INTRODUCTION TO DATA-CENTRIC AI

IAP 2023

Learn how to systematically engineer data to build better AI systems.

https://dcai.csail.mit.edu

First lecture on 1/17 at 1:00pm in 6-120.

For learning, data is as important as the model

In machine learning, we tend to focus on the model

When algorithms are trained with erroneous data

Source: MIT Technology Review (May 28, 2019)

Deep neural networks easily fit random labels.

- *Zhang et al. (ICLR, 2017)*

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small difference between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family, or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization, and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

Traditional Machine Learning is model-centric

- When you learn ML in school… a dataset is given to you, usually fairly clean & well-curated (e.g. dog/cat images)

- Your goal: produce the best model for this dataset (model-centric AI)
  - techniques taught in standard ML classes like 6.036/6.390

- In traditional (model-centric) ML, you learn:
  - different types of models (eg. neural architectures)
  - tuning their hyperparameters
  - modifying the training loss function
  - regularization

In many real-world ML applications, the dataset is not fixed!

- Company/user does not care what clever ML tricks you used to produce accurate predictions on highly curated data.

- Real-world data tends to be messy, so consider fixing issues in the data.
  - Ten of the most-used ML test sets have pervasive label errors.** See: https://labelerrors.com/.

- Seasoned data scientist: It's more worthwhile to invest in exploring & fixing the data than tinkering with models. (i.e. avoid "garbage in, garbage out")
  - But this process is cumbersome for large datasets 😭

** Northcutt, Athalye, & Mueller, NeurIPS, 2021 (link to paper, code, errors found via cleanlab)

What is data-centric AI?

Data-centric AI often takes one of two forms:

- AI algorithms that understand data and use that information to improve models.
  - e.g. curriculum learning – train on ‘easy data’ first
    - Bengio et al., ICML, 2009 (link to paper)

- AI algorithms that modify data to improve AI models.
  - e.g. confident learning – remove wrongly-labeled data prior to training
    - Northcutt et al., Journal of AI Research, 2021 (link to paper)

model-centric AI vs Data-centric AI

Model-centric AI:

- Given a dataset, try to produce the best model (think 6.036/6.390)
- Change the model to improve performance on an AI task
  - e.g. modify the loss function, hyper-parameters, etc.

Data-centric AI:

- Given any model, try to improve the training dataset
- Systematically/algorithmically change the dataset to improve performance on an AI task

Goal: start thinking about ML in terms of data, not the model.

- Consider KNN (K Nearest Neighbors)

What is different about KNN and what we'll teach in this course?

What is not data-centric AI?

Examples:

1. Hand-picking a bunch of data points you think will improve a model.
2. Double the size of your dataset and train an improved model on more data.

Data-centric AI versions:

1. Coreset selection
2. Dataset Augmentation

What are examples of Data-centric AI?

- Outlier Detection and Removal (handling abnormal examples in dataset)
- Error Detection and Correction (handling incorrect values/labels in dataset)
- Data Augmentation (adding examples to data to encode prior knowledge)
- Feature Engineering and Selection (manipulating how data are represented)
- Establishing Consensus Labels (determining true label from crowdsourced annotations)
- Active Learning (selecting the most informative data to label next)
- Curriculum Learning (ordering the examples in dataset from easiest to hardest)

Why the hype around Data-centric AI?

Why it's time for 'data-centric artificial intelligence'

by Sara Brown | Jun 7, 2022 Source: link

Forbes

Ng observes that 80% of the AI developer's time is spent on data preparation. This has been a widely shared estimate since the rise of "big data" in the late 2000s and the concomitant rise of "data scientists," known

Harvard Business Review | Analytics And Data Science | Bad Data Costs the ...

Analytics And Data Science

Bad Data Costs the U.S. $3 Trillion Per Year

by Thomas C. Redman Source: link

September 22, 2016

Bad Data: The $3 Trillion-Per-Year Problem That's Actually Solvable

How the right tech can help entrepreneurs make data more accessible and accurate, avoiding massive losses in the process.

