

ICME Summer Workshops 2021

Intermediate Topics in Machine Learning and Deep Learning



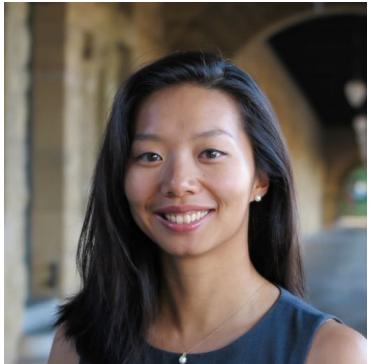
Session 2.1: Transfer Learning

Monday, August 16, 9:30–11:00 AM

Instructor: Sherrie Wang

icme-workshops.github.io/intermediate-ml

About me



Hello
my name is

Sherrie Wang



@sherwang

- **PhD graduate** of ICME (2021)
- **Research focus:** Machine learning methods for remote sensing and applications in sustainability
- **Relevant courses:**
 - CS 221 (Artificial Intelligence)
 - CS 229 (Machine Learning)
 - CS 228 (Probabilistic Graphical Models)
 - CS 230 (Deep Learning)
 - CS 231n (Convolutional Neural Networks)
 - CS 236 (Generative Adversarial Networks)
 - CS 330 (Deep Multi-Task and Meta Learning)
- **Relevant teaching:** CME 250 (Introduction to Machine Learning), Summer Workshop 2019-20

Learning objectives

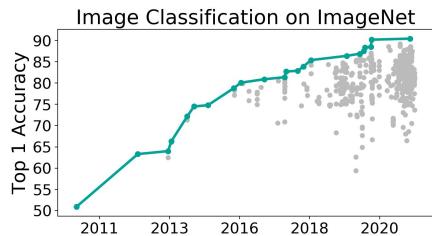
This course is meant to be a continuation of the Introduction to Machine Learning and Introduction to Deep Learning workshops.

Session 2 assumes knowledge of deep learning basics, like fully connected neural networks, CNNs, RNNs, neural network training, hyperparameter tuning.

Now we move toward more realistic learning scenarios, with a focus on what to do when your datasets don't have millions of labels.

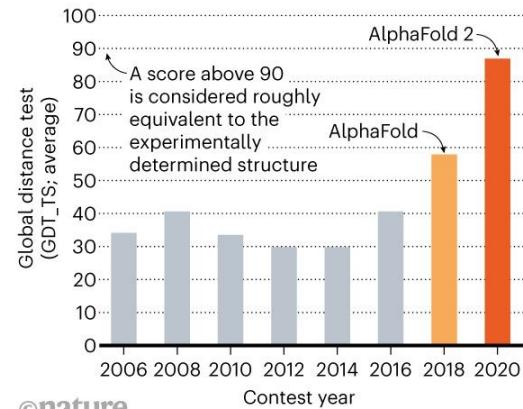
Each topic has a lot of depth on its own: we will provide a survey of transfer learning so that you're familiar with vocabulary and can learn more in the future.

Previously we saw: Deep learning is a game-changer



STRUCTURE SOLVER

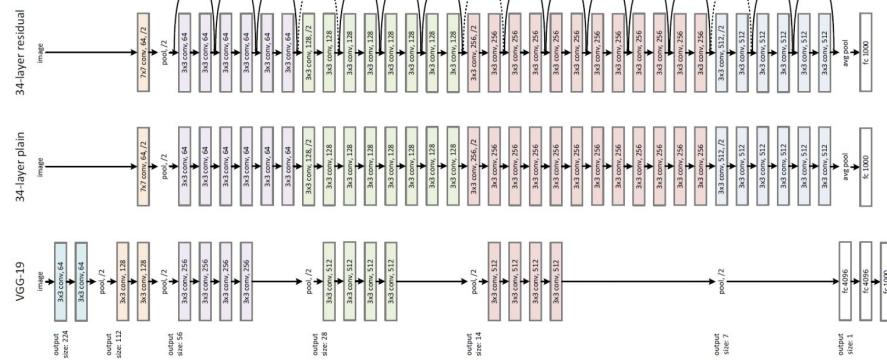
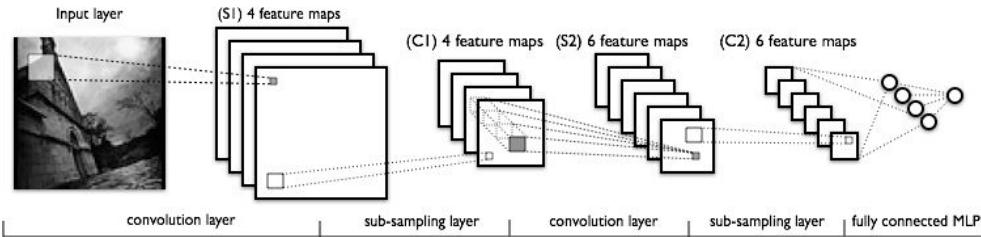
DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



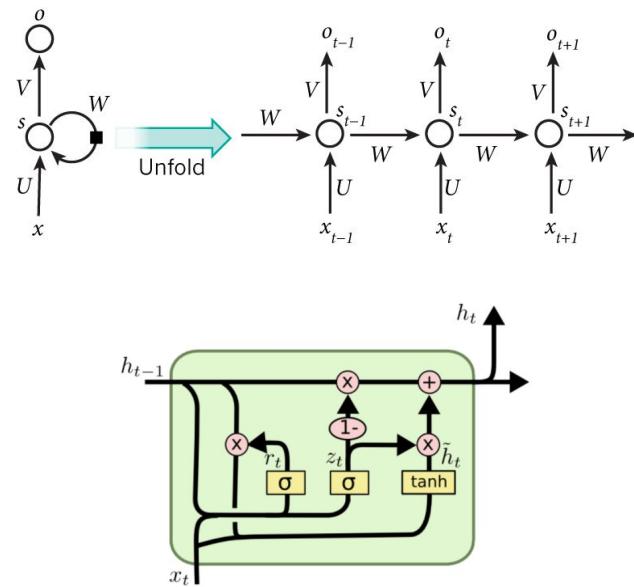
©nature

Previously we saw: Fully connected NNs, CNNs, RNNs

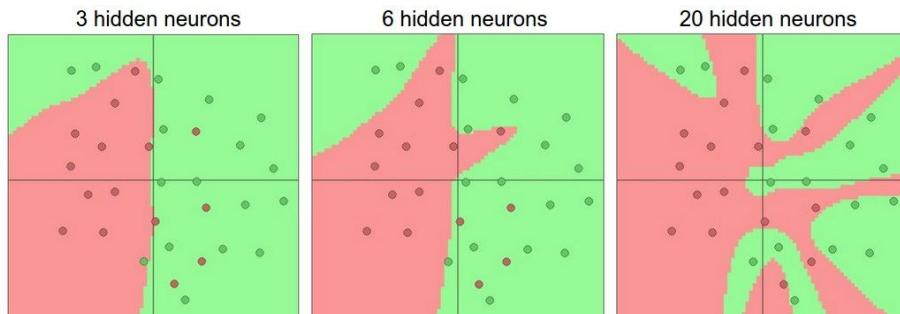
Convolutional neural network



Recurrent neural network

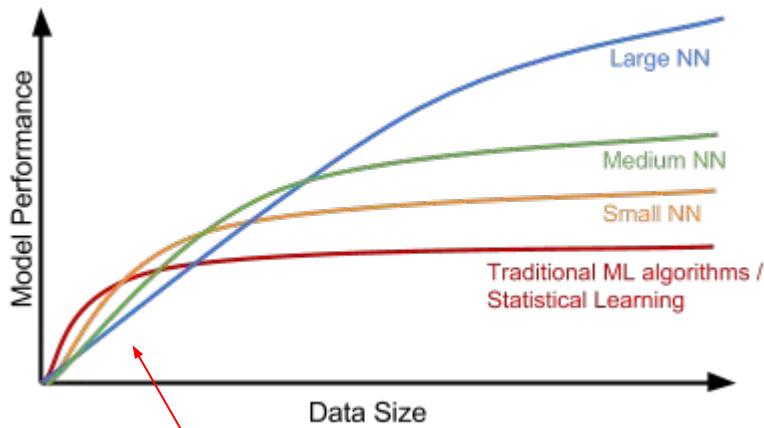


Neural networks can have a lot of learnable parameters



Model	2D-CNN	3D-CNN
	Params	Params
VGG-16	134.7 M	179.1 M
ResNet-18	11.4 M	33.3 M
ResNet-34	21.5 M	63.6 M
ResNet-50	23.9 M	46.4 M
ResNet-101	42.8 M	85.5 M
ResNet-152	58.5 M	117.6 M
DenseNet-121	7.2 M	11.4 M
DenseNet-169	12.8 M	18.8 M

Deep learning needs big data



Notice that at small dataset sizes, could do even worse than traditional ML

Many (most?) tasks don't have *that* much data

- Labels are expensive and time-consuming to generate
- Sometimes labels need experts to generate
- Or there are just a limited number of samples in the world
- Examples:
 - Segmenting an entire road scene takes a long time
 - Segmenting cancer cells needs pathologists to create data
 - Predicting county-level crop yields: there are only 99 counties in, say, Iowa



Cancer cell type	Cancer cell class	Original	Nucleus	Cytoplasm
	Carcinoma in situ			
Abnormal Cells	Mild squamous			
	Moderate squamous			
	Severe squamous			

Hmm... people don't learn everything from scratch

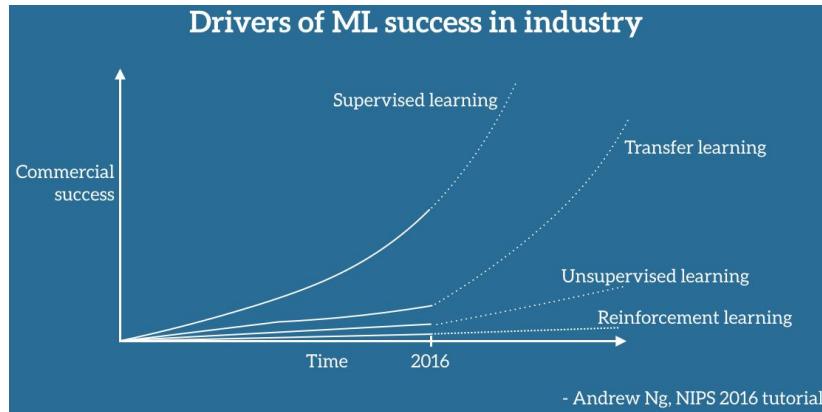


Our world is not made up of a bunch of independent, unrelated tasks. Lots of tasks share skills and knowledge.



What is transfer learning?

A subfield of machine learning. Use knowledge gained while solving one problem and apply it to a different but related problem.



More formally, transfer learning involves the concepts of a **domain** and a **task**.

What is transfer learning?

A domain \mathcal{D} consists of a feature space \mathcal{X} and a marginal probability distribution $P(X)$ over the feature space. Given a domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a task \mathcal{T} consists of a label space \mathcal{Y} and a conditional probability distribution $P(Y|X)$ that is typically learned from the training data $(x_i, y_i) \in X, Y$.

Given a source domain \mathcal{D}_S , a corresponding source task \mathcal{T}_S , and a target domain \mathcal{D}_T and target task \mathcal{T}_T , the objective of transfer learning is to learn the target conditional probability distribution $P(Y_T|X_T)$ in \mathcal{D}_T with the information gained from \mathcal{D}_S and \mathcal{T}_S where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$. In most cases, a limited number of target examples, which is much smaller than the number of labeled source examples, are assumed to be available.

