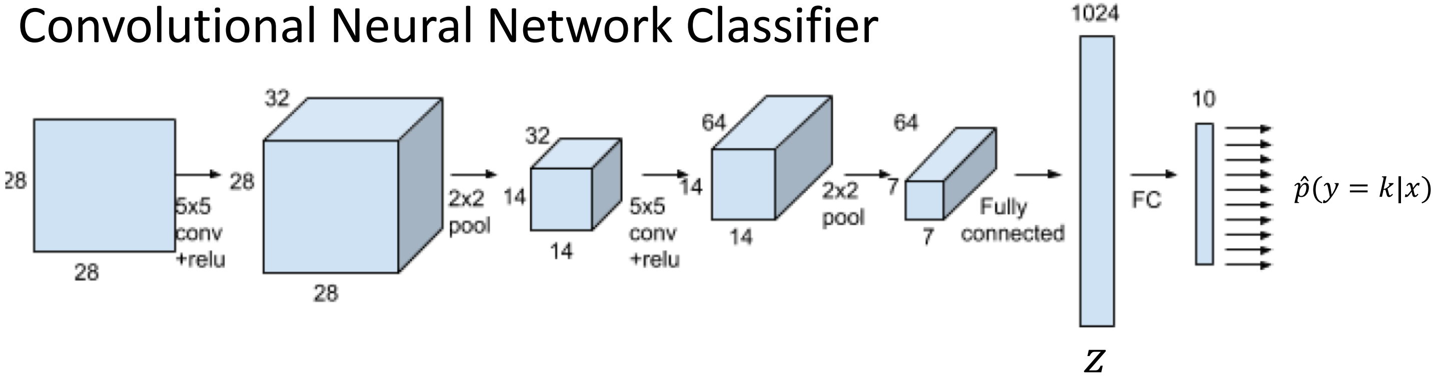
- Principle: Anomaly Detection through Failure
  - Define a task on which the learned system should fail for anomalies
- NavLab self-driving van (Pomerleau, 1992)
  - Primary head: Predict steering angle from input image
  - Secondary head: Predict the input image ("auto-encoder")
  - $A(x_q) = \|x_q - \hat{x}_q\|$
  - If reconstruction is poor, this suggests that the steering angle should not be trusted



Pomerleau, NIPS 1992

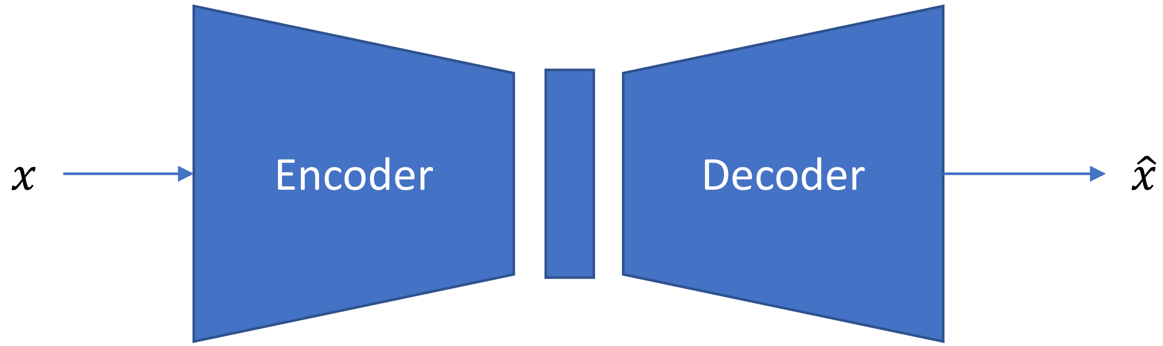
Distance-based Anomaly Detection for Deep Learning

Convolutional Neural Network Classifier



- Let $\mathrm{z} = (z_1, \ldots, z_{1024})$ be the features in the "penultimate" layer of the network.
- Logit score for class k is $\ell_k(z) = \sum_{j=1}^{1024} w_{jk} z_j$
- Probability for class k is $\hat{p}(y = k|x) = \frac{\exp \ell_k(z)}{\sum_{k'} \exp \ell_{k'}(z)}$
- Strategy: Apply distance-based methods to the z vectors

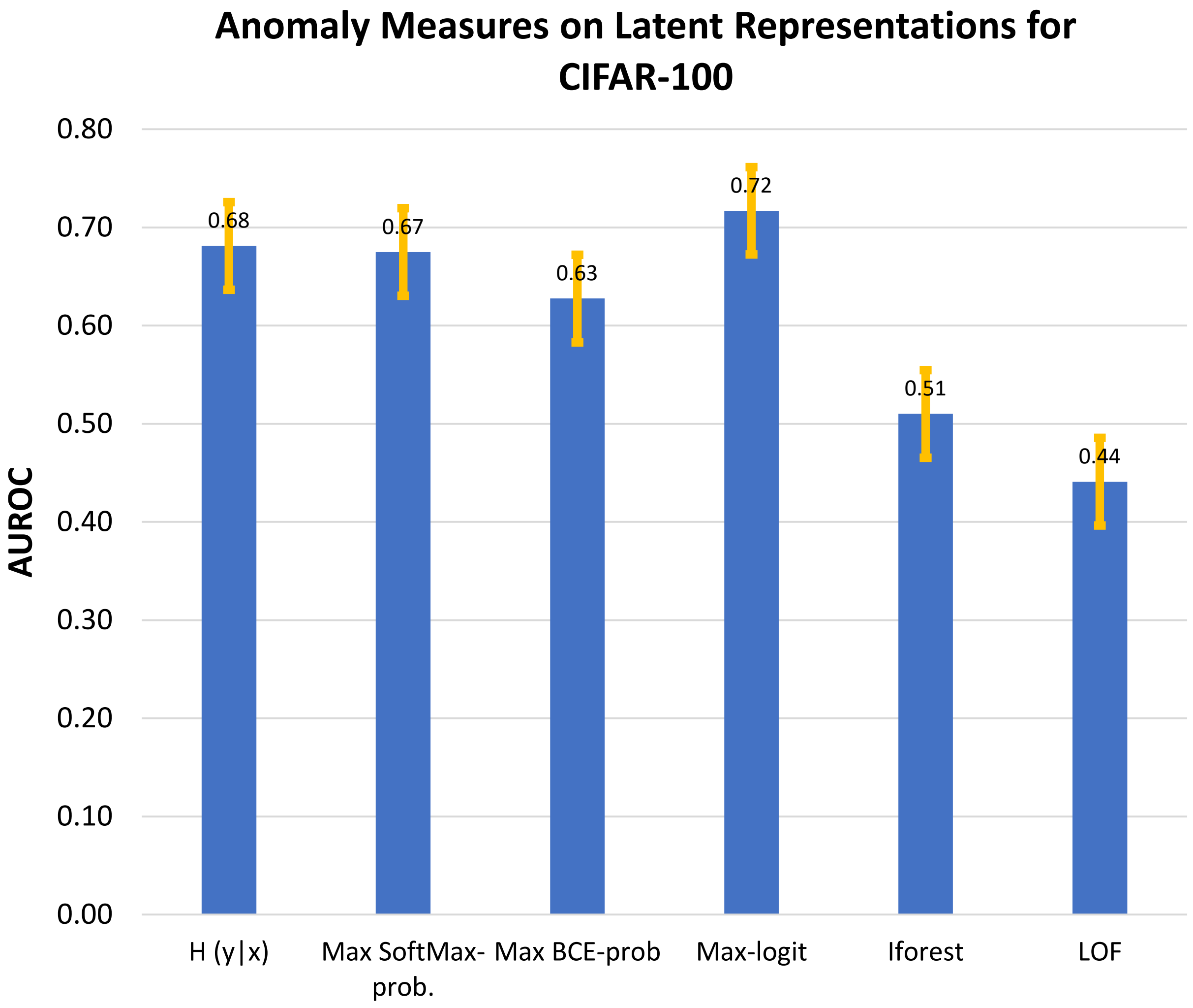
Reconstruction Failure: Deep Autoencoders



- The basic auto-encoder trains an encoder E and decoder D such that $D\left(E(x)\right) \approx x$ by minimizing the image reconstruction error
- The capacity of the bottleneck and of the decoder must be carefully controlled to prevent the network from learning a general-purpose image compression mapping
- Very few people can get this to work
- [but see Haoyang Liu, et al. Class-specific semantic generation and reconstruction for open set recognition. IJCAI 2024]