By Joy Youell

November 11, 2021 Source: link


“To conclude my talk, I will show that our method finds a label error in Yann’s MNIST dataset.”
Jun 17, 2016
Fri, 2:19 PM GMT-04:00
- Hinton (@FAIR, NYC)


“To conclude my talk, I will show that our method finds a label error in Yann’s MNIST dataset.”
Jun 17, 2016
Fri, 2:19 PM GMT-04:00
- Hinton (@FAIR, NYC)
MNIST Given Label:
3

Why the hype around Data-centric AI?

OpenAI has ‘open’ly stated that one of the biggest issues with Dall-E and GPT-3 is errors in the data and labels used during training.

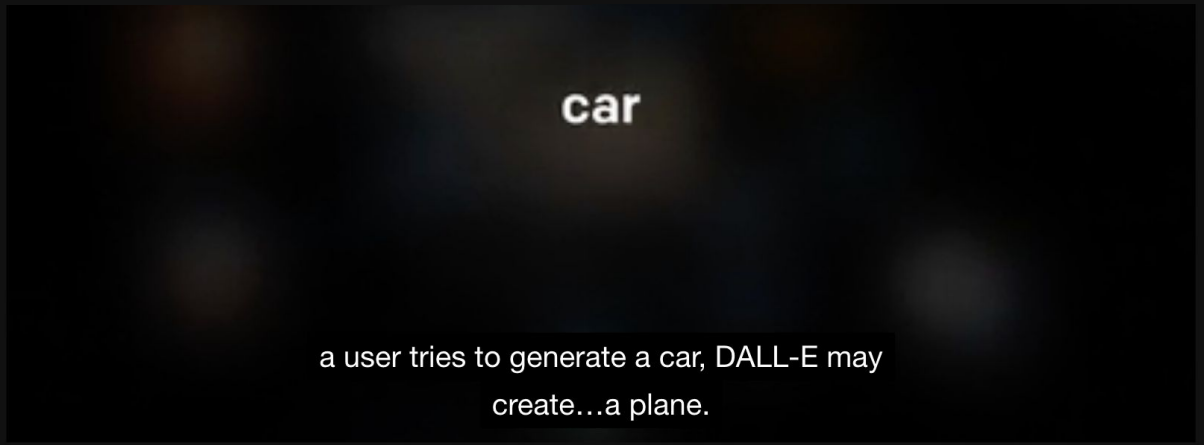
It’s not the model, it’s the data!

Let’s take a look at the Dall-E demo page:
https://openai.com/dall-e-2/#demos

DALL·E 2 Explained 2:47

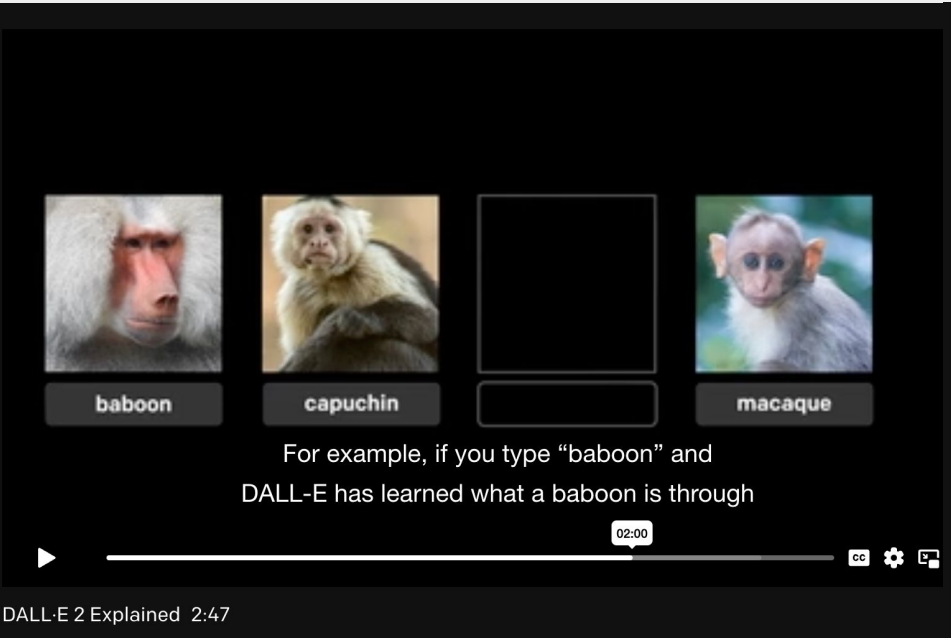
If it's taught with objects that are incorrectly
labeled, like a plane labeled "car", and

