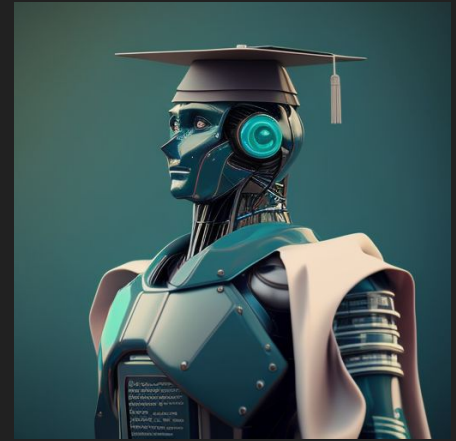


# Data Curation: Transforming Bytes into AI Gold

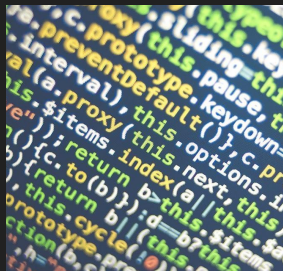
Richard Decal  
Senior ML Scientist at Dendra.io  
2024 AVL AI in Production Conference

Machine Learning is a  
new software  
paradigm that learns  
from data



# Traditional software

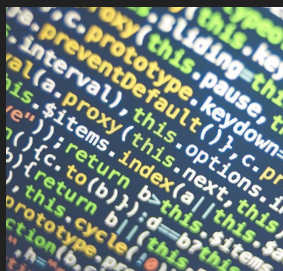
Code



Behavior

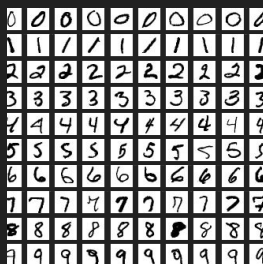
# Machine learning

Code (model)



Behavior

Data

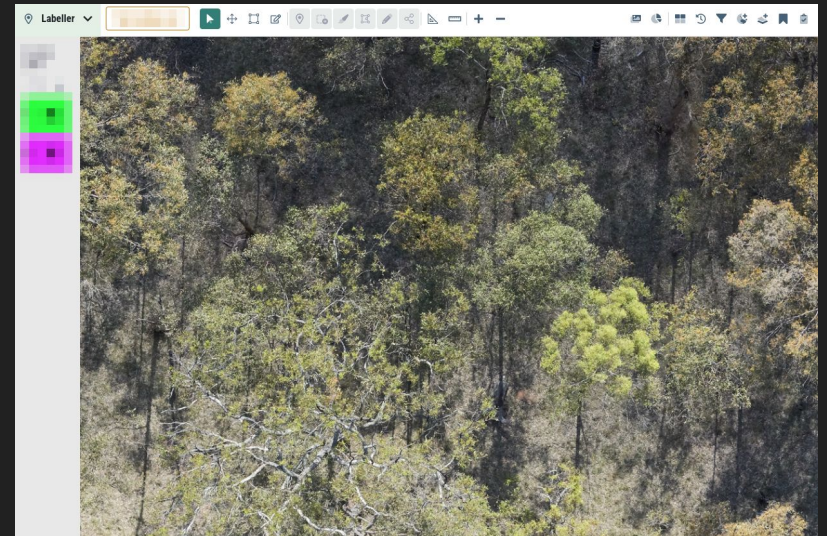


Data is ~code

Data curation is key

# Why me?

- Richard Decal
- Senior ML Scientist at Dendra
- AI for ecosystem restoration
  - Drone swarms for seed spreading at scale
  - Ecosystem monitoring using ultra-high resolution aerial imagery and AI ← my role



# 90%+

Time spent on data curation problems *after* deploying ML model to production

Challenge:  
Training data has to  
represent the cases that  
it will see in production,  
otherwise models can  
misbehave





# A story about domain shift

**BUSINESS  
INSIDER**

TRANSPORTATION

## **A Tesla crashed into a private jet worth up to \$3.5 million while in Smart Summon mode, witness says**

Ryan Hogg Apr 23, 2022, 11:22 AM EDT



*Thankfully, no Tesla Autopilot engineers in the audience...*

# How to detect out-of-distribution data in classification models

- Label-free & domain-agnostic methods
  - Maximum-logit score: detecting if model is blind to a pattern
  - Shannon entropy: detecting if model is perplexed by a sample
  - Max confidence margin score
- Domain-specific hand-crafted rules
  - At Dendra: class frequencies, geofencing species distribution
  - Self-driving cars: object detection “flickering”