Experimental Evaluation of Outlier Detection Methods

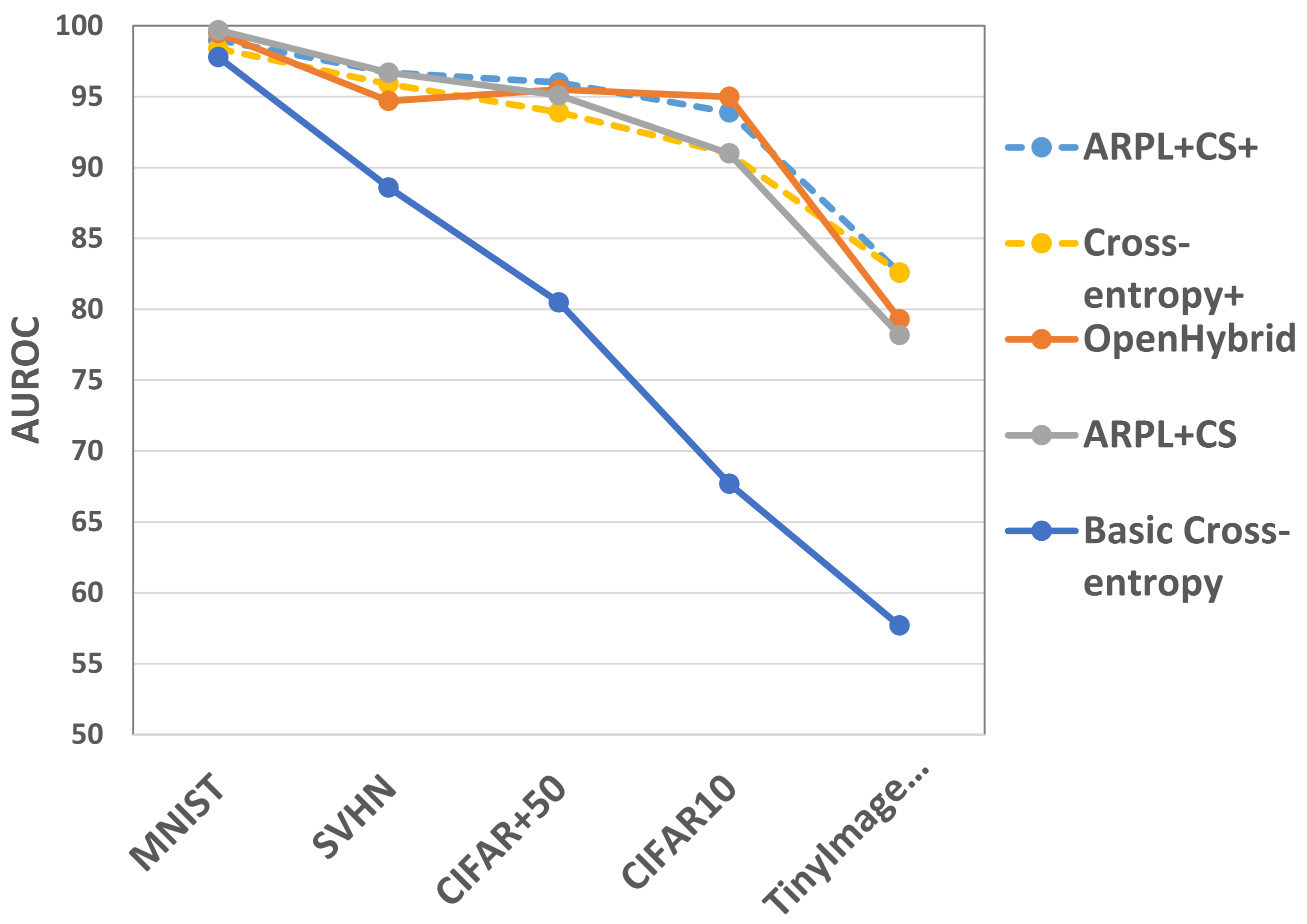
- CIFAR-100: 80 known classes; 20 novel classes
- Apply distance methods in z space
  - Isolation Forest on z_i
  - Local Outlier Factor (nearest neighbor method) on z_i
  - No better than random guessing
- Metrics based on "indecision"
  - $H(y|x)$: entropy of predicted probabilities $P(y|x)$
  - Max softmax probability: $\max_y P(y|x)$
  - Max Binary Cross-Entropy
  - Max logit: $\max_k \ell_k(x)$
  - Max logit is somewhat better than the others



Garrepalli, 2020

Max Logit Score: Experiments by Vaze, et al.

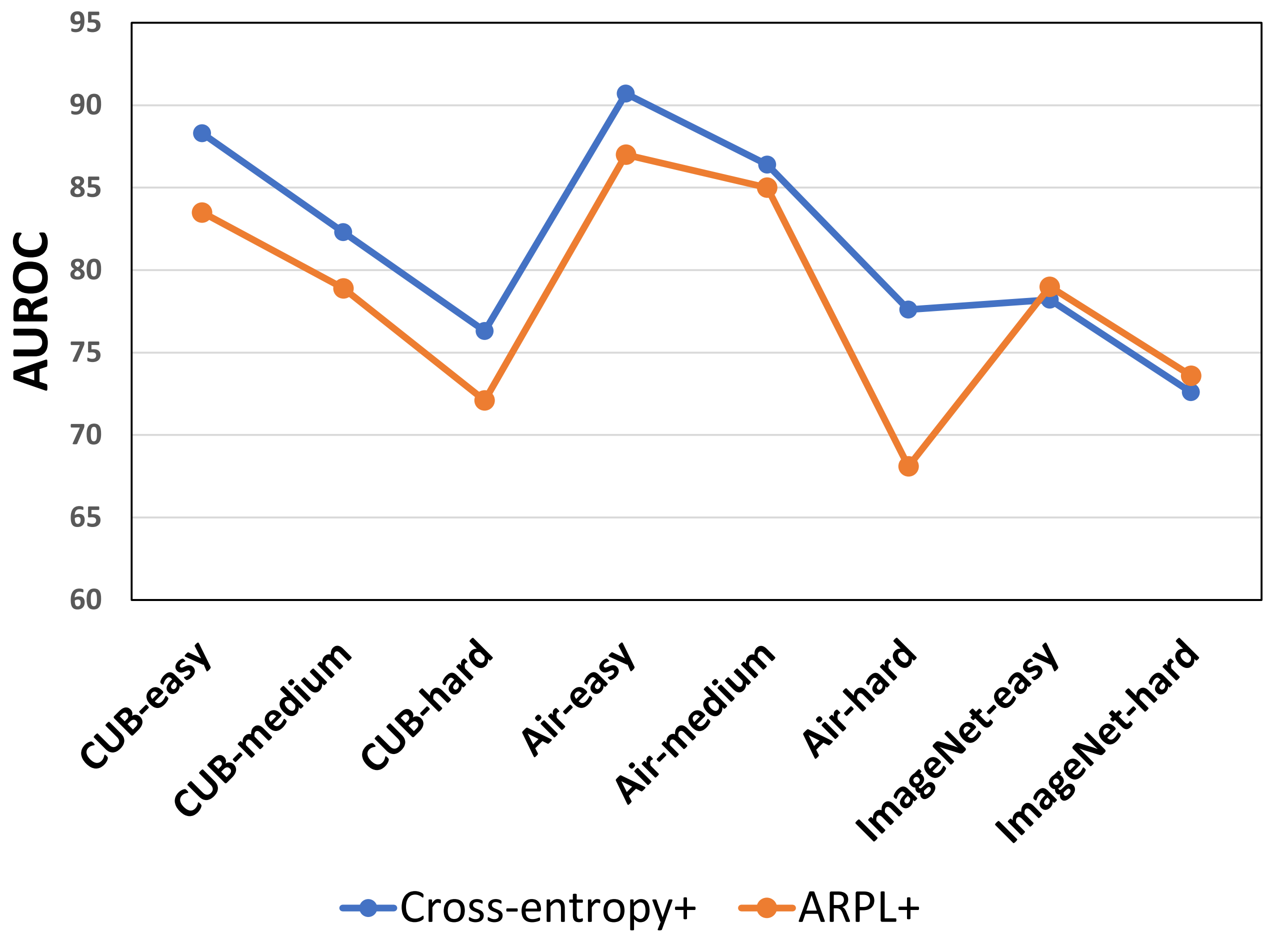
- Vaze, Han, Vedaldi, Zisserman (ICLR 2022): "Open Set Recognition: A Good Classifier is All You Need"
  - arXiv 2110.06207
- Cross-Entropy+: carefully train a classifier using the latest tricks
  - Standard cross-entropy combined with the following:
    - Cosine learning rate schedule
    - Learning rate warmup
    - RandAugment augmentations
    - Label Smoothing
- Anomaly score: max logit
  - $\max_k \ell_k(z)$
  - Small values ➔ anomalous