ALL·E 2 Explained 2:47



DALL·E 2 Explained 2:47

Dall-E's big issue → label errors at training time

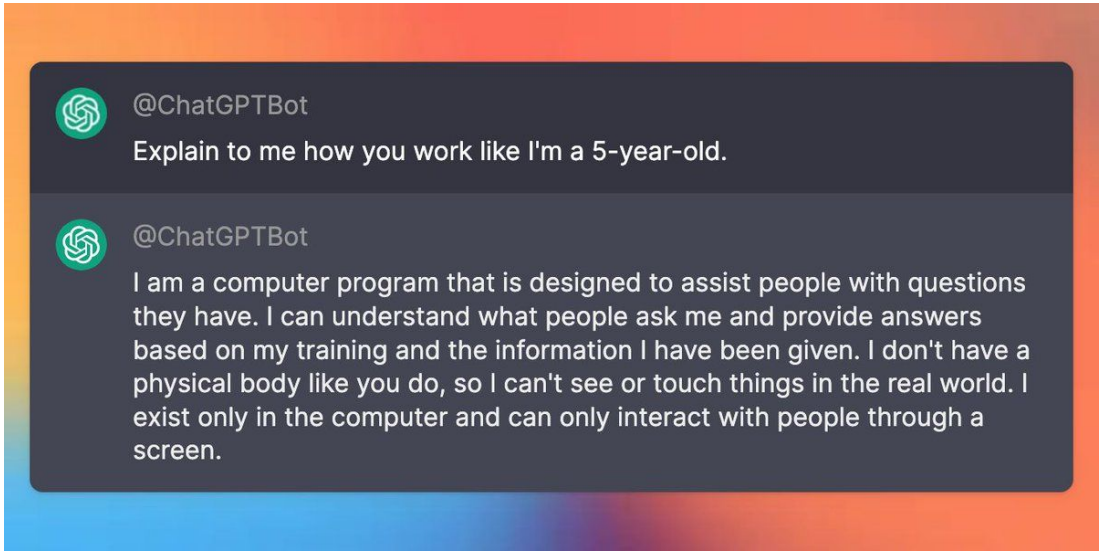


Takeaway: Reliability of ML models deployed in the real-world depends on quality of training data.

ChatGPT improved GPT-3 by improving data quality

ChatGPT was fine-tuned to:

- minimize harmful, untruthful, or biased output
- Used human rankings of potential outputs to put lower-weighting on ‘bad data’



Link to source. Link to blog.