Transfer learning scenarios

$$\mathcal{X}_S \neq \mathcal{X}_T$$

$$P(X_S) \neq P(X_T)$$

$$\mathcal{Y}_S \neq \mathcal{Y}_T$$

$$P(Y_S|X_S) \neq P(Y_T|X_T)$$

$$P(Y_S) \neq P(Y_T)$$

1. Feature spaces of source and target domains differ.	2. Marginal probability distributions of source and target domains differ.	3. Label spaces between the source and target tasks are different. Rarely without #4.	4. Conditional probability distributions of source and target tasks differ.	5. Marginal probability distribution of source and target labels differ.
--	--	---	---	--

Called “domain adaptation”

Sampling bias correction

Example:
Documents are written in 2 different languages.

Example: Documents discuss different topics.

Example:
Documents need to be assigned different target labels.

Example: Documents (e.g. books) get different ratings on different platforms.

Example:
Documents of a label are more prevalent in source vs. target.

Transfer learning scenarios

TABLE 2
Different Settings of Transfer Learning

Transfer Learning Settings	Related Areas	Source Domain Labels	Target Domain Labels	Tasks
<i>Inductive Transfer Learning</i>	Multi-task Learning	Available	Available	Regression, Classification
	Self-taught Learning	Unavailable	Available	Regression, Classification
<i>Transductive Transfer Learning</i>	Domain Adaptation, Sample Selection Bias, Co-variate Shift	Available	Unavailable	Regression, Classification
<i>Unsupervised Transfer Learning</i>		Unavailable	Unavailable	Clustering, Dimensionality Reduction

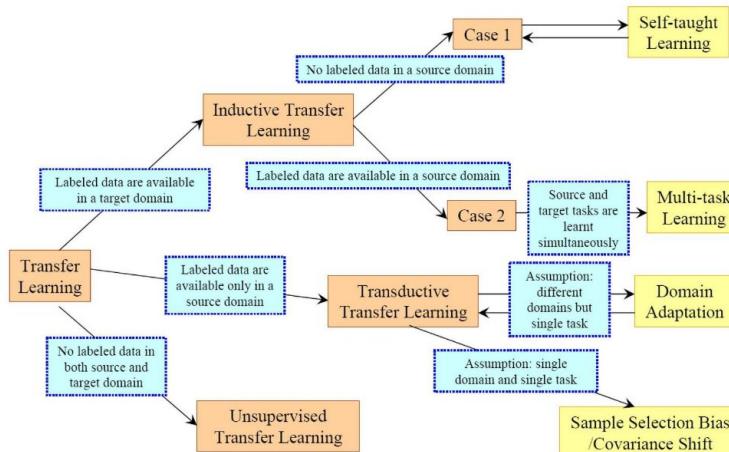


Fig. 2. An Overview of Different Settings of Transfer

Transfer learning scenarios

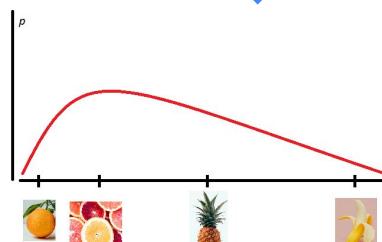
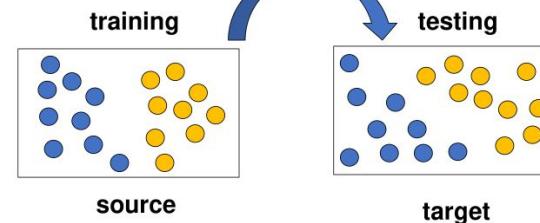
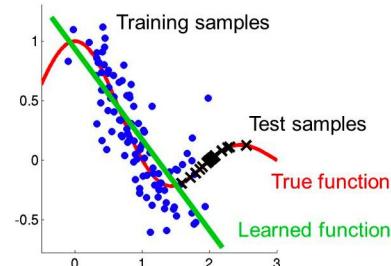
$$\mathcal{X}_S \neq \mathcal{X}_T$$

$$P(X_S) \neq P(X_T)$$

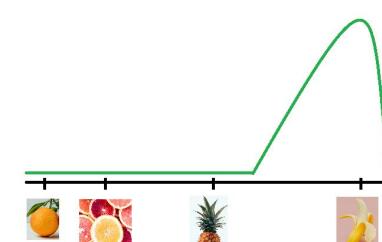
$$\mathcal{Y}_S \neq \mathcal{Y}_T$$

$$P(Y_S|X_S) \neq P(Y_T|X_T)$$

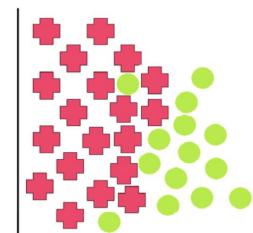
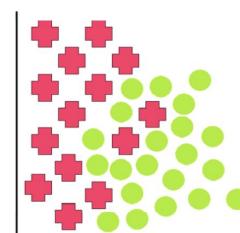
$$P(Y_S) \neq P(Y_T)$$



Distribution the training, validation, and testing data was drawn from



Distribution seen in production



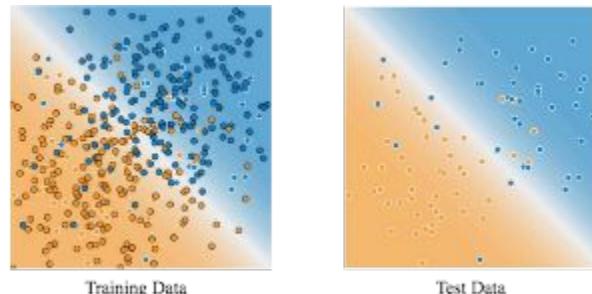
Transfer learning vs. traditional supervised learning

In traditional supervised learning, the assumption is that

$$\mathcal{D}_S = \mathcal{D}_T \quad \text{and} \quad \mathcal{T}_S = \mathcal{T}_T$$

where the source domain is the training set and target is test set.

This allows your learned model to generalize to the test set.

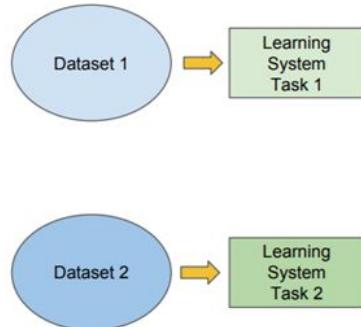


Transfer learning vs. traditional supervised learning

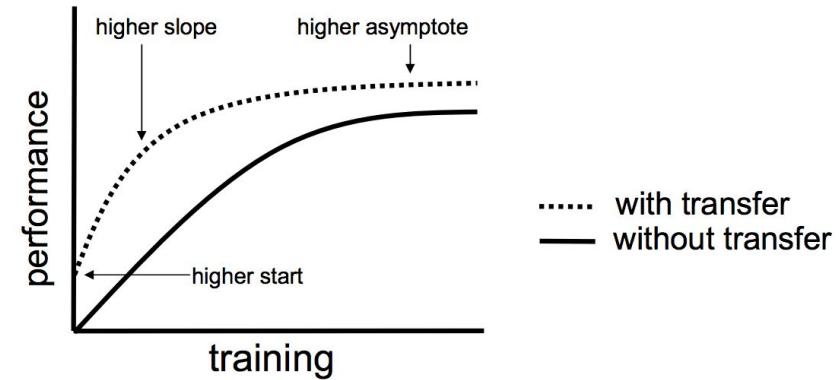
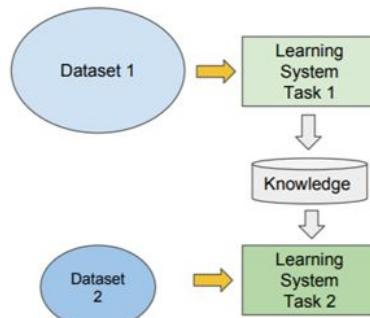
If you have labeled data for your target task, you can always train a neural network from scratch.

If you can find a good source task, however, you often can boost performance.

Traditional supervised learning



Transfer learning



Types of transfer learning

Including but not limited to:

- Direct transfer
- Fine-tuning (sometimes also just “transfer learning”)
- Domain adaptation
- Multi-task learning
- Meta-learning
- Under-sampling, over-sampling, and SMOTE

Direct use of pre-trained models

The simplest strategy: solve a target task by applying a model trained on a source task.

Big corporations often release such models to the public. Examples include ResNet or VGG trained on ImageNet, BERT/XLNet trained on language corpuses, YOLO trained on Pascal VOC.

Also comparatively the least likely to yield good results, because $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

Building detection

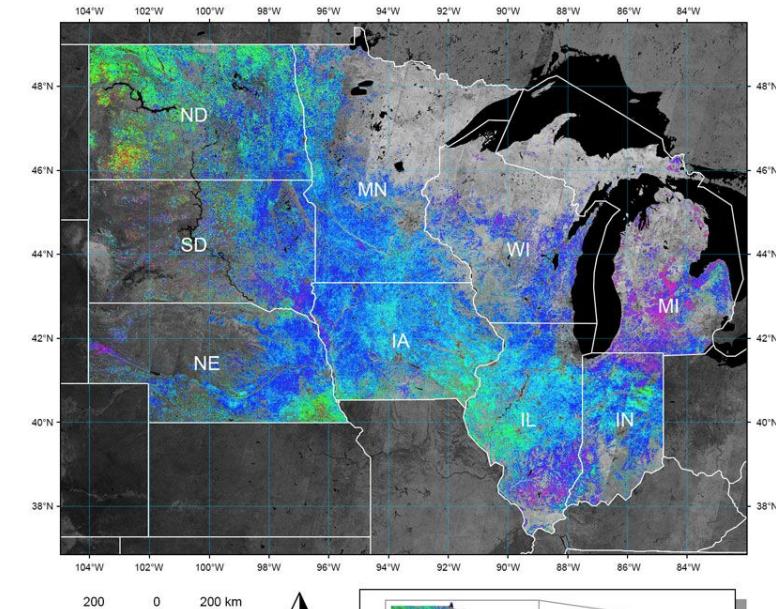
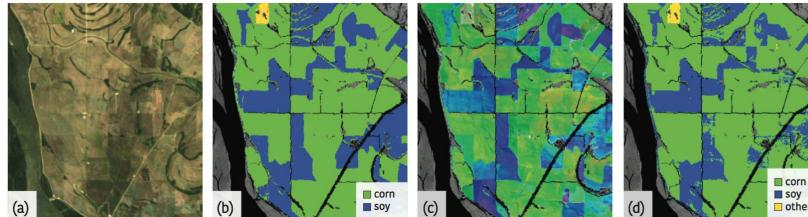


Example: Direct transfer of crop classifier

Goal: Classify crop types from satellite imagery

Challenge: Only have crop type labels in some counties

Solution: Directly transfer a model (random forest)

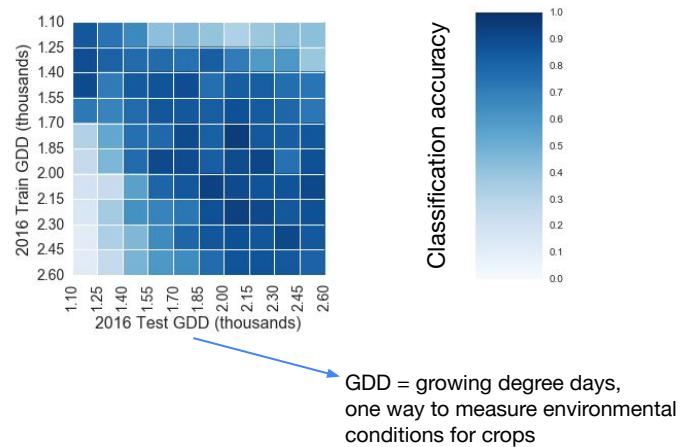


Example: Direct transfer of crop classifier

Use Growing Degree Days (GDD) to measure the climatic conditions of a county.

We found that the more similar two counties were in terms of GDD, the higher the performance of the model when directly applied in the target county.