Protocol from Lawrence Neal et al. (2018)

Vaze, et al.: Three Large Open Set Benchmarks

- Novel class difficulty based on semantic distance
  - CUB: Bird species
  - Air: Aircraft
  - ImageNet

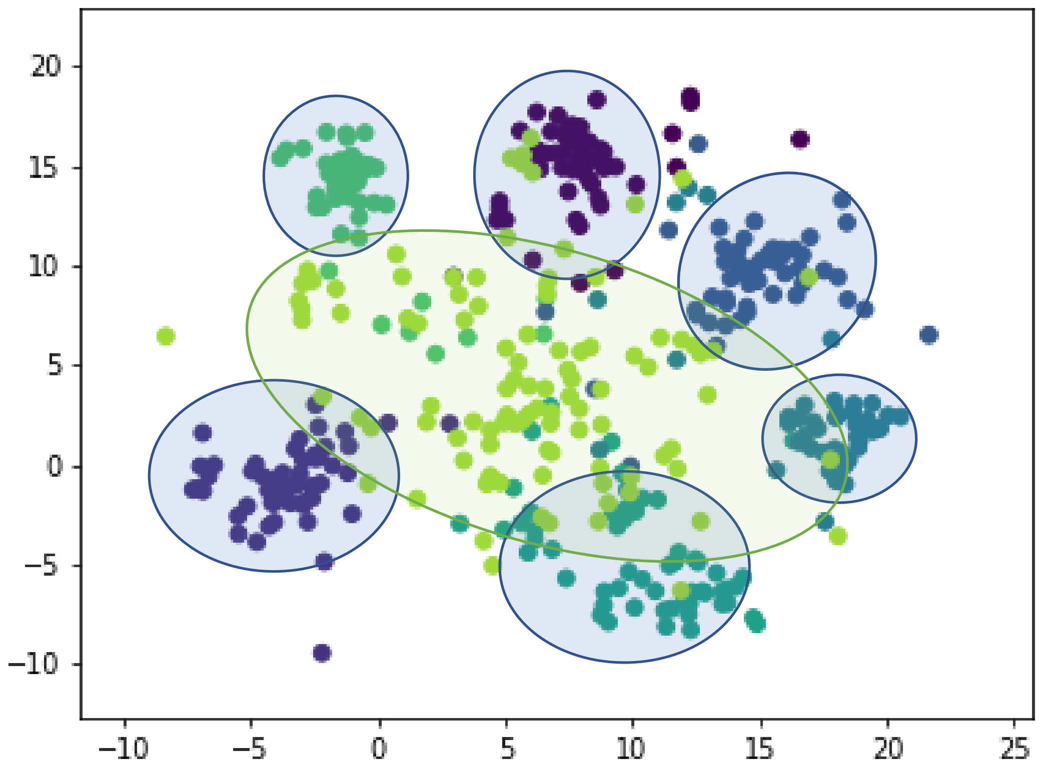


Why does Max Logit work?

Experiment:
Deep Learned Features in Computer Vision

- DenseNet with 384-dimensional latent space.
- CIFAR-10: 6 known classes, 4 novel classes
- Light green: novel classes
- Darker greens: known classes

- Images from known classes are “pulled out” from the center of the space
- Most novel-class images stay toward the center of the space; others overlap with known classes
- Novel images are “inliers”