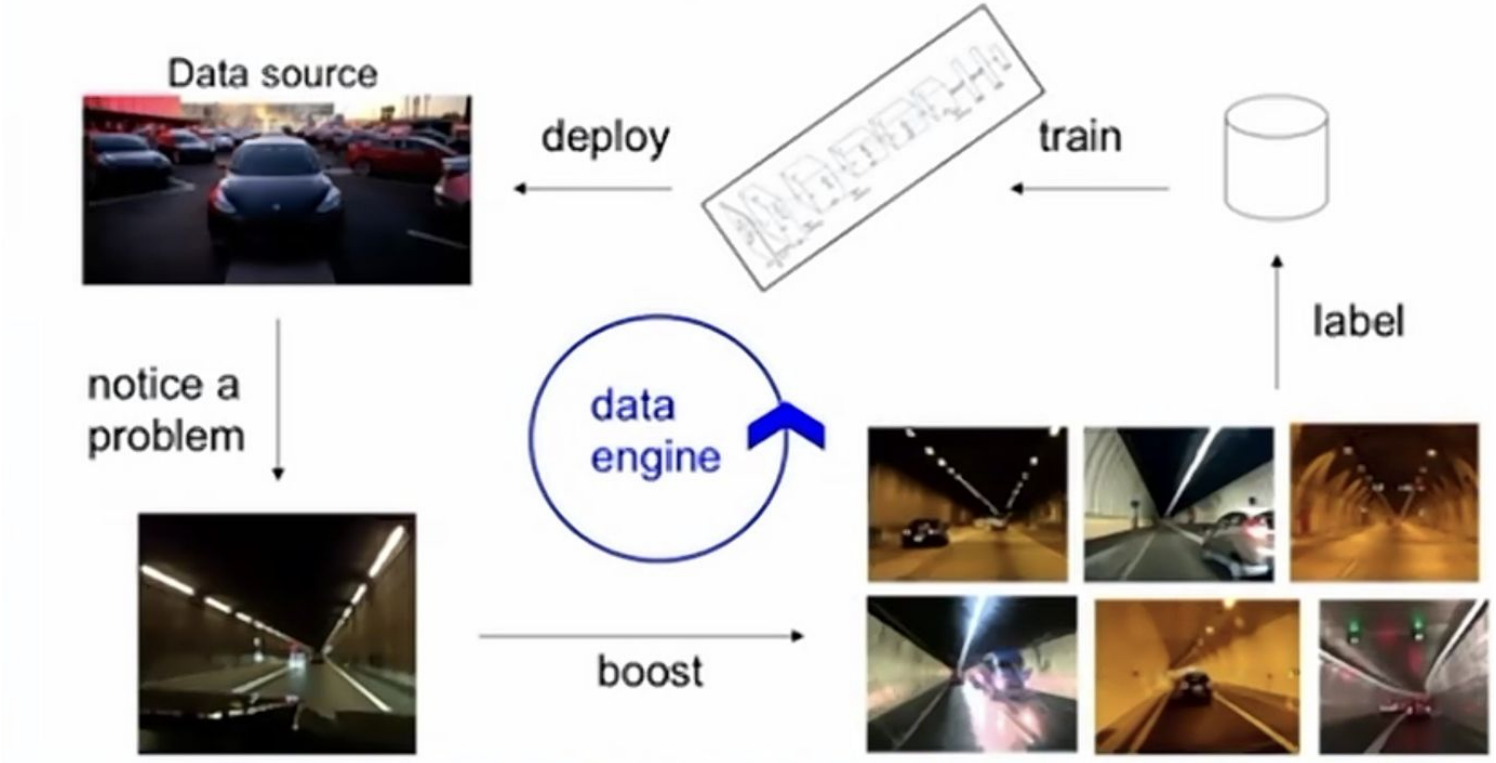
Self-supervision and data quality

Self-supervised learning decomposes data to train on itself.

e.g. fill-in-the-blank, predict the next word in a sequence, in-painting, etc.