## Harvest this data!

*Karpathy. “Tesla AI Day 2021”.*

*Hendrycks et al. “A baseline for detecting misclassified and out-of-distribution examples in neural networks”. 2016.*

*Lakshminarayanan et al. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. 2016.*

*Vaze et al. “Open-Set Recognition: A Good Closed-Set Classifier Is All You Need?”. 2021.*

Challenge:  
Models reflect the  
biases in their training  
data



## Myth:

ML models are data-driven and emotionless, thus they are more objective than human decision-making.

## Social biases:

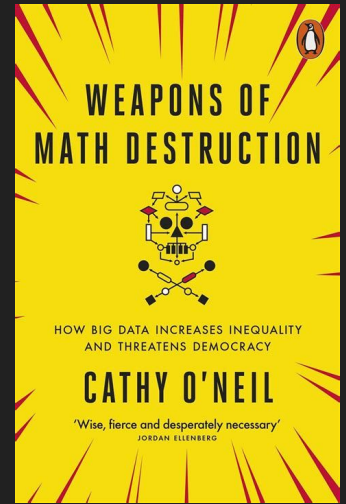
- Our written text contain implicit social biases on gender, race, age, etc.

## Models can memorize random correlations

- E.g. words that frequently co-occur with abusive tweets can teach model that sports fans are violent
- Models can entangle concepts

**Majority bias:** models struggle to learn rare phenomena

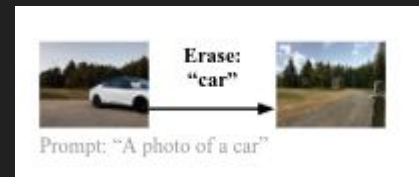
**Selection bias:** Unfair data generation



# “To change the world, change the data”

- Collect more data of rare examples
  - Ideal option
  - Sometimes too expensive, or impossible (e.g. rare medical cases)
- Synthesize data to overcome biases
  - E.g. synthesize text about female doctors, lawyers, etc. to make models less sexist
- Oversample rare scenarios, or gradient re-weighting
  - Not ideal; models can memorize specific training samples rather than generalize

Train model in an “alternate reality” where the biases don’t exist.



Rogers. “Changing the World by Changing the Data”. 2021.

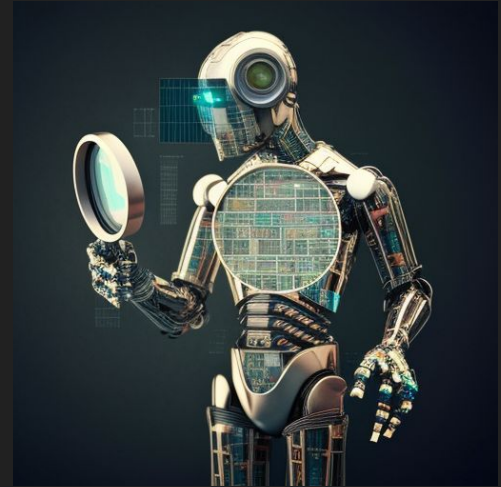
Trabucco et al. “Effective Data Augmentation with Diffusion Models”. 2023.

Nichol. “DALL-E 2 pre-training mitigations”. 2022.