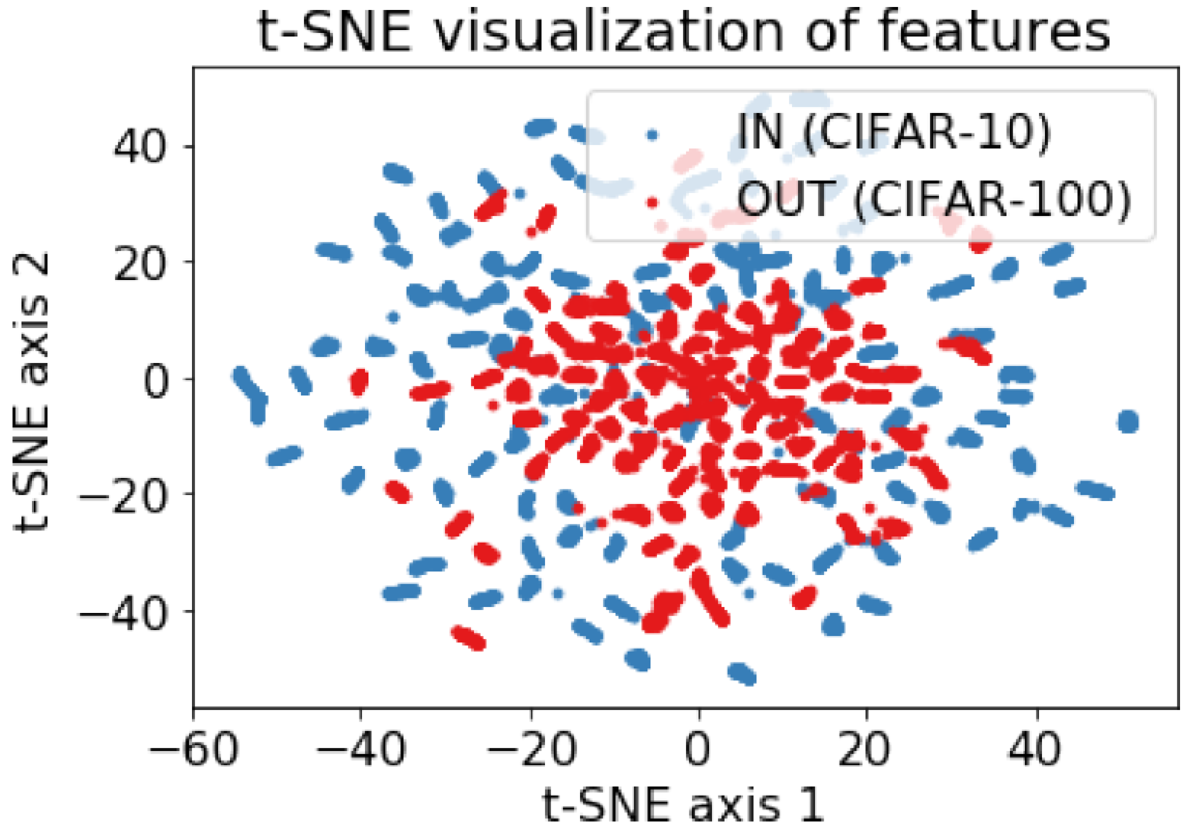


6 Known Classes

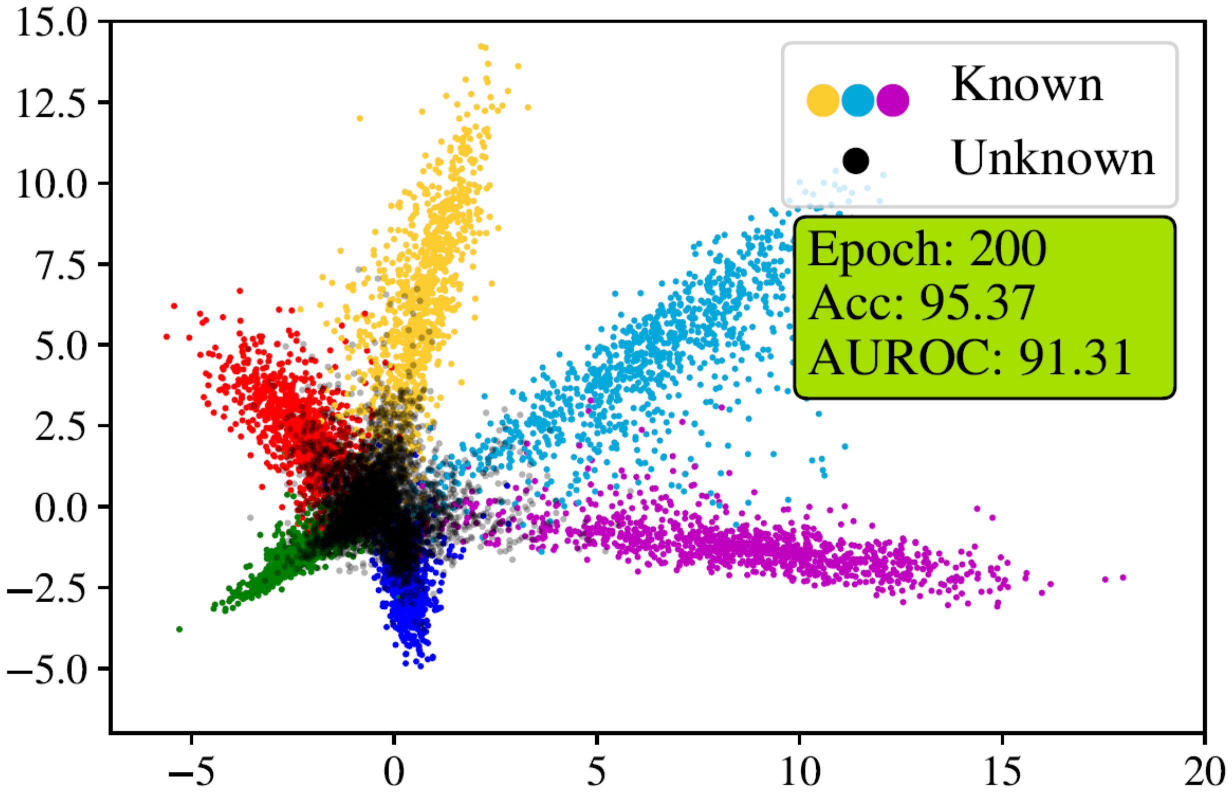
4 Novel Classes

Dietterich & Guyer, 2022

Similar Results from Other Groups



[Tack, et al. NeurIPS 2020]



[Vaze, et al. arXiv 2110.06207]

The Familiarity Hypothesis

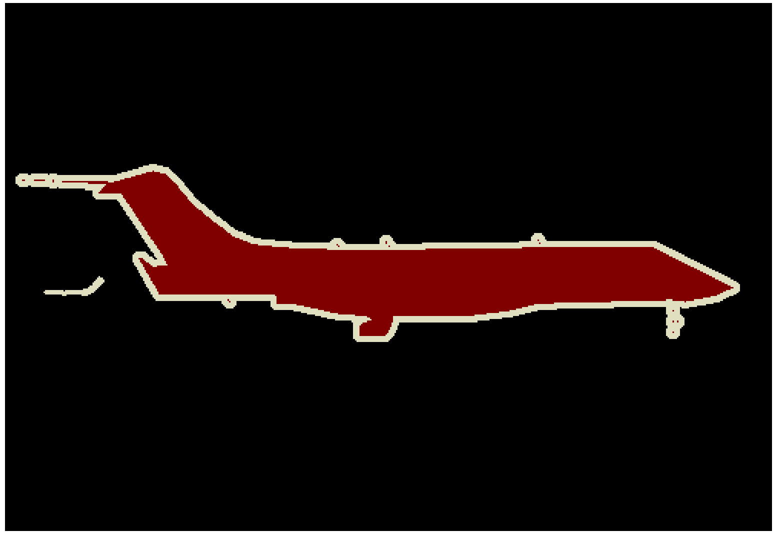
The network doesn't detect novelty, it detects the absence of familiarity

- Convolutional neural network learns "features" that detect image patches relevant to the classification task
- The logit layer weights these features to make the classification decision
- Novel classes activate fewer of these features, so their activation vectors are smaller
- Hypothesis: The networks don't detect that an elephant is novel because of trunk and tusks but because its head doesn't activate known features

Which features are responsible for the drop in activation?

Are they features "on" the object vs. the background?

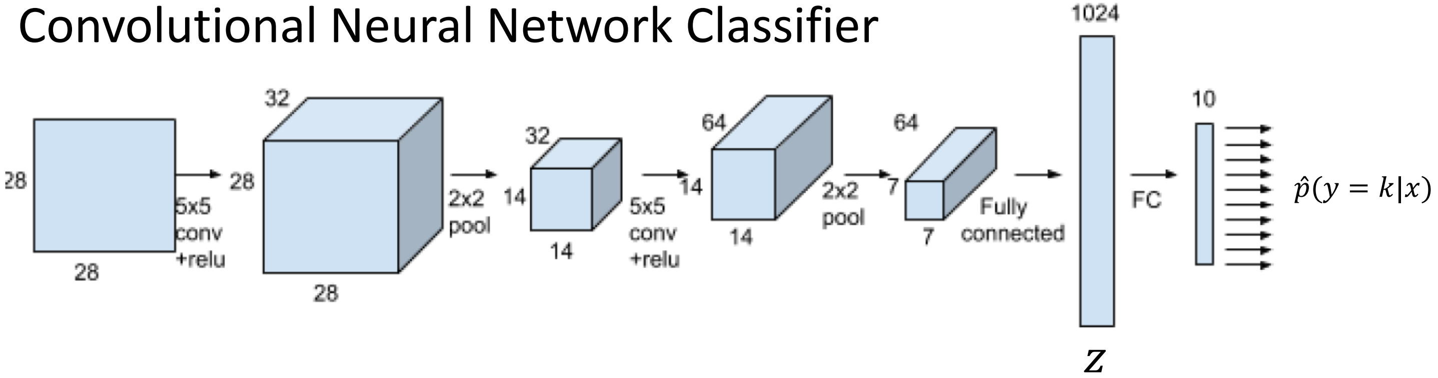
- Strategy: blur the object and see how the feature activations change
    - activations that change must be on the object
- Details:
    - PASCAL VOC Segmented Images
    - Blur the original image (31x31 kernel; sd=31)
    - Form composite image where blurred region replaces the segmented region



https://www.peko-step.com/en/tool/blur.html

Four Types of Features

Convolutional Neural Network Classifier



- Logit score for class k is $\ell_k(z) = \sum_{j=1}^{1024} w_{jk} z_j$
- $z_j \geq 0$ for all node functions in common use

- Presence features: Blurring causes their activation to drop
- Absence features: Blurring causes their activation to rise
- Positive features: $w_{jk} > 0$
- Negative features: $w_{jk} < 0$

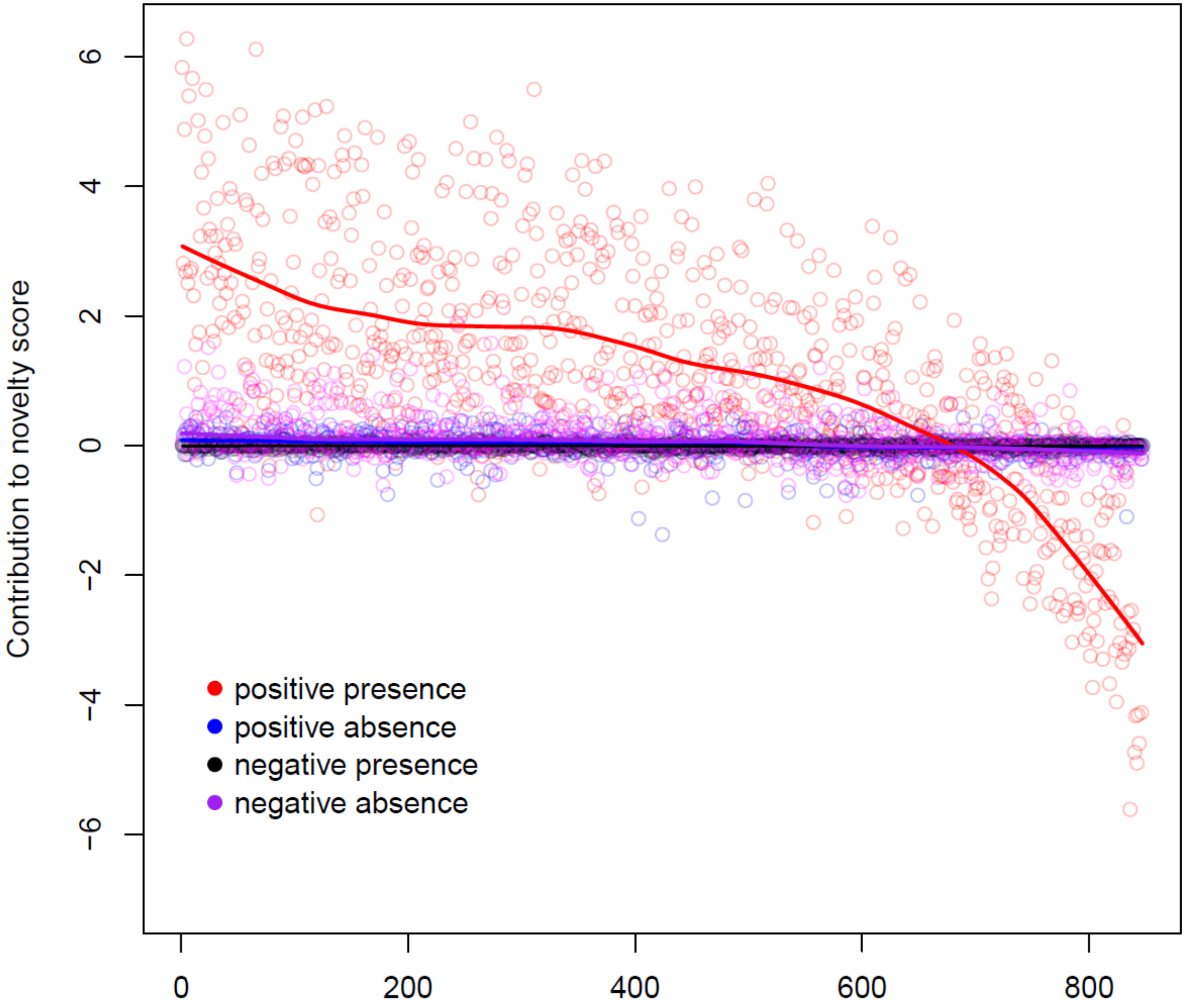
Class k	$w_{jk} > 0$	$w_{jk} < 0$
Presence	positive presence	negative presence
Absence	positive absence	negative absence