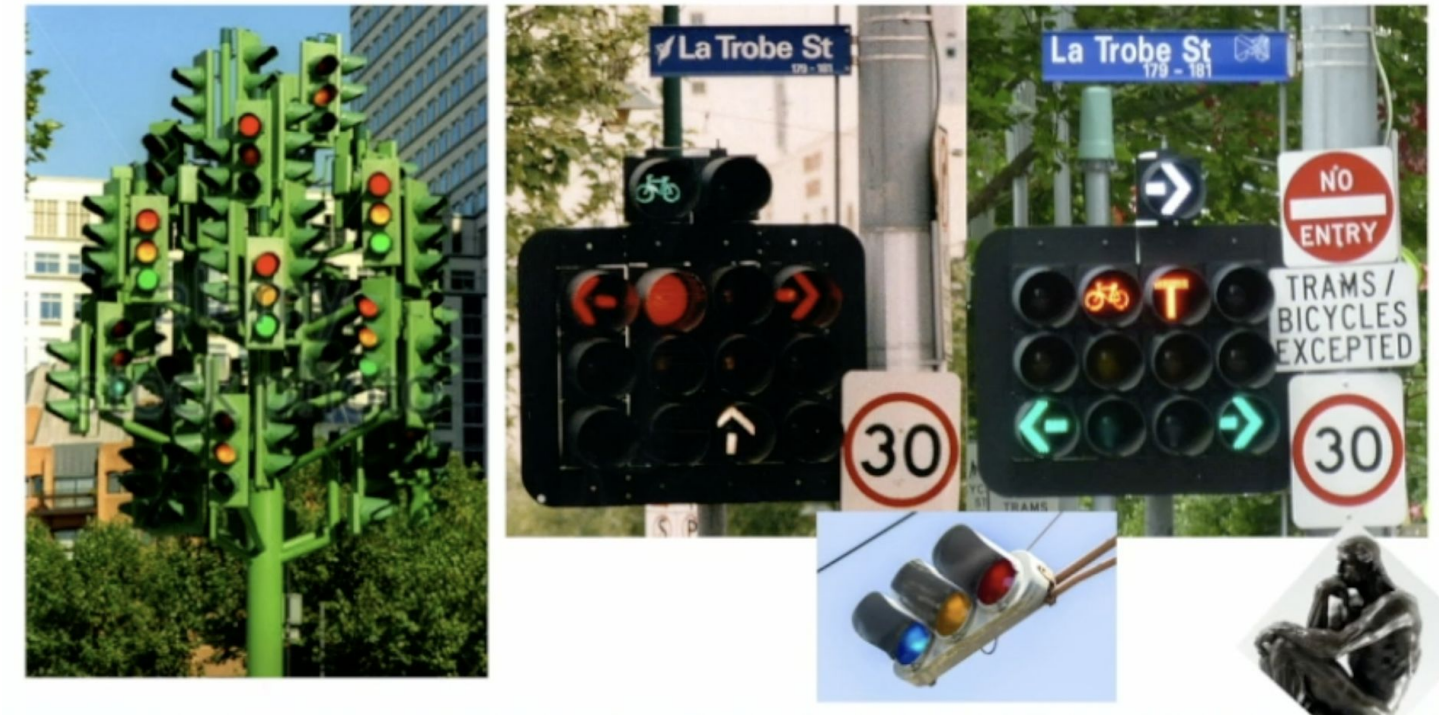
Here data issues are also label issues!