Random forest transfer performance

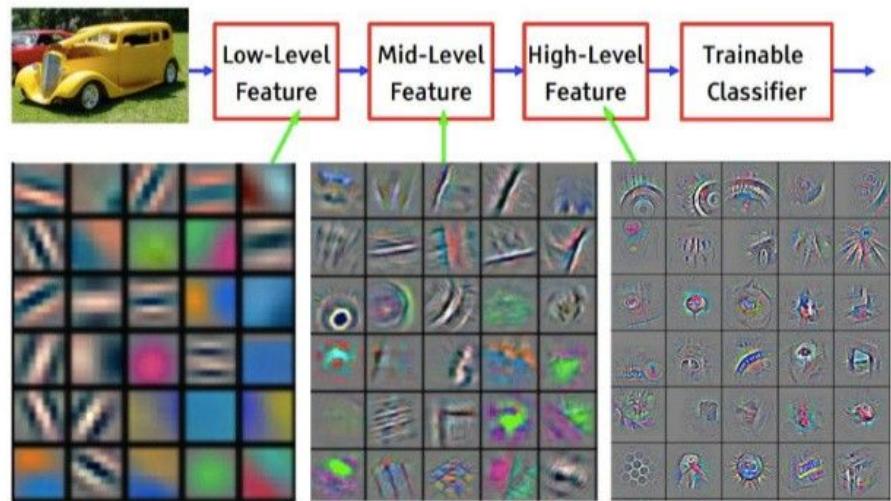


Fine-tuning

Instead of using a pre-trained model end-to-end, we can use the model as a **feature extractor**.

Rationale: Lower-level features in the source task's neural network can generalize to the target task.

“Fine-tuning” refers to making small changes to the neural network.

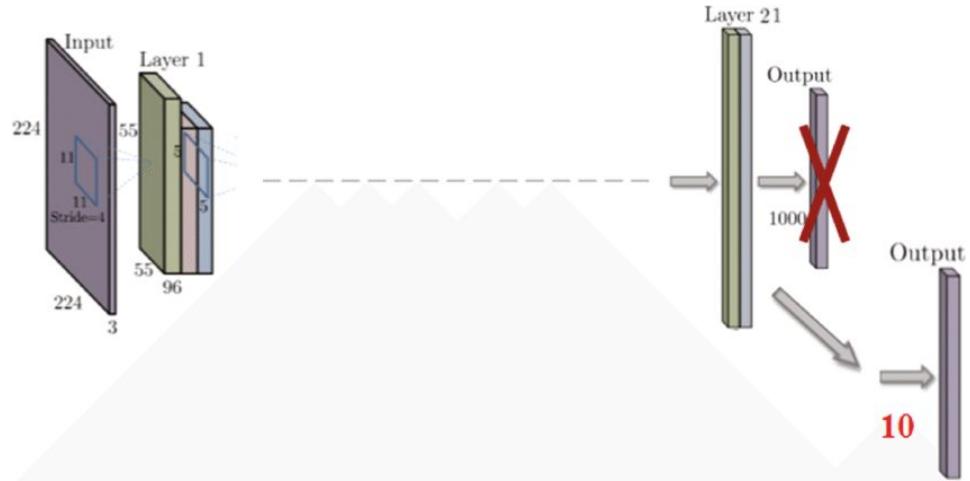


Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

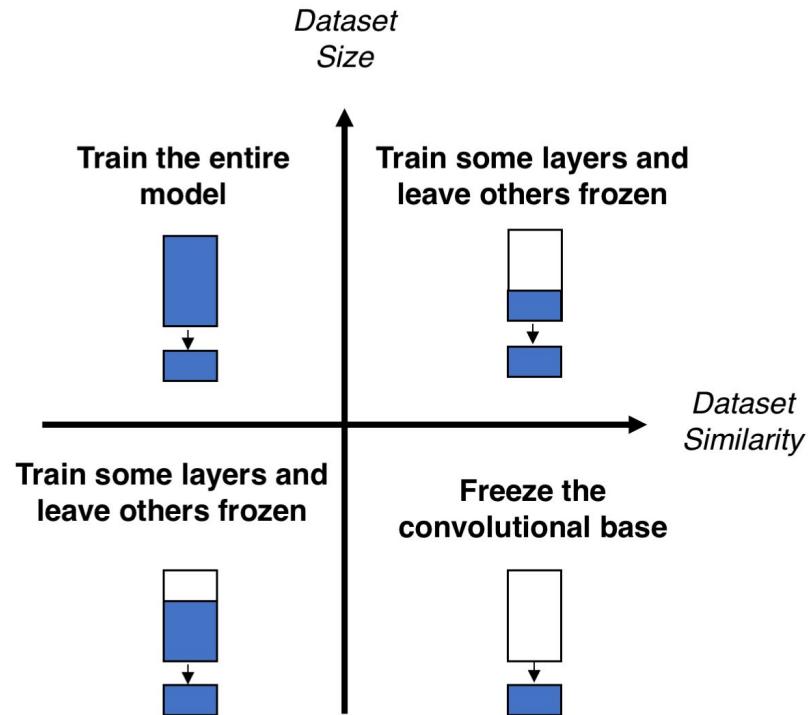
Fine-tuning

At the very least, the last layer (classifier) of the neural network is trained on the target task.

Example: The ImageNet task classifies among 1000 image classes. For a different image classification task, we can remove the last layer, “freeze” all other layers, add on a randomly initialized classifier, and “fine-tune” the last layer on the target task.



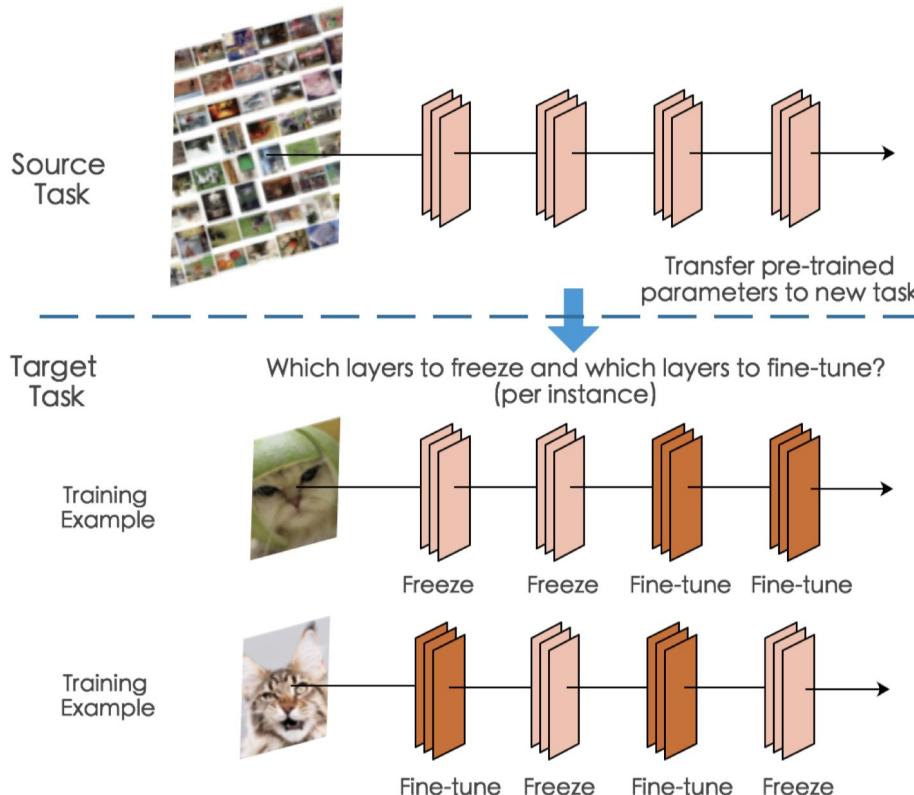
Pre-training and fine-tuning spectrum



Note: training the entire model is like using the pre-trained model as weight initializations.

Note #2: Frozen layers' weights are not updated during backpropagation.

Pre-training and fine-tuning spectrum



IBM and collaborators designed a fine-tuning method called SpotTune that automatically decides which layers of a model should be frozen or fine-tuned. (CVPR 2019)

Example: ImageNet feature transfer

Researchers at Google Brain investigated whether pre-training on ImageNet helps performance on medical imaging datasets (e.g. diabetic retinopathy).

Conclusion: Helps final performance slightly, helps convergence speed a lot.

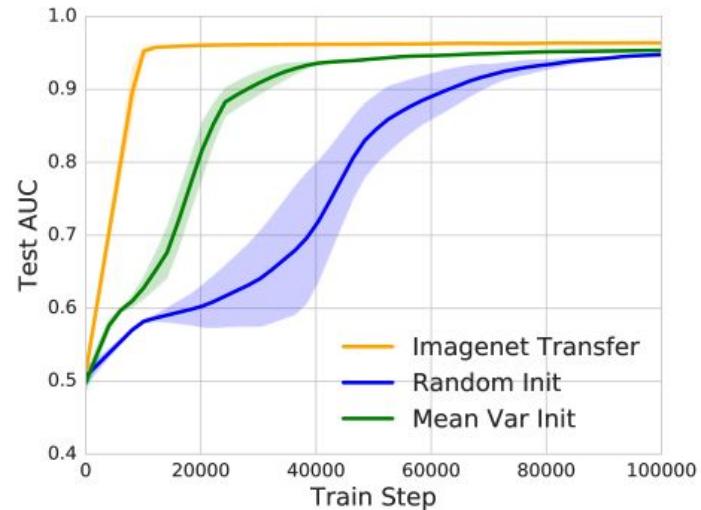


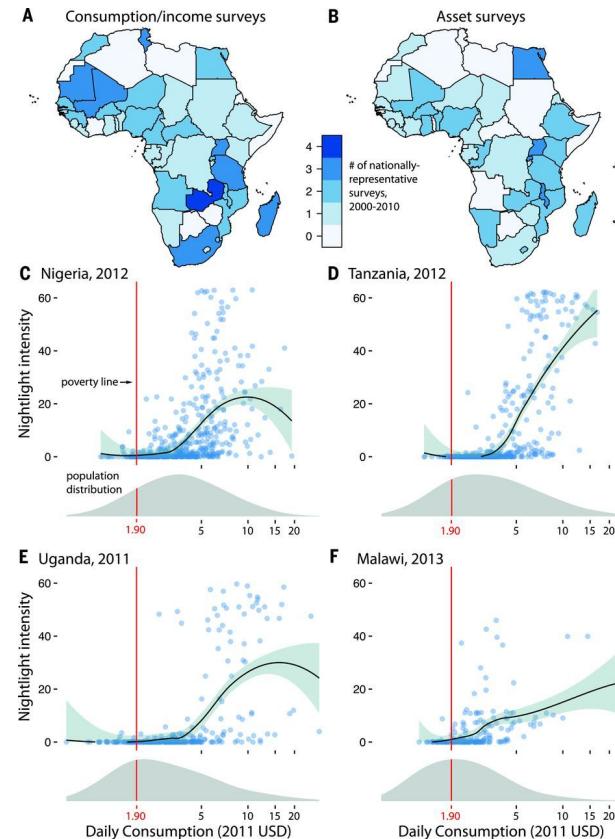
Figure 1: Example images from the Imagenet, the fundus photographs, and the ChestXray14 datasets, respectively.

Example: Transfer learning to predict poverty

Goal: Predict wealth level in a region from satellite imagery.

Challenge: In the world's poorest places, data on wealth is scarce.