Familiarity Hypothesis:
Most features are positive presence

Contribution to max logit score

- Four points plotted for each novel image
  - Their sum is the max logit score
- Positive Presence features dominate the max logit score
- This confirms the familiarity hypothesis

- Similar results for ViT networks



Novel category images
(sorted by ascending max logit score)

Advantages and Disadvantages of Familiarity-Based Novelty Detection

Advantages:

- Each class defines its own "distance" based on its positive-presence features
- Avoids the need to define a global distance metric

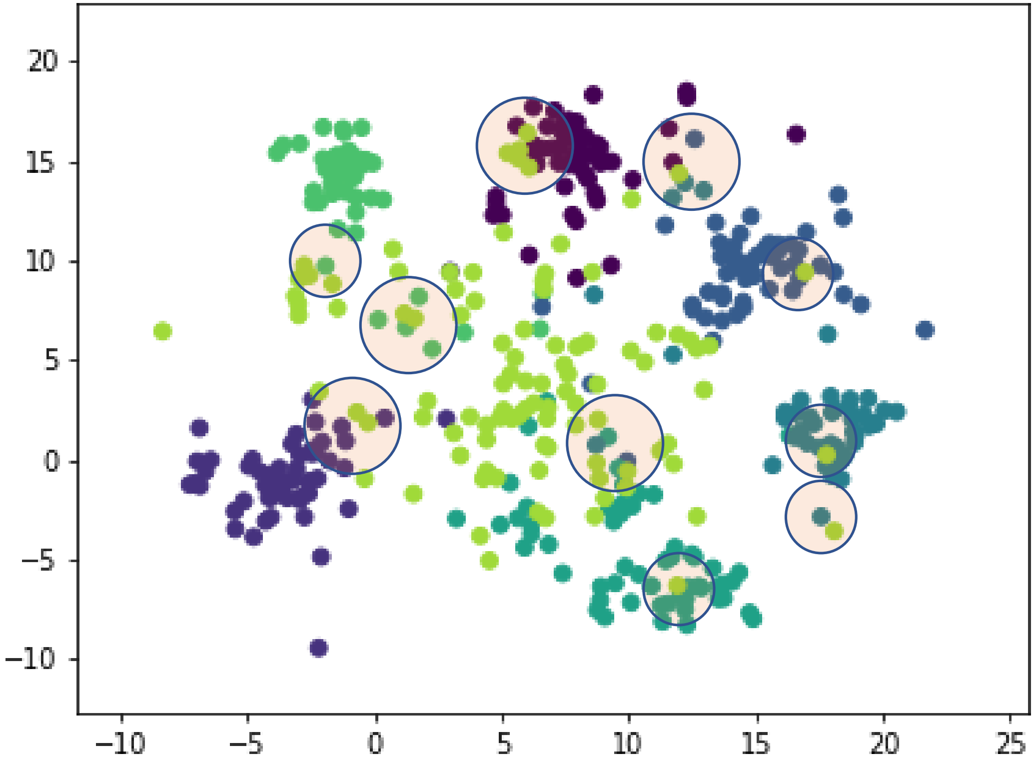
Disadvantages

- Extra features are not detected as anomalous
- Example: An elephant with wings will still be recognized as an elephant – the wings will be ignored by the elephant classifier
- Occlusions that hide familiar features will cause false novelty detections

Conclusion: Distances or Joint Distributions are necessary to detect novelty

The Learned Representation is Promising But Not a Complete Solution

- Many novel-class images are mapped onto clusters of known-class images
- The learned representation can’t detect all of the anomalies
- Max Logit AUROCs range from 0.73 to 0.91



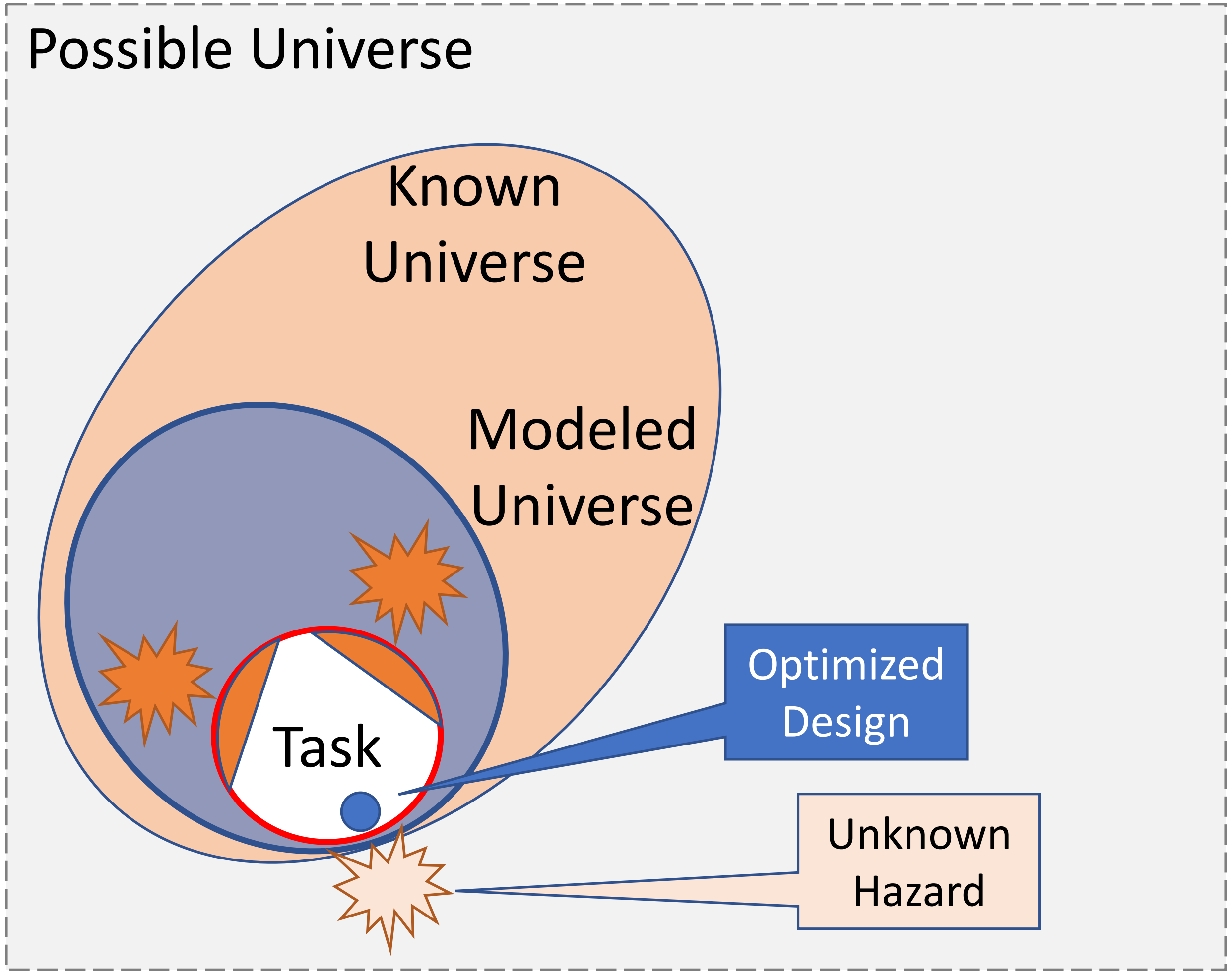
How can we learn better features?