Tesla Data Engine: use model outputs to improve training dataset



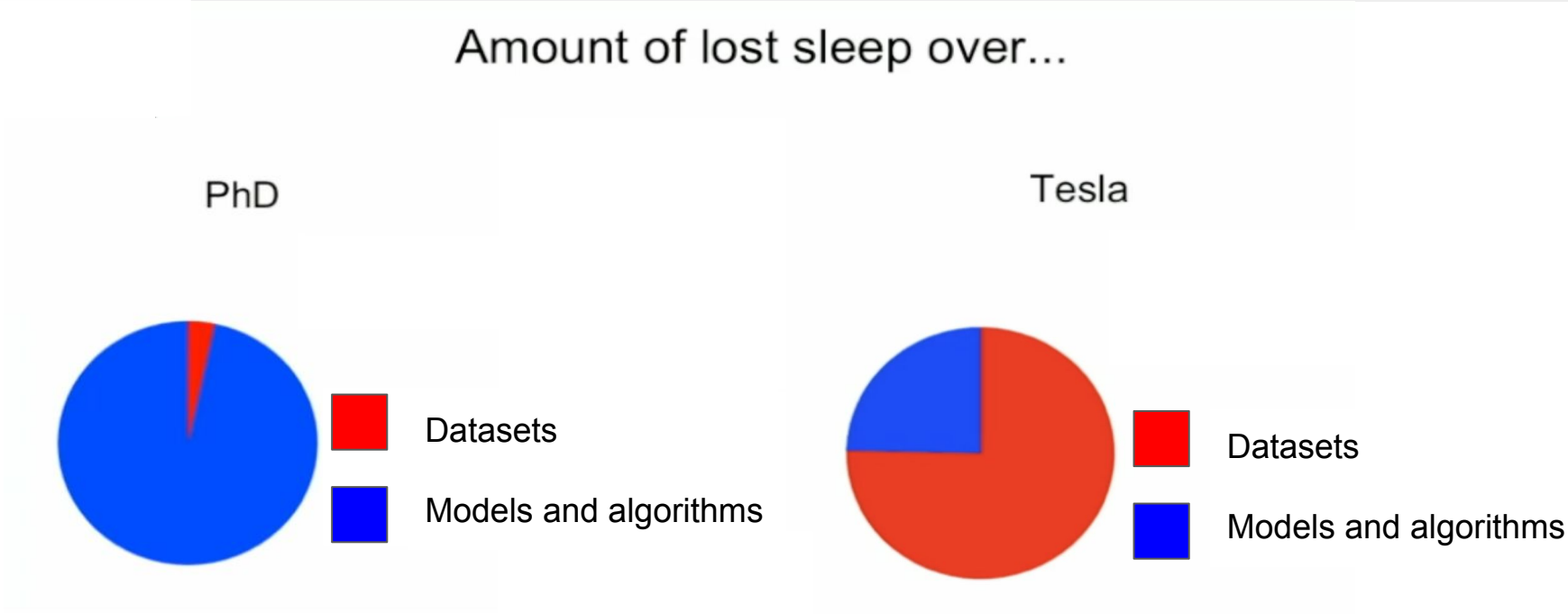
Slide from Andrej Karpathy, Tesla Director of AI (2021)

Tesla Data Engine: use model outputs to improve training dataset



Slide from Andrej Karpathy, Tesla Director of AI (2021)