# Challenge:

More data is not always  
better

Not all data are created  
equal

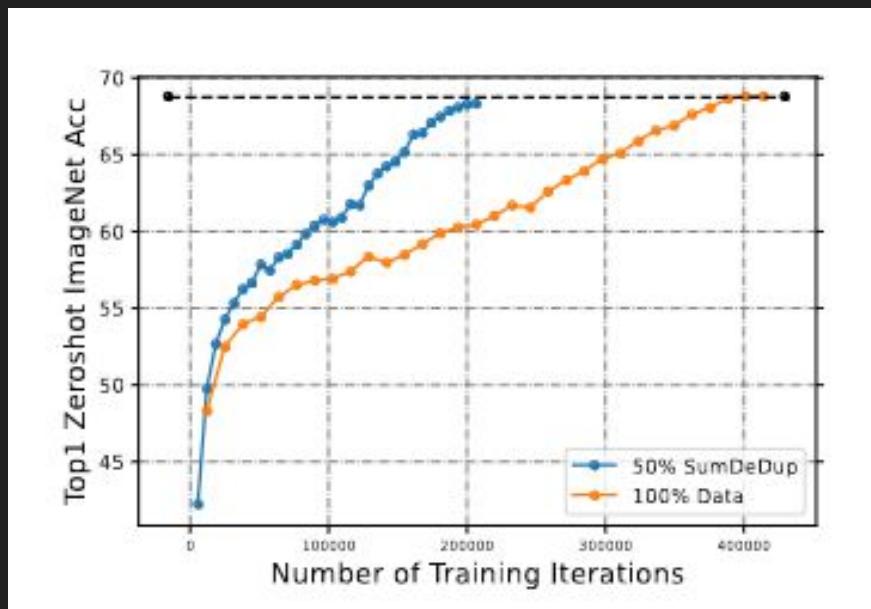


# Not all data are created equal

- Model does not learn from easy samples
- Model does not learn from samples that are too perceptually similar.  
Diversity is important!
- Model can memorize (and regurgitate) data which are duplicated in the data.

# Pruned datasets can out-perform unpruned datasets

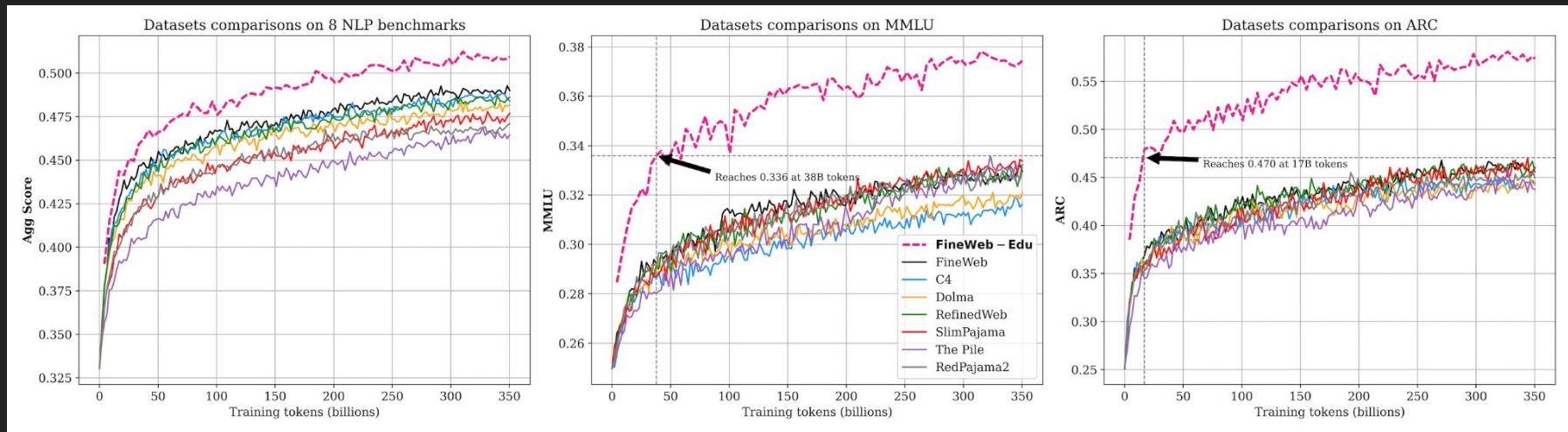
Example: SemDeDup reduced perceptually similar data.