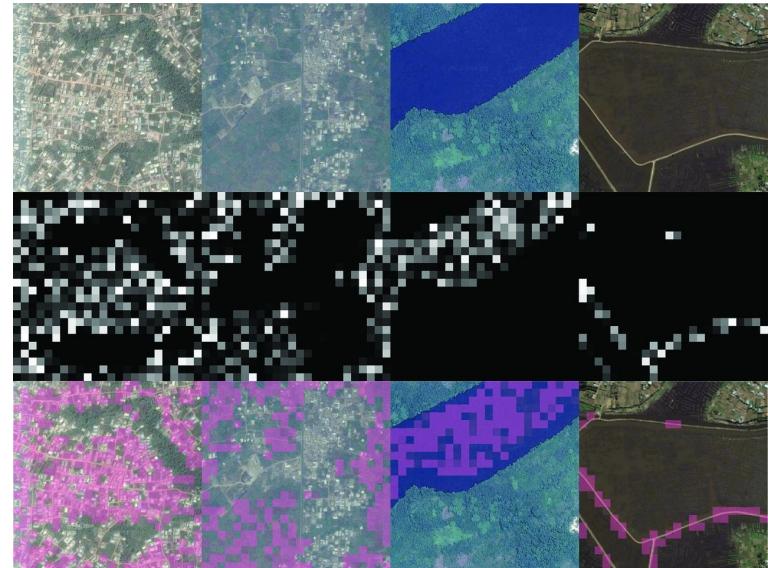
Innovation: Use nighttime lights, which are abundant, as a proxy task (source task).



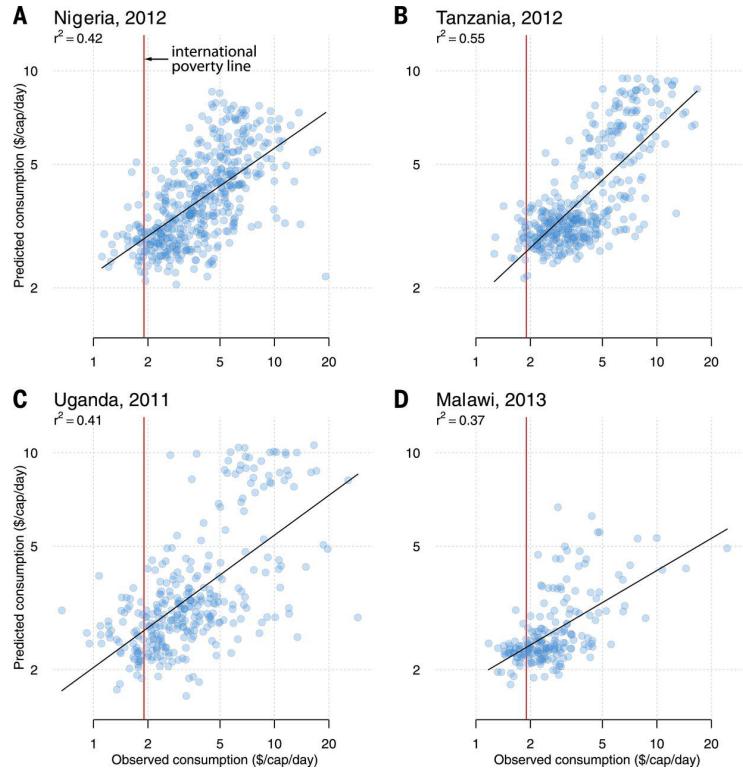
Example: Transfer learning to predict poverty

Features learned through training on nighttime lights capture the presence of buildings, roads, and other features that correlate with wealth.

Authors then removed the last layer of their night-light neural network, froze the previous layers, added a ridge regression, and trained the new classifier to predict wealth.



Example: Transfer learning to predict poverty



Positive, negative, neutral transfer

Note: Transfer learning is not guaranteed to improve performance.

- **Positive transfer:** When learning from one task improves performance on another task
- **Negative transfer:** When learning from one task worsens performance on another task
- **Neutral transfer:** When learning from one task neither improves nor worsens performance on another task

When will transfer learning help?

Answer #1: The more similar source and target tasks and domains are, the more transfer should help. The less data you have in the target task, the more transfer should help.

Answer #2: It's hard to predict exactly and something that you have to try.

Common pre-training datasets

Image classification: ImageNet

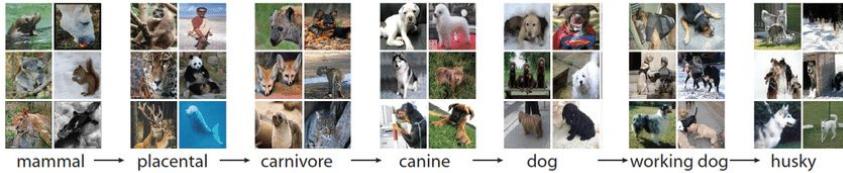
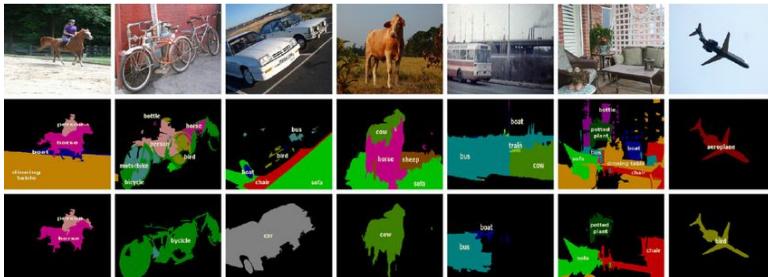


Image segmentation: Pascal VOC, MS-COCO



NLP: BookCorpus, English Wikipedia, WikiText-2, Amazon reviews, Reuters corpus, Tweet sentiments, ...

Check out more datasets:
https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

Dataset examples



Learning from simulations

Simulations can help when gathering labels in the real world is expensive, time-consuming, or too dangerous.

Examples include self-driving cars and robotics.



Domain adaptation

Modify the source domain to bring the distribution of the source closer to that of the target domain (or vice versa), thereby enabling model transfer.

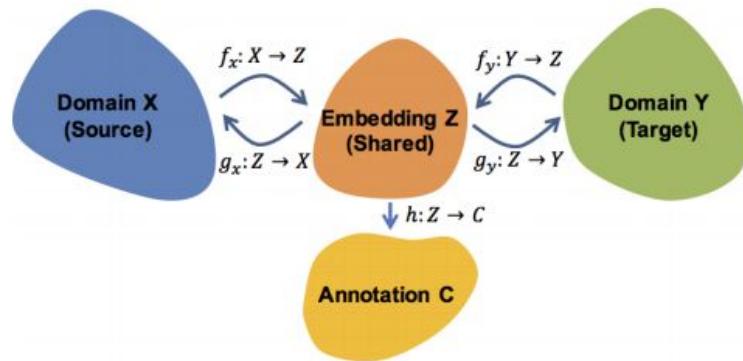
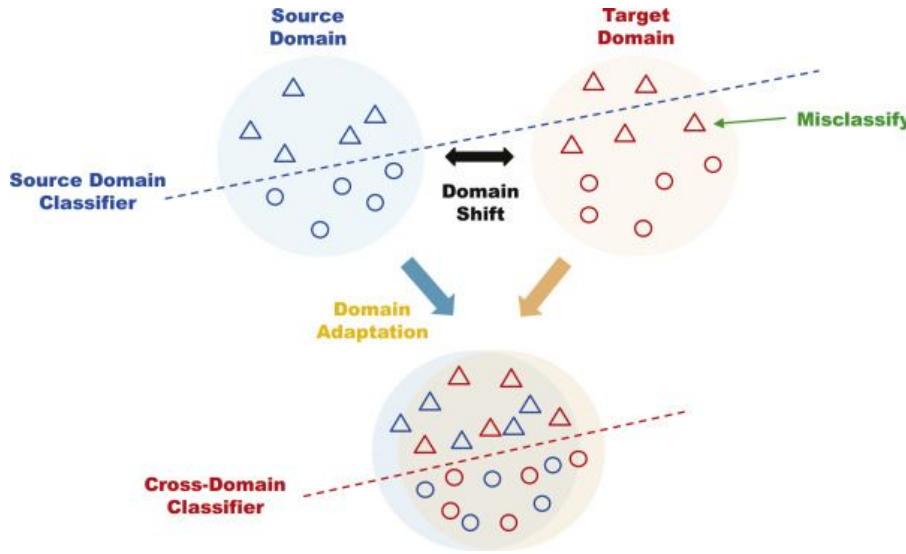


Figure 2. The source, target, annotation, and shared embedding spaces with the corresponding mappings between them.

Example: MNIST \longleftrightarrow SVHN

Both the MNIST and SVHN datasets have digit recognition as the task.

MNIST: Binarized images of handwritten digits.

SVHN: House numbers from Street View.



Source: MNIST



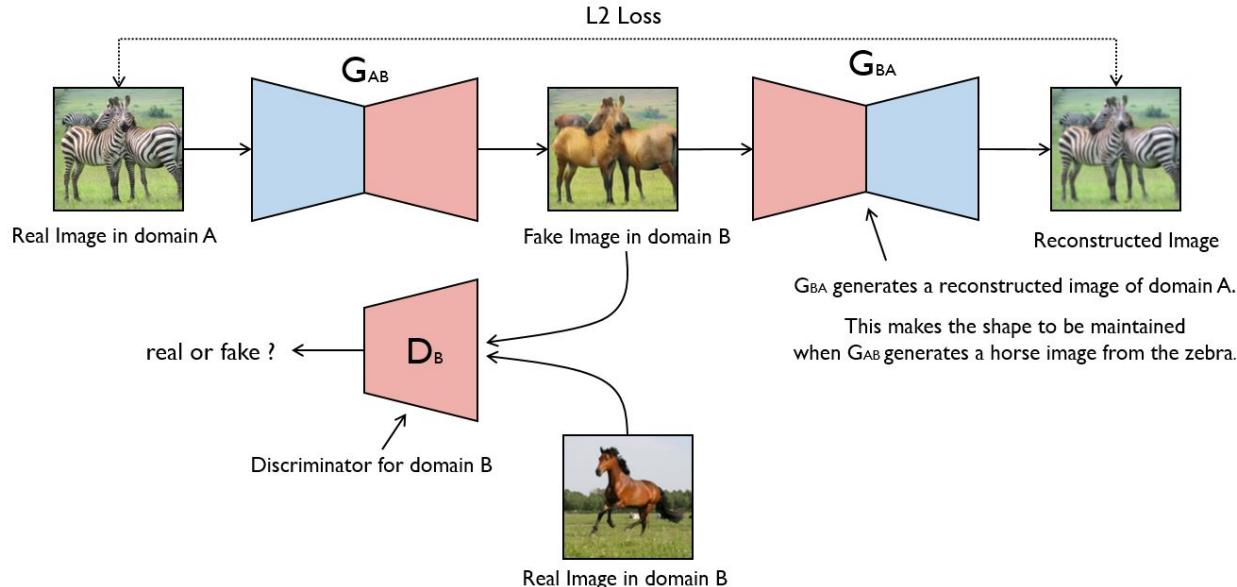
Target: SVHN

Or vice versa!

Example: MNIST \longleftrightarrow SVHN

One way to perform domain adaptation is via adversarial learning.

CycleGAN:



Example: MNIST \longleftrightarrow SVHN

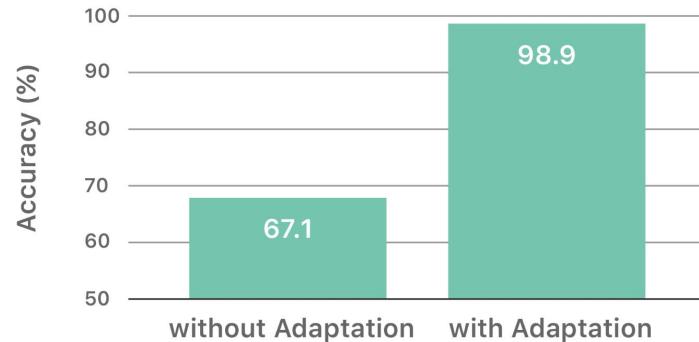


Example: MNIST \longleftrightarrow SVHN

A CNN can achieve 98% accuracy when trained on SVHN.