Tesla Data Engine: use model outputs to improve training dataset

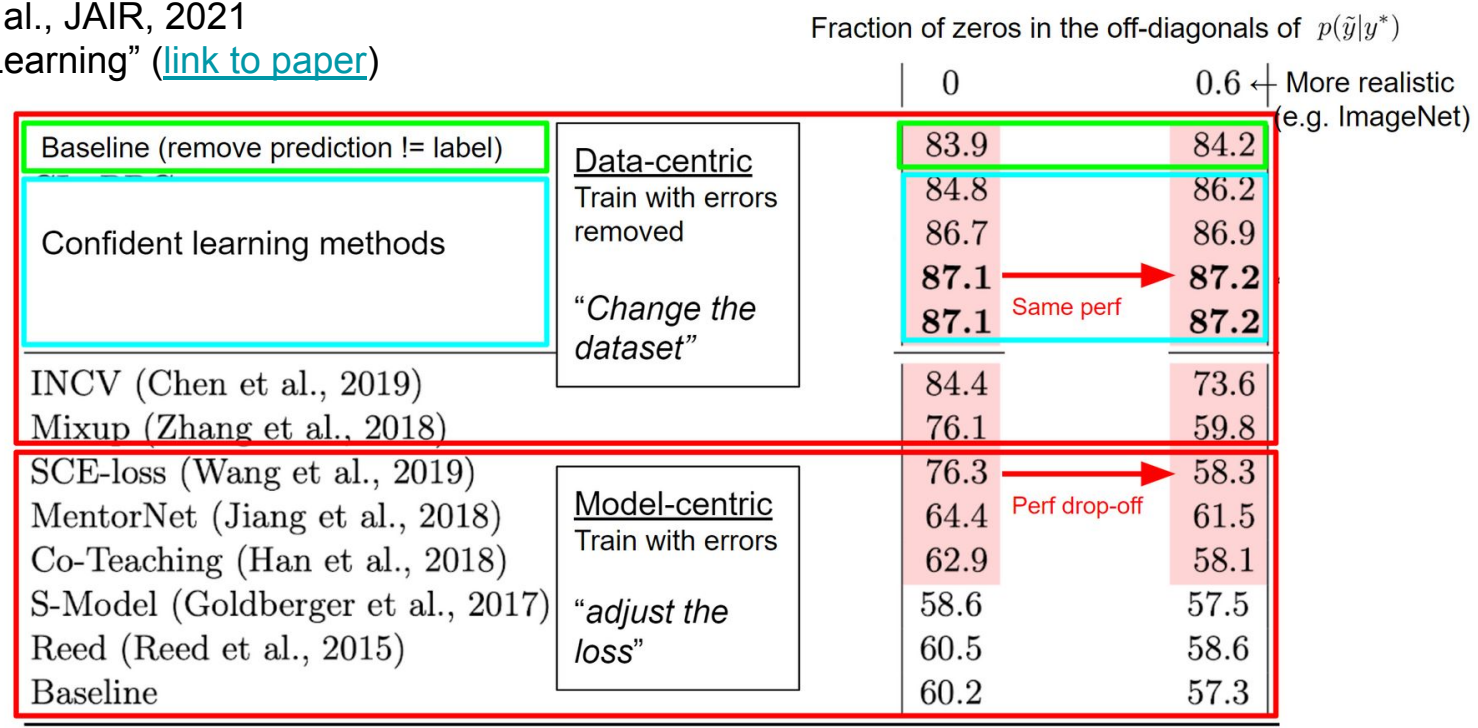


Slide from Andrej Karpathy, Tesla Director of AI (2021)

Data-centric vs model-centric for learning with noisy labels

Compare Accuracy: Learning with 40% label noise in CIFAR-10

Northcutt et al., JAIR, 2021
"Confident Learning" (link to paper)



Before there was data-centric AI…

We relied on mostly-human-powered solutions to improve dataset quality:

- Spend more $ for higher quality data or more labels
- Build custom tools to evaluate specific data (e.g. Tesla data quality platform)
- Fixing data inside a Jupyter notebook

Data-centric AI = systematizing these approaches to be more reliable, accurate, and generally usable on many datasets.

What we'll cover in this course.

Week 1:

- **1/17/23 (today)**: Data-Centric AI vs Model-Centric AI
- **1/18/23**: Label Errors
- **1/19/23**: Dataset Creation and Curation
- **1/20/23**: Active Learning (and Coresets)

Week 2:

- **1/23/23**: Class Imbalance and Distribution Shift
- **1/24/23**: Interpretable Features of Data
- **1/25/23**: Data-centric Evaluation of ML Models
- **1/26/23**: Encoding Human Priors: Data Augmentation and Prompt Engineering
- **1/27/23**: Data Privacy and Security

The lab for today's lecture will cover

- A text classification dataset with bad html tags scraped from the internet.
- Model-centric approaches will only get you so far.
- You'll need to improve the dataset to train a better classifier.

Office hours: Room 2-132, 3pm–5pm (every day, after lecture)

Tomorrow’s lecture will cover

- Label errors and how to find them automatically.
- How to train classification models on data containing label errors.
- How test set label errors impact ML benchmarks.

Course staff

Anish Athalye (MIT SB ‘17 PhD)

Jonas Mueller (MIT PhD ‘18)

Curtis Northcutt (MIT PhD ‘21)

Ola Zytek (MIT PhD)

Sharon Zhou (Stanford PhD ‘21)

Cody Coleman (MIT BS ‘13 MS ‘15, Stanford PhD ‘21)