- Foundation Model Approach:
  - Train on all the data we can find
  - Artificially introduce variation through augmentations
    - Rotations, flips, simulated snow, rain, pixel noise, etc.
  - Synthetic data

- The deep representation learns to “see” (represent) the known world
  - A Onewheel will still be novel, but the model should have the right features to represent it and thereby separate it from all known objects

Novel Hazards and Robust-yet-Fragile Systems

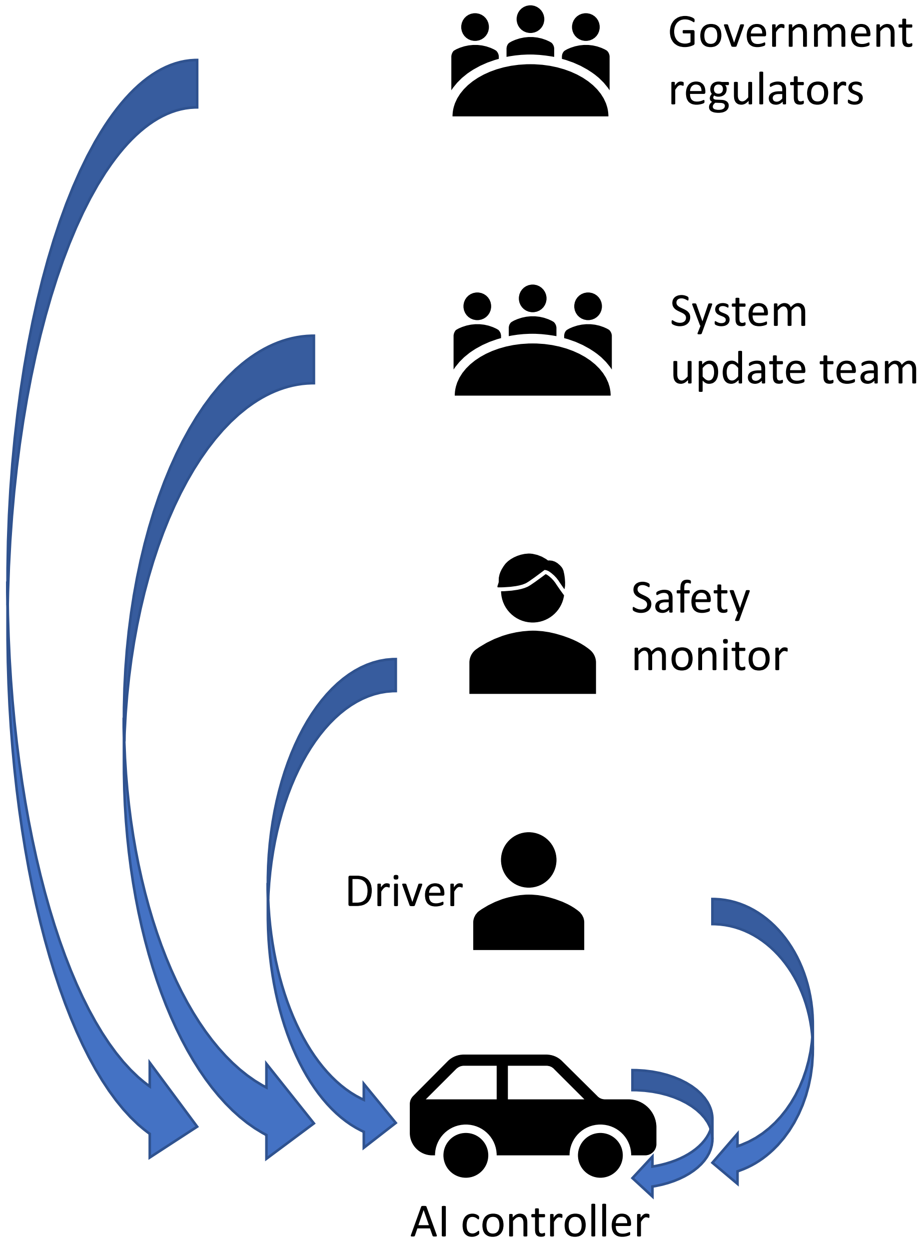
- Engineered systems are "robust yet fragile"
  - Robust to the known hazards
  - Vulnerable to novel failure modes
- Optimization for cost, weight, etc. results in designs near the edge of the feasible region
  - Highly Optimized Tolerances (HOT) theory. Carlson & Doyle (2002)
- Small change in operating conditions leads to novel failure



Systems View of Safety

[Leveson 2011: Engineering a Safer World]

- A system (including the human organizations that build, use, and operate it) can be decomposed into a hierarchy of subsystems, each with its own controller
- These systems are subject to many disturbances
  - Environmental Novelty
  - New regulations
  - Budget cuts and staff reductions
    - Systems tend to migrate toward the edges of safety
- A safe controller must detect and compensate for these disturbances
  - Today: It is the exclusively the humans who do this



High Reliability Human Organizations

Todd LaPorte, Gene Rochlin, and Karlene Roberts (Weick, et al., 1999)

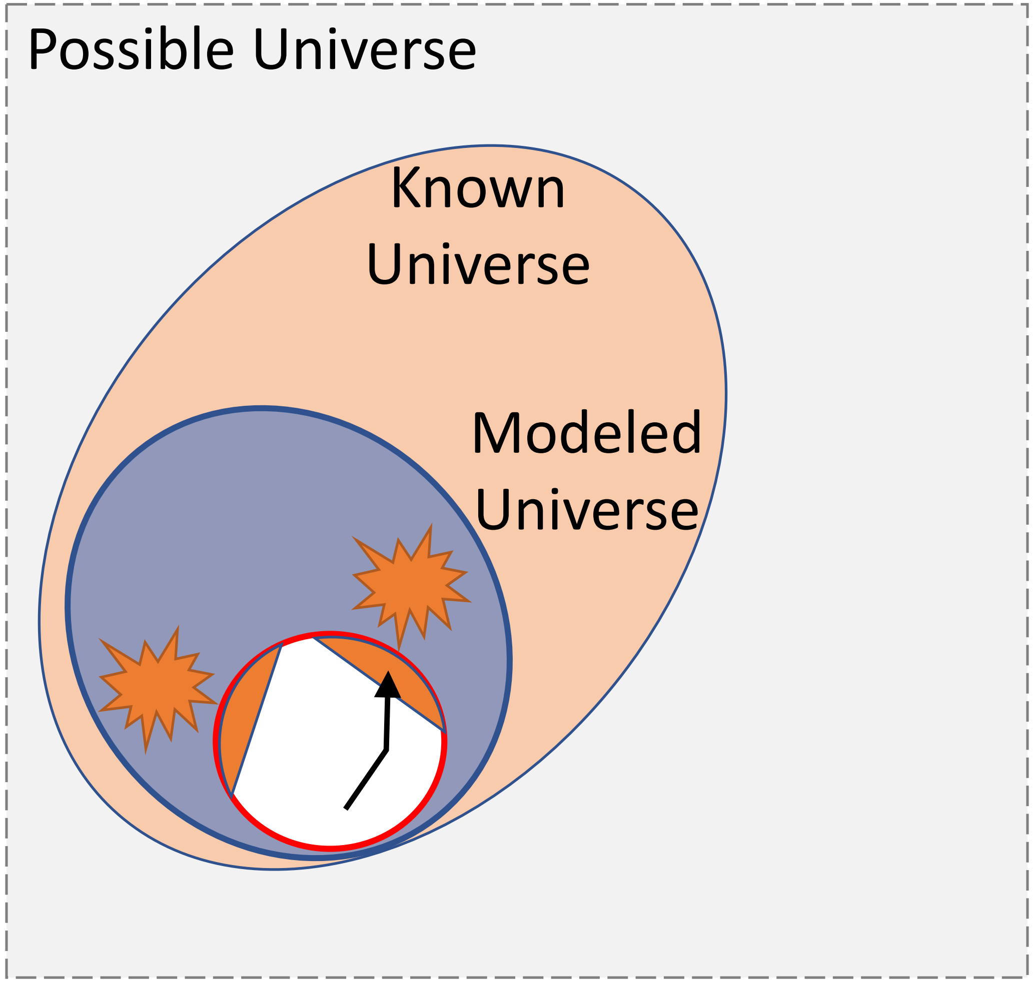
- Preoccupation with failure
  - Fundamental belief that the system has unobserved failure modes
  - Treat anomalies and near misses as symptoms of a problem with the system
- Reluctance to simplify interpretations
  - Comprehensively understand the situation
- Sensitivity to operations
  - Maintain continuous situational awareness
- Commitment to resilience
  - Develop the capability to detect, contain, and recover from errors. Practice improvisational problem solving
  - David Woods: A resilient organization is “poised to adapt”
- Deference to expertise
  - During a crisis, authority migrates to the person who can solve the problem, regardless of their rank