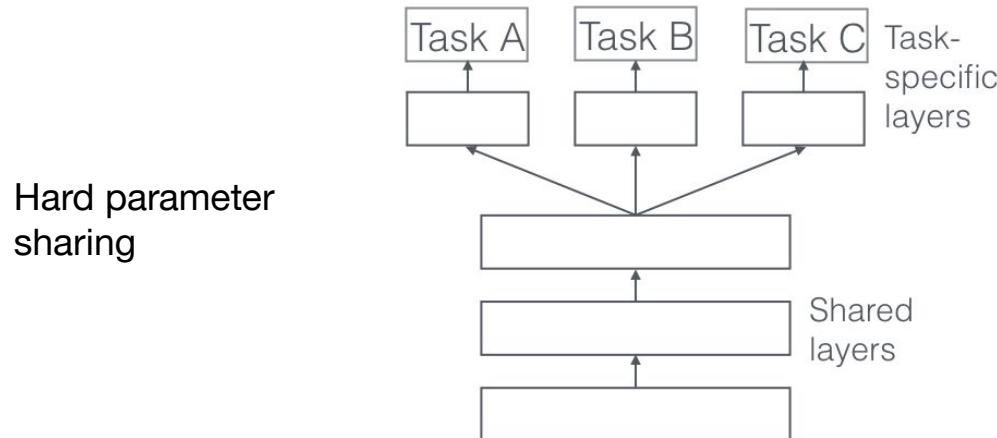
When applied to MNIST, accuracy is 67%.

With domain adaptation, the same CNN's performance becomes 99%.



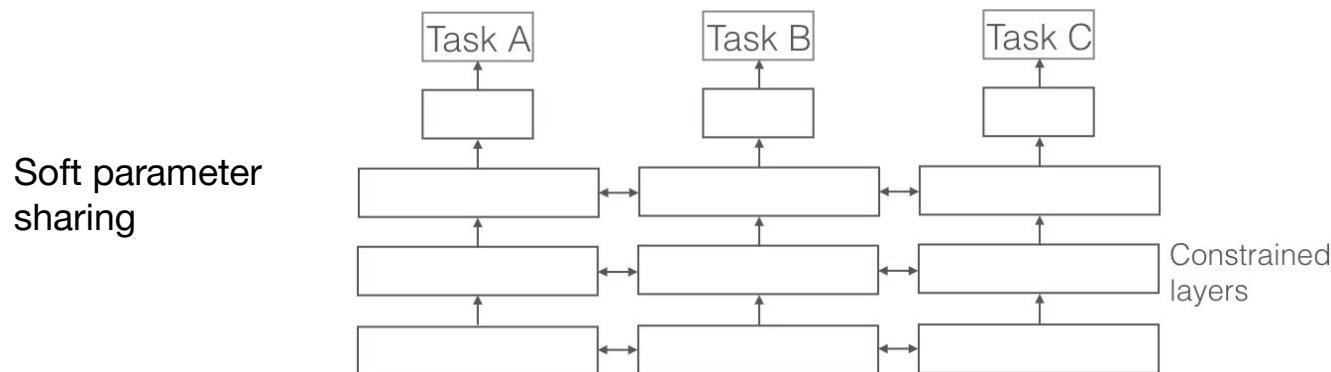
Multi-task learning

Since the knowledge from one task can benefit another, we can train them together, simultaneously, while sharing neural network weights.



Multi-task learning

Since the knowledge from one task can benefit another, we can train them together, simultaneously, while sharing neural network weights.



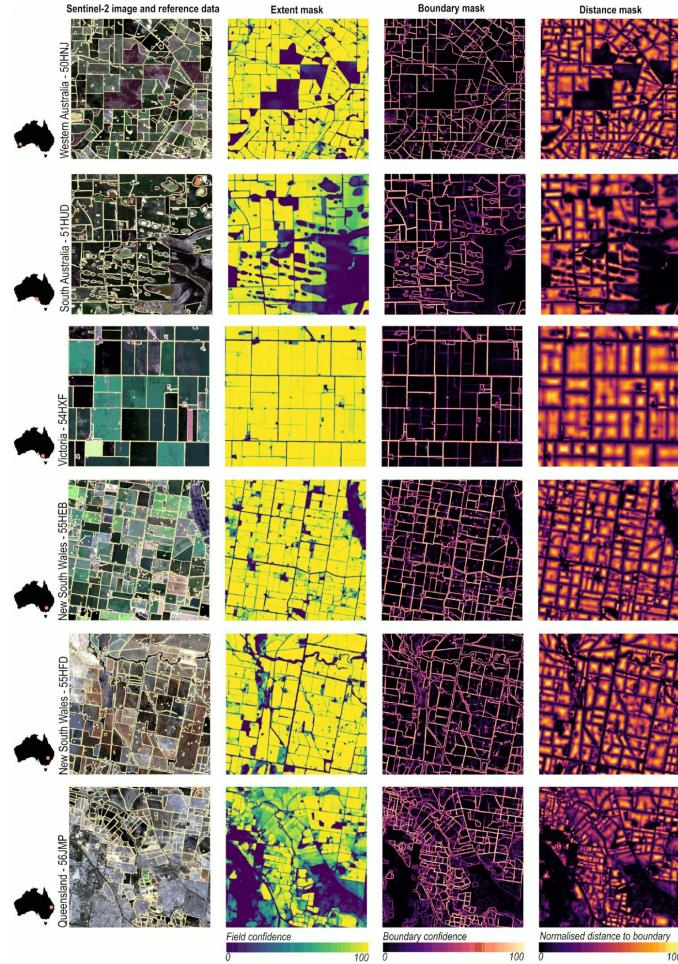
Example: Segmenting crop fields

Often, additional tasks can be devised for the sole purpose of improving performance on the target task.

Goal: Segment fields in satellite imagery.

Challenge: Only training on field boundaries yields unsatisfactory performance.

Innovation: Train on a distance task and image reconstruction task as well.



Meta-learning

Meta-learning aims to create models that can learn new tasks and adapt to new environments quickly (with few training examples); AKA “learn how to learn”

Single task learning

learn and
perform tasks



Pre-training and fine-tuning

learn tasks



refresh task of
interest



Meta-learning

learn to learn tasks



quickly learn
new task



Adapted from <https://meta-world.github.io>

Model-agnostic meta-learning (MAML)

“Model agnostic” because the algorithm can be used to train any model that is normally trained with gradient descent-based methods.

MAML is a type of initialization-based meta-learning algorithm. It tries to find an initialization that will allow for fast adaptation to a new task.

Note: using pre-training and fine-tuning is also using an initialization to improve target task performance. But the pre-trained weights are not explicitly designed to transfer well to new tasks.

Training with regular gradient descent

Pre-training gradient steps

Algorithm 1: Regular Gradient Descent

$p(\mathcal{D})$: distribution over data points;

α : step size hyperparameters;

randomly initialize ϕ ;

repeat

sample $D \sim p(\mathcal{D})$;

evaluate $\mathbf{g} = \nabla \mathcal{L}(f_\phi, D)$;

update parameters $\phi \leftarrow \phi - \alpha \mathbf{g}$;

until *convergence*;

Model-agnostic meta-learning (MAML) algorithm

Algorithm 2: Model-Agnostic Meta-Learning

$p(\mathcal{T})$: distribution over tasks;
 α, β : step size hyperparameters;
randomly initialize θ ;

repeat

- sample batch of tasks $\tau \sim p(\mathcal{T})$;
- foreach** $\tau_i \in \tau$ **do**

 - initialize ϕ_i with θ ;
 - sample $\{D_{\text{support}}, D_{\text{query}}\} \sim p(\tau_i)$;
 - evaluate $\mathbf{g} = \nabla_{\phi_i} \mathcal{L}_{\tau_i}(f_{\phi_i}, D_{\text{support}})$;
 - adapt parameters $\phi_i \leftarrow \phi_i - \alpha \mathbf{g}$;
 - evaluate test loss $\mathcal{L}_{\tau_i}(f_{\phi_i}, D_{\text{query}})$;

- end**
- update $\theta \leftarrow \theta - \beta \sum_{\tau_i \sim p(\tau)} \nabla_{\theta} \mathcal{L}_{\tau_i}(f_{\phi_i}, D_{\text{query}}^{\tau_i})$;

until convergence;

MAML gradient steps

θ : initialization
 ϕ : adaptation

Model-agnostic meta-learning (MAML)

Sine wave prediction experiment:

Train a regression model to predict the full sine curve from $k=5$ or $k=10$ random samples from the curve.

Test on new sinusoidal functions that have not been seen by the model before.

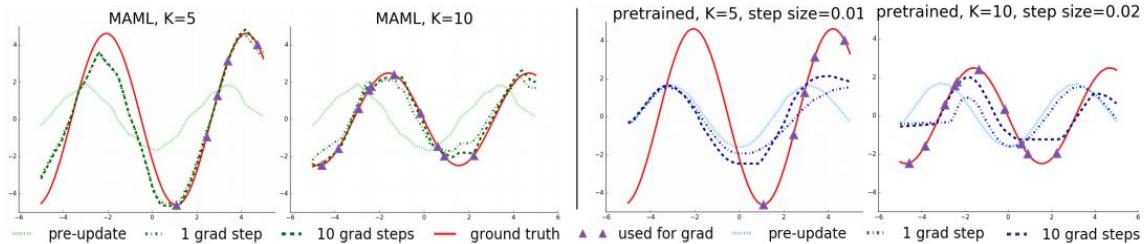


Figure 2. Few-shot adaptation for the simple regression task. Left: Note that MAML is able to estimate parts of the curve where there are no datapoints, indicating that the model has learned about the periodic structure of sine waves. Right: Fine-tuning of a model pretrained on the same distribution of tasks without MAML, with a tuned step size. Due to the often contradictory outputs on the pre-training tasks, this model is unable to recover a suitable representation and fails to extrapolate from the small number of test-time samples.

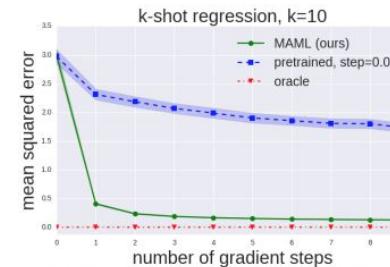


Figure 3. Quantitative sinusoid regression results showing the learning curve at meta test-time. Note that MAML continues to improve with additional gradient steps without overfitting to the extremely small dataset during meta-testing, achieving a loss that is substantially lower than the baseline fine-tuning approach.

Summary

Use transfer learning if:

- You don't have that much labeled data
- You don't have the money or time to train models from scratch
- You can find a source dataset with positive transfer

The type of transfer learning you use will depend on the source and target domain/task. From low training to high training involvement:

- Direct transfer
- Fine-tuning
- Domain adaptation, multi-task learning
- Meta-learning



ICME Summer Workshops 2021

Intermediate Topics in Machine Learning and Deep Learning



Session 2.2: Generating Labels

Monday, August 16, 9:30–11:00 AM

Instructor: Sherrie Wang

icme-workshops.github.io/intermediate-ml

What is data labeling and why is it important?