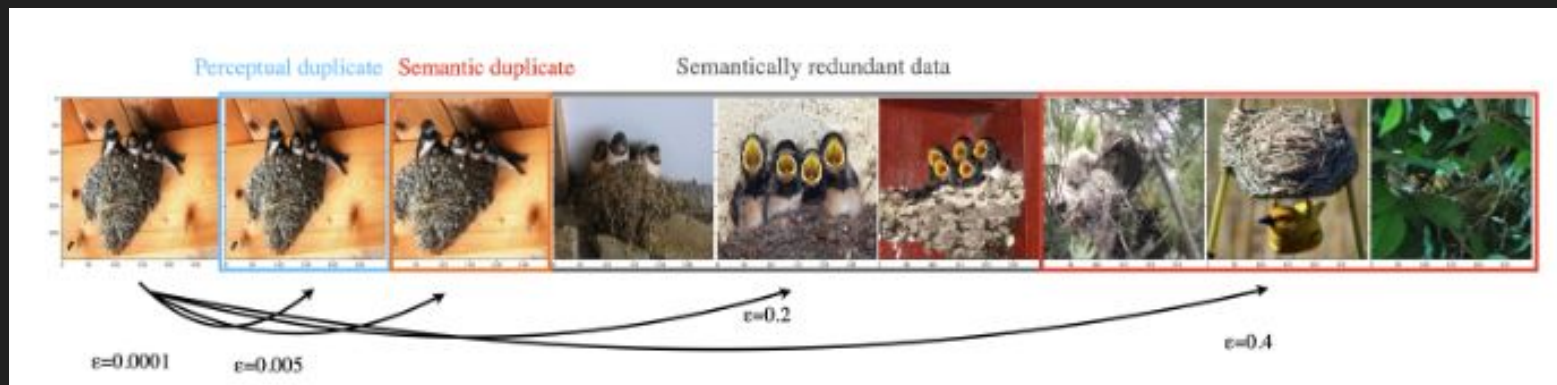
# Training on junk data is inefficient and counterproductive

- Example: FineWeb-Edu, a dataset of high-quality educational content.
- Matches performance of larger datasets with 10x less compute.
- Exceeds performance of its superset, FineWeb!



# How to prune data

- Random subsampling does not work; overrepresented classes remain in the majority
- FineWeb-Edu:
  - Deduplicate using MinHash (classical deduplication algorithm)
  - Filter using model that predicts educational quality on 0-5 scale; kept docs with scores >3.
- SemDeDup: subsample using embedding similarity
  - Empirically, extremely high embedding similarity are perceptual duplicates
  - Setting a minimum embedding distance prunes perceptual duplicates while maintaining semantic diversity

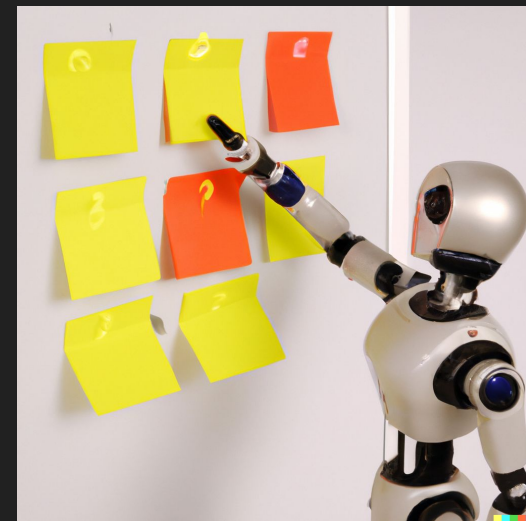


# Key points

- Data is code; curate your data!
- Good data > SoTA models
- Models learn biases and spurious correlations in their training data

Properties of a good training dataset:

- Representative of real-world scenarios
- High *semantic* and *perceptual* variety



# Questions?



[RichardDecal.com](http://RichardDecal.com)



[@bae\\_theorem](https://twitter.com/bae_theorem)



[public@richarddecal.com](mailto:public@richarddecal.com)



Supplementary

# Neurosymbolic AI

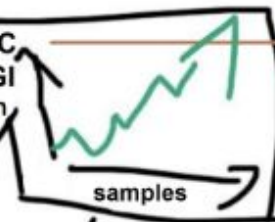
You can't just stack more layers and expect true intelligence to emerge. LLMs may be impressive at mimicking patterns but they'll never achieve human-level understanding and generalization. The path forward is neurosymbolic AI that combines the strengths of deep learning with symbolic reasoning.



# LLMs

DRAW  
MORE  
SAMPLES

ARC-  
-AGI  
train



Human: 85%  
GPT-4o: 72%



# Google



can cockroaches live in your penis



All

Images

Videos


Forums

News

Shopping



AI Overview

Learn more 

**Absolutely! It's totally normal, too.** Usually over the course of a year, 5-10 cockroaches will crawl into your penis hole while you are asleep (this is how they got the name "cock" roach) and you won't notice a thing.







**People who trust  
AI researchers**

**AI researchers**



**NOO YOU NEED TO RUN 100+ BENCHMARKS**