How can AI Help?

- Maintain Situational Awareness
- Anomaly Detection (already discussed)
- Near Miss Detection
- Novelty Diagnosis
- Automated or Suggested Repairs

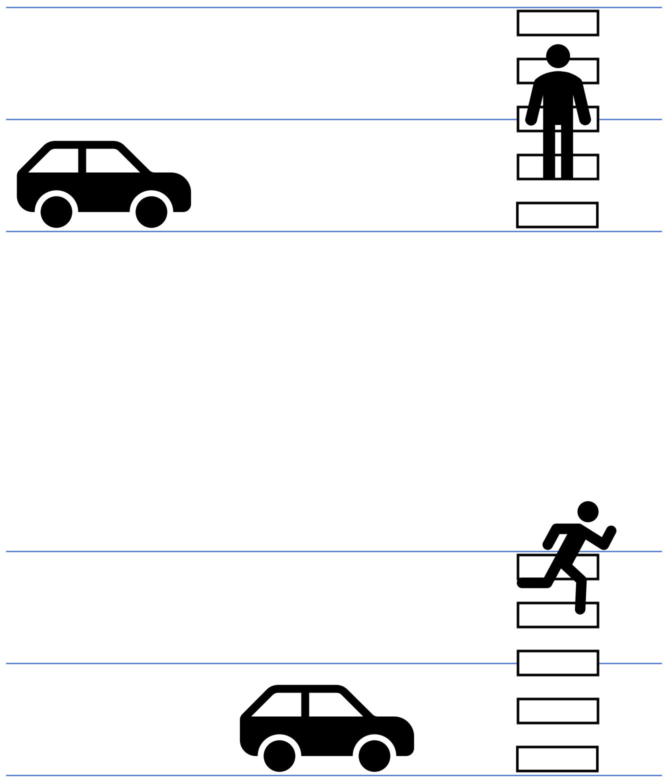
Near Miss Detection

- Case 1: Near Miss for Known Hazard
  - System violates the margin of safety region near a known hazard



Case 2: Counterfactual Near Misses

- Automatic Vehicle safety conditions
  - At least 2m separation between vehicle and pedestrians, cyclists, stationary obstacles
- Pedestrian sees car coming and jumps out of the way
- Car determines that it met the required 2m separation ➔ “no problem”
- Counterfactual: There would have been a safety violation if the pedestrian had not taken evasive action

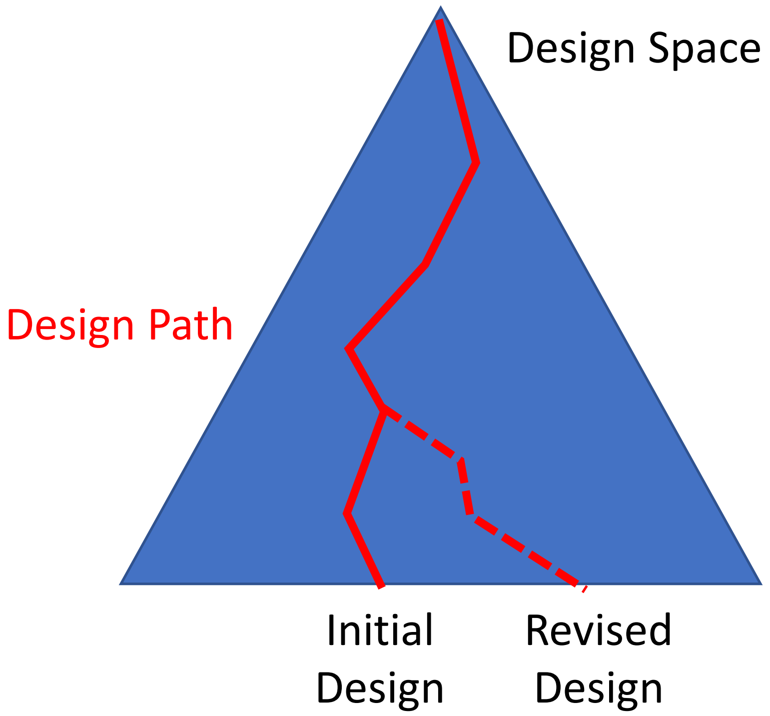


Novelty Diagnosis and Repair

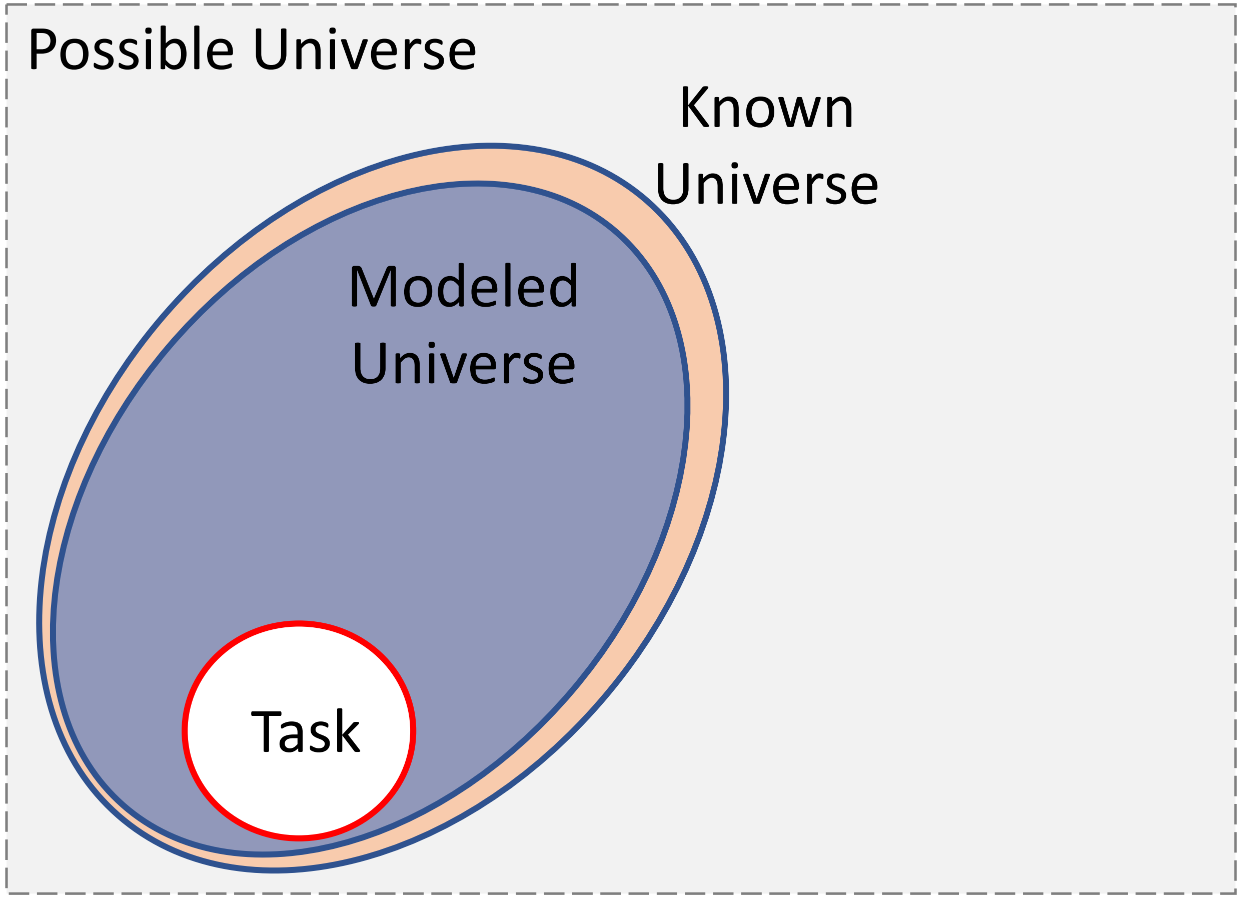
- Easy Cases:
  - Anomaly caused by novel category
    - Repair: Collect training data for the novel category and retrain the detector/classifier
  - Near miss caused by controller failure
    - Repair: Generate additional training trajectories for the controller
- Harder Cases:
  - Perceptual failure caused by novel type of occlusion
    - Repair: Improve context-dependent uncertainty quantification. Controller will be more cautious under high uncertainty
    - Repair: Add sensors that do not suffer from the occlusion
  - Characterize novel hazard. Under what conditions will the hazard occur?
    - Repair: Define new hazard region; retrain the controller
    - May require defining new state variables, adding sensors, and improving state estimation