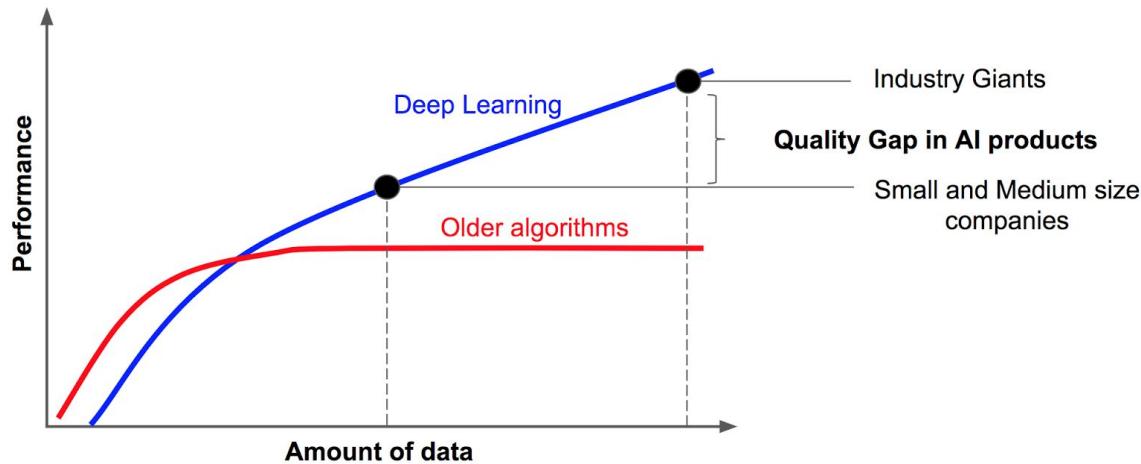
In machine learning, data labeling is the process of identifying raw data (images, text files, videos, etc.) and adding one or more meaningful labels to provide context so that a machine learning model can learn from it

Today, most practical machine learning models use **supervised learning**, which applies an algorithm to map one input to one output

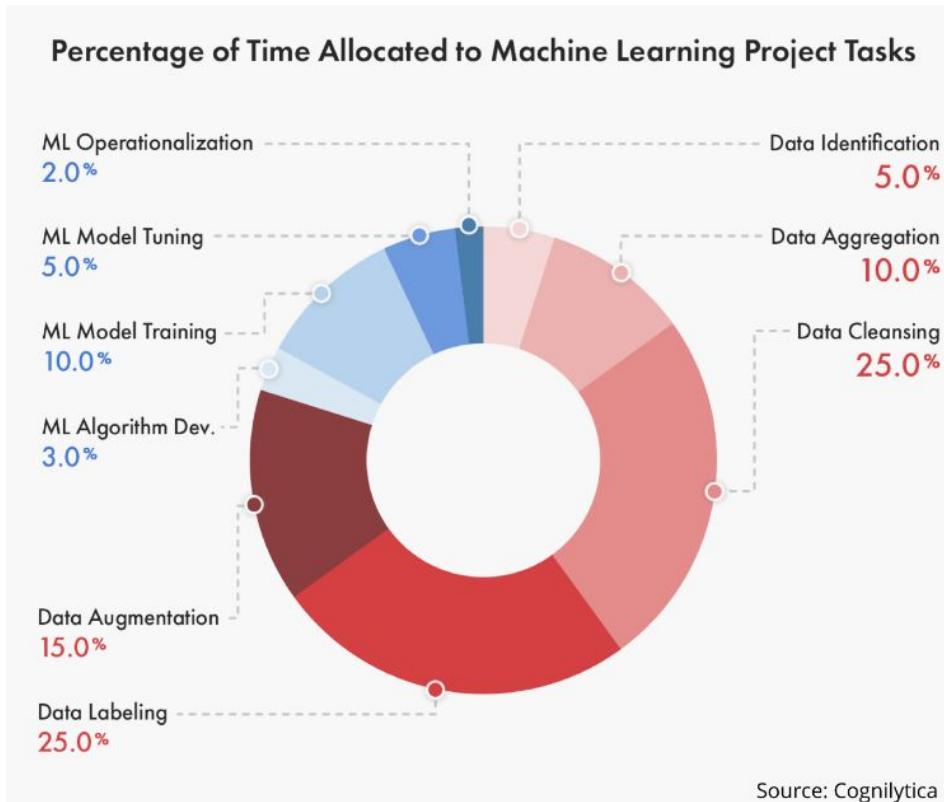
For this you need a labeled set of data that the model can learn from

What is data labeling and why is it important?

To achieve the best performance possible, deep learning needs a lot of labels



Data munging is time consuming



Types of labels

Computer vision

- Image classification
- Object detection
- Semantic segmentation
- Instance segmentation

Natural language processing

- Sentiment analysis
- Parts of speech
- Text in images

Audio

- Source classification
- Audio to text

Examples

Three Type of Classification Tasks

Binary Classification



- Spam
 - Not spam

Multiclass Classification



Multi-label Classification

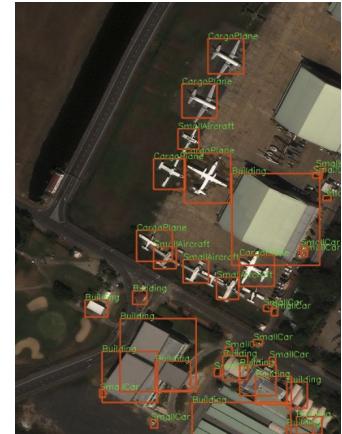


- Dog
 - Cat
 - Horse
 - Fish
 - Bird

Semantic segmentation



Bounding boxes



A black silhouette of five people standing side-by-side against a white background.

Semantic Segmentation

A graphic showing five stylized human figures in red, green, yellow, blue, and purple. Each figure is labeled with its corresponding number: Person 1 (red), Person 2 (green), Person 3 (yellow), Person 4 (blue), and Person 5 (purple). The figures are arranged in a horizontal line.

Instance Segmentation

Sentiment Analysis



My experience so far has been fantastic!

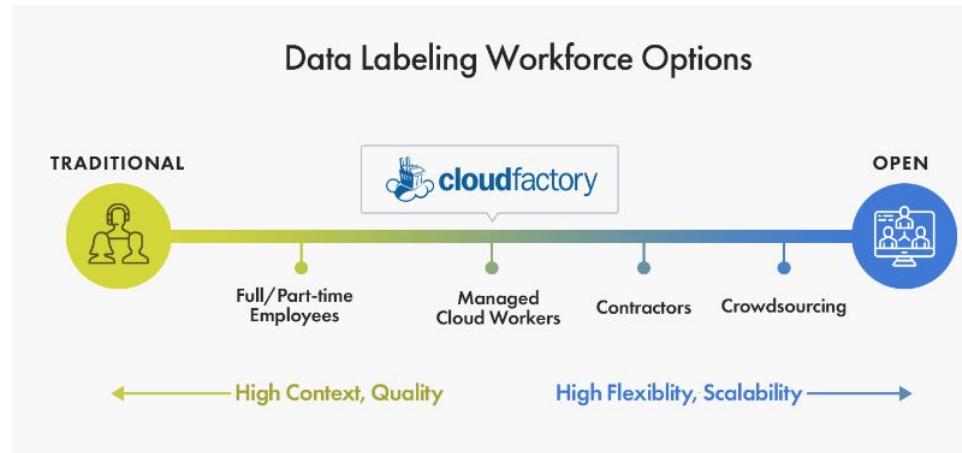
The product is
ok I guess



Your support team is
useless

How do companies label data today?

- Employees - on payroll, full-time or part-time. Their job description may or may not include data labeling.
- Managed teams - vetted, trained, actively managed data labelers.
- Contractors - temporary or freelance workers.
- Crowdsourcing - third-party platform to access large number of workers at once.



Proliferation of labeling companies / tools

Commercial Data Annotation Tool	Annotation Supported								Deployment Model	
	Computer Vision					NLP				
	2D	3D	Video	Audio	Transcription	Text	Transcription	On-premise		
Annotell	✓	✓								
Dataloop AI	✓			✓	✓	✓				
Datasaur AI						✓	✓			
Deepen AI	✓	✓								
Hasty	✓				✓					
Hivemind					✓	✓	✓	✓		
LightTag						✓				
UnderstandAI	✓	✓								
V7 Labs Darwin	✓	Volumetric	✓							

Managed Labeling Solutions / Platforms

Active Learning API support

Labeling Functions (Weak supervision)

5 considerations when labeling data

1. Label quality
2. Scalability
3. Pricing and incentives
4. Tooling
5. Security

Consideration 1: Label Quality

Label quality is determined by:

- **Workforce knowledge and context** - “We’ve found that workers label data with far higher quality when they have context, or know about the setting or relevance of the data they are labeling.”
- **Agility** - “A flexible data labeling team can react to changes in data volume, task complexity, and task duration.”
- **Communication** - incorporating feedback into labeling

Consideration 1: Label Quality

How to measure quality?

1. **Gold standard / benchmark** - There's a correct answer for the task.
Measure quality based on correct and incorrect tasks.
2. **Sample review** - Select a random sample of completed tasks. A more experienced worker reviews the sample for accuracy.
3. **Consensus** - Assign several people do the same task, and the correct answer is the one that comes back from the majority of labelers.

Consideration 2: Scalability

Elements of scalability:

- **Workforce size:** maximum number of workers labeling at once
- **Elasticity:** ability to scale up and down as needed
- **Worker productivity:** volume of completed work, accuracy/consistency, and worker engagement
 - “On the worker side, strong processes lead to greater productivity. Combining technology, workers, and coaching shortens labeling time, increases throughput, and minimizes downtime. We have found data quality is higher when we place data labelers in small teams, train them on your tasks, and show them what quality work looks like.”

Consideration 3: Pricing and Incentives

Pricing is typically either:

- Per hour
- Per annotation

If you pay data labelers per task, it could incentivize them to rush through as many tasks as they can.

Managed workers have more incentive to get things right.

There is often a cost vs. quality trade-off.

Consideration 3: Pricing and Incentives

HiveMind study: per hour vs. per annotation

Task: Transcribing easy text

“Overall, on this task, the crowdsourced workers had an error rate of more than 10x the managed workforce.”

Error Type	Crowdsourced (0.40/iteration)	Crowdsourced (0.80/iteration)	Managed (1.00/minute)
Non-numeric	1	0	0
Year Incorrect	24	9	1
Country Incorrect	6	7	0
Both Incorrect	6	4	0
No Match	6	8	1

Consideration 3: Pricing and Incentives

Task 2: Extracting information from text

“Workers used a title and description of a product recall to classify the recall by hazard type, choosing one of 11 options... The crowdsourced workers’ accuracy was 50% to 60%, regardless of word count. Managed workers achieved higher accuracy, 75% to 85%.”

Consideration 3: Pricing and Incentives

Aim for pricing that is:

- Predictable, so you know what data labeling will cost as you scale
- Pay only for what you need to get high-quality datasets
- Flexible to make changes as your data features and labeling requirements change

Consideration 4: Tooling

You can either:

- Build
 - If you need something very custom
 - Data security
 - Can address bias more easily
- Buy
 - There are funded entities that are vested in the success of that tool
 - Flexibility to use more than one tool, based on your needs
 - Tool provider supports the product

Consideration 5: Data Security

Security can be compromised when workers:

- Access data from an insecure network or using device without malware protection
- Download or save some of your data (e.g. screen captures, flash drive)
- Label your data as they sit in a public place
- Work in a physical or digital environment that is not certified to comply with data regulations you must observe (e.g. HIPAA)

What to look for in a labeling tool

DIY / team

- Intuitive and fast UI
- Ability to generate desired type of labels
- Easy label export
- Built-in quality control
- Ability to add team members

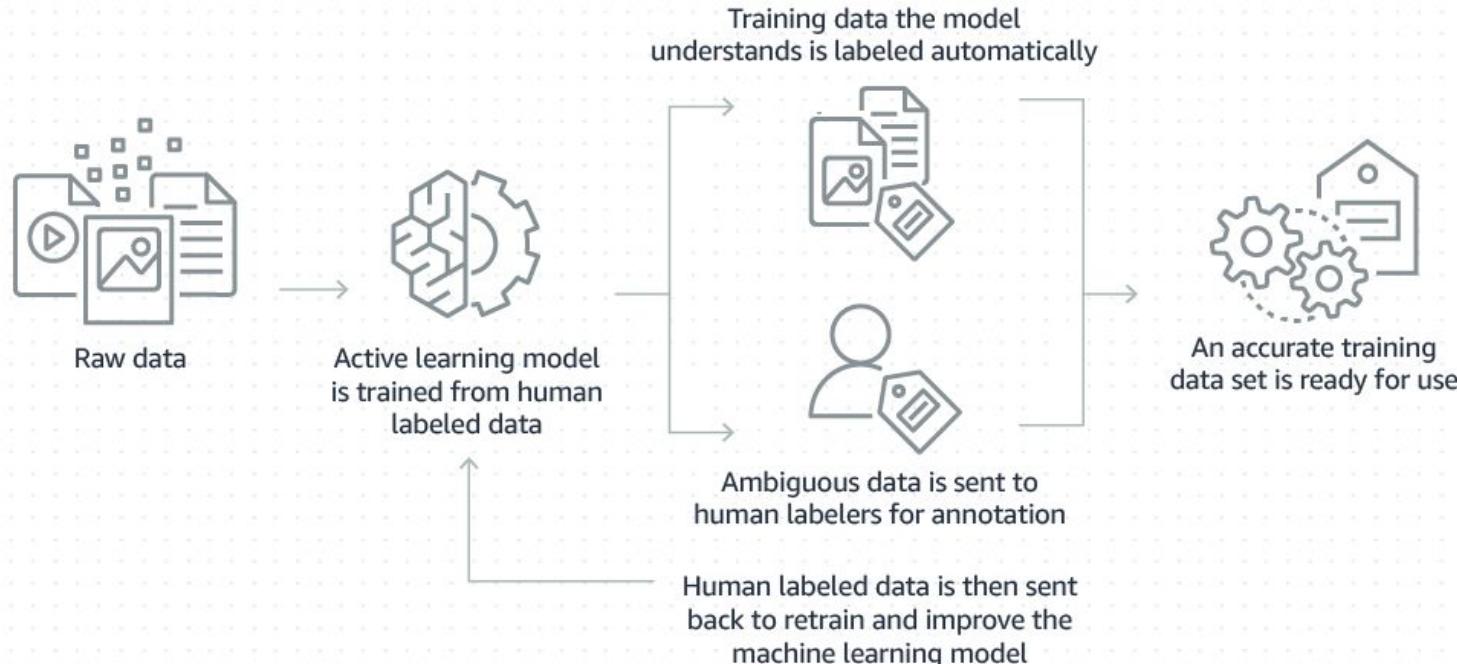
Workforce / crowdsourcing

- Curated, high-quality workers familiar with your type of task
- Easy way to provide feedback on labels
- Label correction / feedback
- Quality metrics

Citizen science

- Build your own UI
- Recruiting channels
- Incentive / gamification

Efficient labeling pipeline



DIY tools

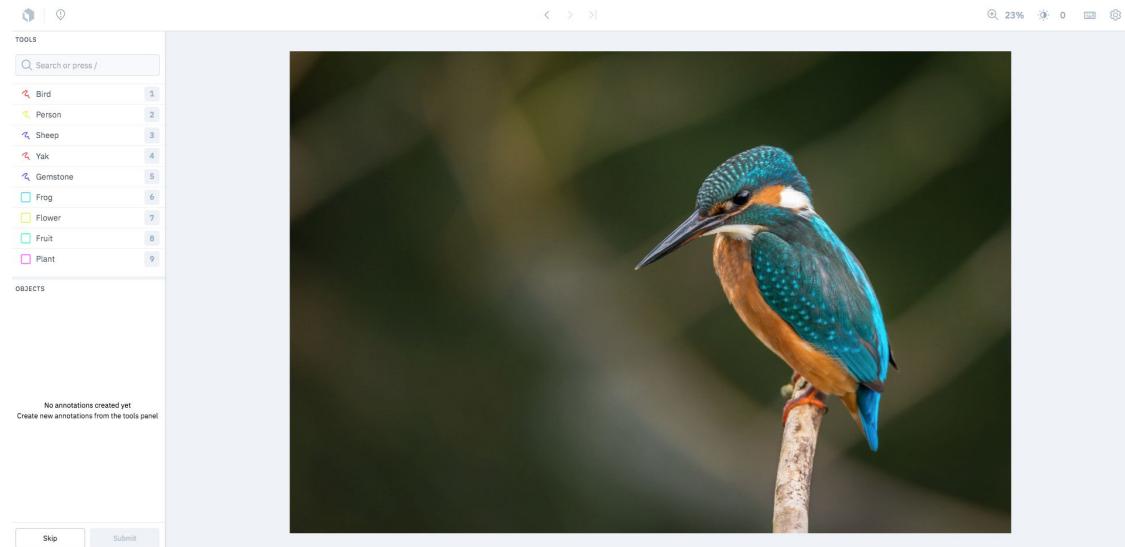
Appropriate when:

- Need a relatively small number of labels (< 100s-1000s)
- Only experts can provide labels
- Data cannot be shared / seen by others
- You are broke
- You have time

Labelbox

Pros: Tons of tooling, fast
(little lag)

Cons: Import your own
imagery, add back geospatial
context later



Labelbox label ontology

(Ontology = formal naming and definition of categories, properties and relations between data)



The image shows an aerial photograph of a residential neighborhood. Several objects have been annotated with colored shapes: a large rectangular area on a roof is highlighted in green, several cars parked along a street are highlighted in pink, and some trees are highlighted in yellow. The Labelbox interface is overlaid on the image.

TOOLS

Search or press /

- Roof 1
- Car 2
- Tree 3
- Street 4
- Skylight 5

Are there pedestrians?

- Yes 6
- No

OBJECTS

- ▶ Roof (5)
- ▶ Car (11)
- ▶ Tree (4)

Skip Submit

Labelbox dashboard

Sample Project

Demonstrating image segmentation and object detection

[Start labeling](#)

[Overview](#) [Labels](#) [Performance](#) [Issues](#) [Export](#) [Settings](#)

Progress

Overall Mine

Submitted	11
Remaining	13
Skipped	0
Complete	45%

Labels created

Weekly ▾

Object count ▾

Object	Count	Share
Flower	13	42%
Bird	5	16%
Person	4	13%
Plant	4	13%
Sheep	3	10%
Gemstone	1	3%
Frog	1	3%
Yak	0	0%
Fruit	0	0%

Training data quality

Review step disabled

Reviews

5

Labelbox Consensus and Benchmark features

The **Consensus** tool allows you to compare the label of one member against the labels of other members to automatically calculate the overall consensus.

The screenshot shows the 'SETTINGS' tab selected in the top navigation bar. On the left, a sidebar lists 'Setup', 'Data Source', 'Labeling Interface', 'Collaborators', 'Quality' (which is highlighted in blue), 'Intelligence', and 'Danger Zone'. The main area is titled 'Quality Control' and contains two sections: 'Benchmark' (with a description and 'Learn more' link) and 'Consensus' (selected, with a description and 'Learn more' link). Below these are sliders for 'How many assets?' (set to 10%) and 'How many times?' (set to 3). A 'Summary' section at the bottom shows a calculation: 'Original Labels' (34) + 'Consensus Labels' (6) = 'Total Labels' (40). A 'CONFIRM' button is at the bottom right.

OVERVIEW LABELS PERFORMANCE EXPORT SETTINGS

Setup Quality Control

Data Source

Labeling Interface

Collaborators

Quality

Intelligence

Danger Zone

Benchmark

Measure quality with the help of a gold standard you define. Recommended for larger projects where an absolute measure of quality is necessary.
[Learn more](#)

Consensus

Measure the level of agreement among Labelers on the data set. Randomly distributes assets to be labeled more than once.
[Learn more](#)

How many assets?

Choose the proportion of the total number of assets

% 10 %

How many times?

Choose the number of times each asset will be labeled

x 3

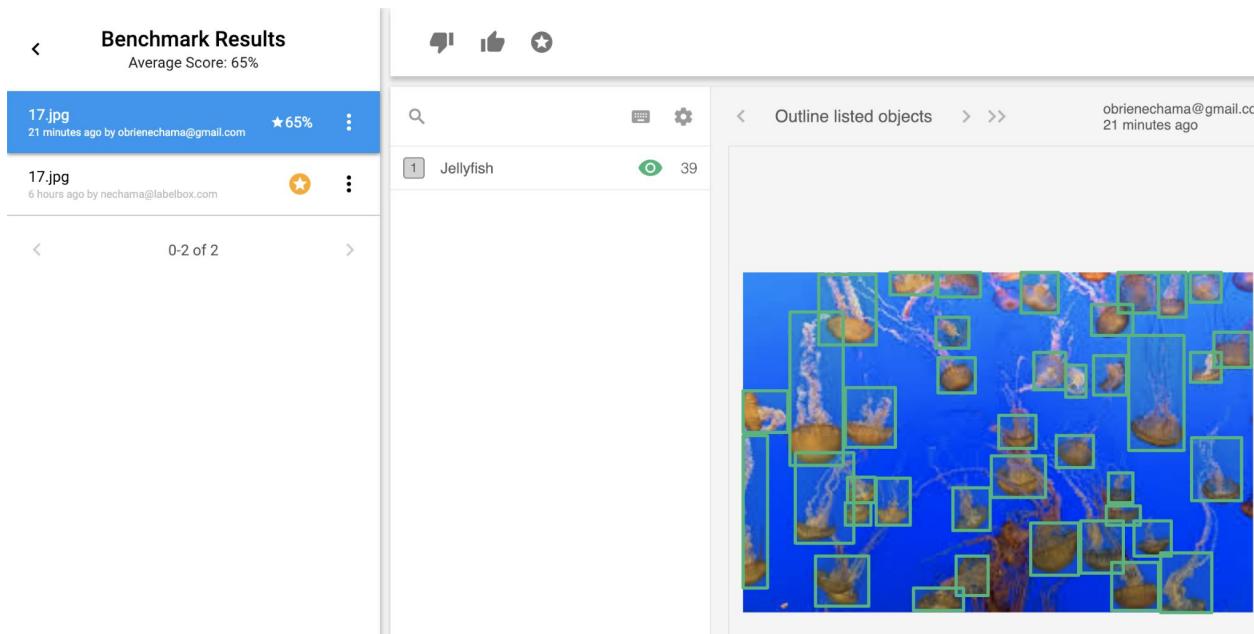
Summary

34
Original Labels + 6
Consensus Labels = 40
Total Labels

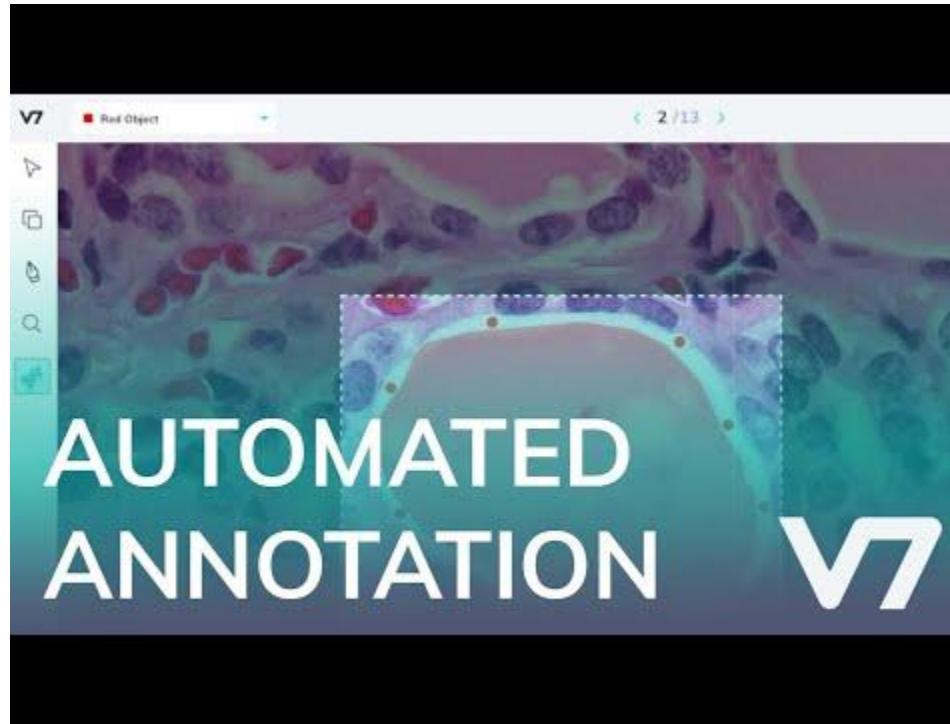
CONFIRM

Labelbox Consensus and Benchmark features

The **Benchmarks** tool allows you to set a “gold standard” annotation and compare your labelers' annotations against it.