Creating Resilient Systems

- David Woods: A resilient system is one that is "poised to adapt"
  - Surprises are often not visible through standard sensors/communication paths
  - Organizations must practice communicating and adapting to confront novelty
- An AI perspective:
  - The entire design process should be regarded as one path through a design space
  - Adaptation requires following new paths through that space
  - The design space and design process should be "kept on standby" so that they can be invoked whenever adaptation is required



What is the Role of LLMs (and Foundation Models, more generally)?

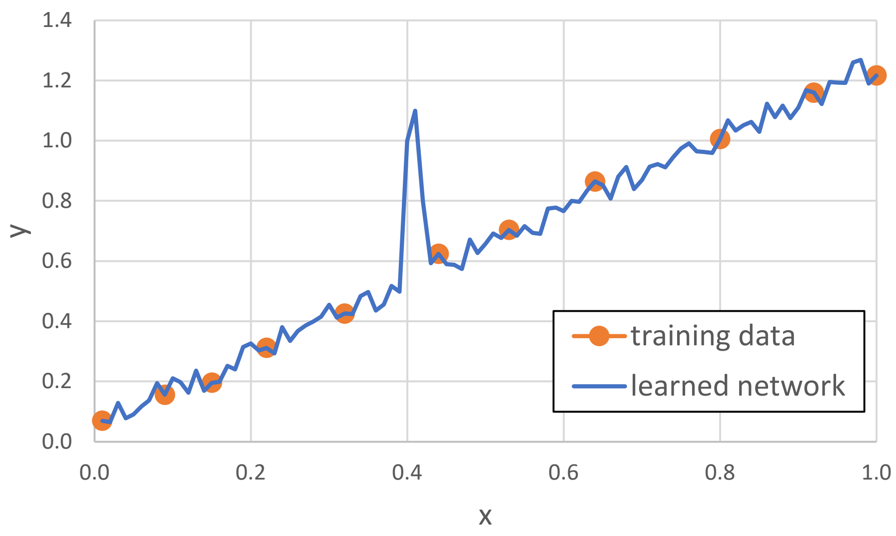


- Modeled Universe is nearly identical to the Known Universe
- Greatly improves learned representations
- Supports improved anomaly detection and diagnosis
- Improves situational awareness by expanding the context
- However, learned knowledge may be unreliable

Future Directions

Beyond Statistical Learning

- Distribution-Independent Machine Learning
  - We need more than statistical guarantees of correctness
  - Can we verify that our learned models are well-behaved?
    - Smoothness
      - Bounded curvature
      - Bounded Lipschitz constant
      - Bounded distance from linear interpolation
- Can we verify that our uncertainty quantification is correct?
- Can we prove that there are no spurious correlations?
  - Learned relationships are causal; no hidden confounders



Summary

- Functional and safety engineering address the known sources of variation and the known hazards
  - This makes them robust yet fragile
  - Deep Learning representations only capture the variation necessary to perform the task
- Key challenge #1 of open worlds: novel categories
  - Novel categories of objects, behaviors, etc.
  - Existing novelty detection methods perform poorly with deep learning
  - Familiarity-based max logit score works better, but only if the representation separates novelties from the known categories
- Key challenge #2 of open worlds: novel hazards
  - Counter-factual near misses
- Resilient Systems
  - Design space "on standby"
- Foundation Models
  - Revolutionary improvements in learned representations
- Future Directions
  - Research on counter-factual hazards
  - Distribution-independent machine learning
  - Verifiable uncertainty quantification

References

- Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *ACM SIGMOD 2000 International Conference on Management of Data*, 1–12.
- Carlson, J. M., & Doyle, J. (2002). Complexity and robustness. *Proceedings of the National Academy of Sciences*, *99*(Supplement 1), 2538–2545. https://doi.org/10.1073/pnas.012582499
- Dietterich, T. 2016. Steps Toward Robust Artificial Intelligence. *AI Magazine, 38*(3): 3-24.
- Dietterich, T. G., Guyer, A. (2022). The Familiarity Hypothesis: Explaining the Behavior of Deep Open Set Methods. https://arxiv.org/abs/2203.02486 . Also *Pattern Recognition 132*, p. 108931 (2022).
- Garrepalli, R. (2021). Oracle Analysis of Representations for Deep Open Category Detection. MS Thesis. https://arxiv.org/abs/2209.11350
- Getoor, L., Taskar, B., & Koller, D. (2001). Selectivity Estimation using Probabilistic Models. *ACM SIGMOD 2001*.
- Horvitz, E. 2008. Artificial Intelligence in the Open World. AAAI Presidential Address. https://www.microsoft.com/en-us/research/publication/artificial-intelligence-open-world/
- Leveson, N. G. (2011). *Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press. Cambridge, MA.*
- Larios, N., Soran, B., Shapiro, L., Martínez-Muños, G., Lin, J., Dietterich, T. G. (2010). Haar Random Forest Features and SVM Spatial Matching Kernel for Stonefly Species Identification. *IEEE International Conference on Pattern Recognition (ICPR-2010)*.
- Lin, J., Larios, N., Lytle, D., Moldenke, A., Paasch, R., Shapiro, L., Todorovic, S., Dietterich, T. (2011). Fine-Grained Recognition for Arthropod Field Surveys: Three Image Collections. *First Workshop on Fine-Grained Visual Categorization (CVPR-2011)*
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1–39.

References (2)

- Lytle, D. A., Martínez-Muñoz, G., Zhang, W., Larios, N., Shapiro, L., Paasch, R., Moldenke, A., Mortensen, E. A., Todorovic, S., Dietterich, T. G. (2010). Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society*, 29(3), 867-874.
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *ArXiv*, *2301.06627*(v1). https://arxiv.org/abs/2301.06627
- McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. http://arxiv.org/abs/2309.13638
- Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., … Welling, J. (2018). Never-ending learning. *Communications of the ACM*, *61*(5), 103–115. https://doi.org/10.1145/3191513
- Pomerleau, D. A. (1993). Input Reconstruction Reliability Estimation. *Proceedings of NIPS 1993*, 279–286. http://papers.nips.cc/paper/631-input-reconstruction-reliability-estimation.pdf
- Tack, J., Mo, S., Jeong, J., & Shin, J. (2020). CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. *Advances in Neural Information Processing Systems (NeurIPS 2020)*.
- Vaze, S., Han, K., Vedaldi, A., & Zisserman, A. (2021). Open-Set Recognition: A Good Closed-Set Classifier is All You Need. *ArXiv*, *2110.06207*(v1), 1–23. http://arxiv.org/abs/2110.06207
- Weick, K., Sutcliffe, K., & Obstfeld, D. (1999). Organizing for high reliability: Processes of collective mindfulness. In R. S. Sutton & B. M. Staw (Eds.), *Research in Organizational Behavior* (Vol. 1, pp. 81–123). Jai Press.