Automatic Annotation



DIY Pricing

Company	Founded	Tiers
Labelbox	2018	Free unlimited under educational license Developer: 2500 free labels per year Custom Pro and Enterprise tiers
Azavea (GroundWork)	2000 (2020)	Free: 10 projects, 10GB storage, 5 collaborators/project Pro: \$10k/year, 50 projects, 50GB storage, 50 collaborators/project
SuperAnnotate	2018	Free trial: 14 days Starter: \$62/mo, 10 users, 10,000 images Pro: request demo, unlimited

Workforce tools

Appropriate when:

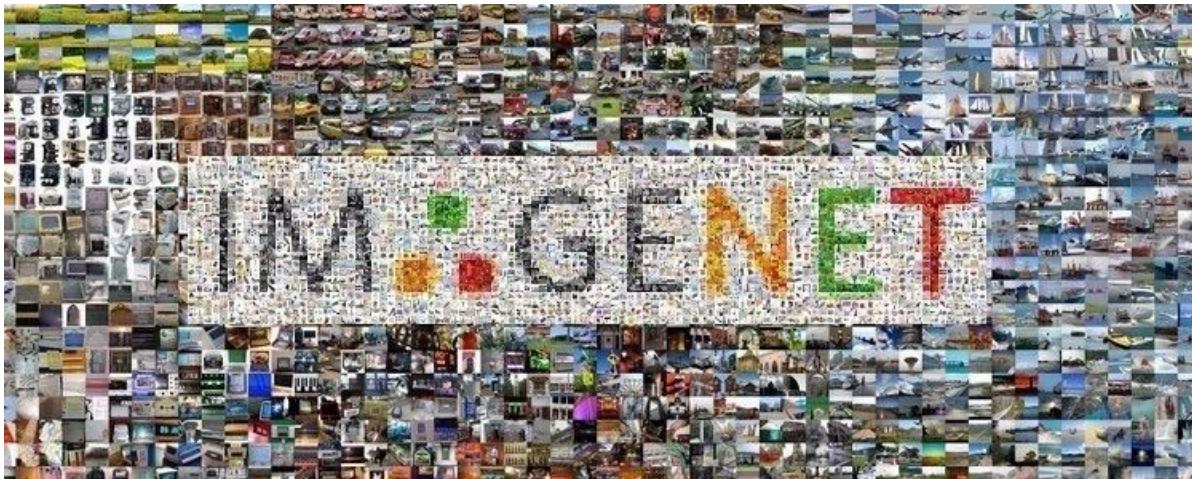
- Need to scale (≥ 1000 s of labels)
- Task can be completed by trained non-experts
- Data is not sensitive
- You have funds

Amazon Mechanical Turk

Pros: high worker capacity, cheap

Cons: Quality issues, have to manage yourself

“By 2018, research had demonstrated that while there were over 100,000 workers available on the platform at any time, only around 2000 were actively working.”



Started 2007, by 2009 had
3.2 million images labeled

Final version 14 million

Amazon Mechanical Turk

Define a “human intelligence task” or HIT

```
question = open(name='questions.xml',mode='r').read()

new_hit = mturk.create_hit()
    Title = 'Is this Tweet happy, angry, excited, scared, annoyed or
upset?',
    Description = 'Read this tweet and type out one word to describe
the emotion of the person posting it: happy, angry, scared, annoyed
or upset',
    Keywords = 'text, quick, labeling',
    Reward = '0.15',
    MaxAssignments = 1,
    LifetimeInSeconds = 172800,
    AssignmentDurationInSeconds = 600,
    AutoApprovalDelayInSeconds = 14400,
    Question = question,
)

print "A new HIT has been created. You can preview it here:"
print "https://workersandbox.mturk.com/mturk/preview?groupId=" + new\_hit\['HIT'\]\['HITGroupId'\]
print "HITID = " + new\_hit\['HIT'\]\['HITId'\] + " \(Use to Get Results\)"

# Remember to modify the URL above when you're publishing
# HITs to the live marketplace.
# Use: https://worker.mturk.com/mturk/preview?groupId=
```

The screenshot shows the 'Create HIT' page. At the top, there are sorting options ('Sort by: HITs Available (most first)', a 'GO!' button, and links to 'Show all details' and 'Hide all details'). Below this, the 'Items per Page' dropdown is set to 10. The main area displays the HIT details: Requester (Taneem Talukdar), HIT Expiration Date (May 8, 2017), Reward (\$0.15), Time Allotted (10 minutes), and HITs Available (1). A link 'View a HIT in this group' is also present.

The screenshot shows the HIT preview page. It includes a timer at the top left (00:00:00 of 10 minutes) and a button 'Want to work on this HIT? Accept HIT'. On the right, it shows 'Total Earned: \$55.21' and 'Total HITS Submitted: 253'. The main content area contains the question 'Is this Tweet happy, angry, excited, scared, annoyed or upset?' and the instruction 'Type in one word to describe the main emotion in the message. If it is unclear, type in "unclear".' Below the question, it says 'Requester: Taneem Talukdar' and 'Qualifications Required: None'. At the bottom, it shows reward information: 'Reward: \$0.15 per HIT', 'HITs Available: 1', and 'Durations: 10 minutes'.

The screenshot shows the HIT preview page with a completed answer. The question 'Is this Tweet happy, angry, excited, scared, annoyed or upset?' is followed by the response 'Tweet: "I am really looking forward to the next Seahawks game!"'. Below the tweet, there is a text input field with placeholder 'Type in your answer here...' and a note 'You must ACCEPT the HIT before you can submit the results.' The overall layout is identical to the previous screenshot but with the answer filled in.

Please write a few sentences describing the last movie you saw in theaters:

sdfkjwer

Amazon MTurk: known quality issues like bots

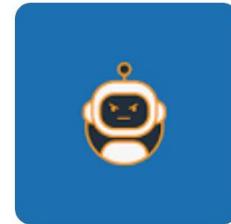


WIRED

[A Bot Panic Hits Amazon Mechanical Turk](#)

A Bot Panic Hits Amazon's Mechanical Turk. Concerned social scientists turned their analytical skills onto one of their most widely used research ...

Aug 17, 2018



Advice for avoiding low quality workers and bots:

1. Use 99+% approval rating. Pay a fair wage and you will get enough workers.
2. Use location USA only. Social security numbers verified by Amazon for federal tax purposes - each worker ID is attached to a single participant.
3. Use HITs approved >1000 which will remove new accounts that could be compromised.
4. Use a master block list. Once blocked, always block.
5. Add a simple captcha or two to your study like "What is 12-8?". If you are using Qualtrics you can incorporate reCAPTCHA.
6. If the simple captcha does not seem secure enough, you can write your question in a jpeg.
7. A specific instruction on writing a sentence below. Not only does this screen out inattentive participants, it also screens out bots because if they do write something, it is usually nonsense ("VERY GOOD STUDY" etc).

Amazon SageMaker Ground Truth

Pros: fully managed, higher quality, large workforce

Cons: possibly not the easiest to set up and work with

Pricing details

You are charged for the number of dataset objects that are labeled. A dataset object is defined as an atomic unit of data and can include images, video frames, text documents, audio files, etc.

Number of labeled objects per month	Price per labeled object
Less than 50,000 objects	\$0.08
50,000 to 1,000,000 objects	\$0.04
Greater than 1,000,000 objects	\$0.02

Amazon SageMaker Ground Truth

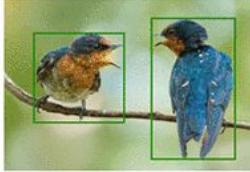
Can create label verifying task

Task type [Info](#)

Task category
Select the type of data being labeled to view available task templates for it or select 'Custom' to create your own.

Image

Task selection
Select the task that a human worker will perform to label objects in your dataset.

- Image classification**
Get workers to categorize images into specific classes. [Info](#)
 - Basketball
 - Soccer
- Bounding box**
Get workers to draw bounding boxes around specified objects in your images. [Info](#)
- Semantic segmentation**
Get workers to draw pixel level labels around specific objects and segments in your images. [Info](#)
- Label verification**
Get workers to verify existing labels in your dataset. [Info](#)
 - Correct label
 - Incorrect label

Label verification tool

Preview 

Provide instructions to help workers identify correct and incorrect labels. Workers will refer to these instructions for each task to verify existing labels. You can add up to 30 labels for workers to choose from. See guidelines for [creating high-quality instructions](#)

Existing labels

Review the existing labels on the objects and choose the appropriate option.

About existing labels
Please draw a tight bounding box around each human that you see in the image below.

Good example
Provide instructions to help workers understand how the task was supposed to be done.

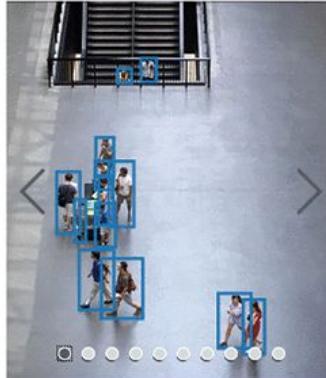
Bad example
Provide examples of mislabeled items that should be rejected.


Select an option
Add up to 30 labels

Label Correct

Label Incorrect

Add label



Workforce Pricing

Company	Per Hour	Per Annotation
Amazon MTurk	--	You set
Amazon Sagemaker Ground Truth	--	\$0.02-0.08 depending on scale
General Blockchain	--	Custom \$0.12/crop field
Labelbox	\$6+/hour	--
Scale	--	\$0.08/image + \$0.08/annotation

Who are the workers?



A.I. Is Learning From Humans. Many Humans.

Artificial intelligence is being taught by thousands of office workers around the world. It is not exactly futuristic work.

Who are the workers?

THE NEW NEW WORLD

How Cheap Labor Drives China's A.I. Ambitions



17



Workers at the headquarters of Ruijin Technology Company in Jiaxian, in central China's Henan Province. They identify objects in images to help artificial intelligence make sense of the world. Yan Cong for The New York Times

Who are the workers?

“Crowd work is generally not well paid. A 2018 study led by Carnegie Mellon University pegged the median wage at around \$2 an hour although workers can push that higher by being careful about the tasks they select.”

WIRED

BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

SIGN IN

Newly Unemployed, and Labeling Photos for Pennies

People who've lost jobs and are stuck indoors are turning to crowd work—filling out online surveys and transcribing audio for less than the minimum wage.



ILLUSTRATION: ELENA LACEY; GETTY IMAGES

Who are the workers?



[Our model](#)

We are proud to be partnering with some of the leading organizations which provide digital skills trainings and livelihoods support in the Middle East



[Syria & Turkey](#)

Roia Foundation

Since 2019 we are partnering with the Roia Foundation to bring online remote job opportunities to conflict-affected people in Turkey and Syria. In Turkey, the Foundation works with asylum-seekers, including people with disabilities and medical doctors. In Syria, annotators come from conflict-affected areas such as Aleppo, Raqqa, and Idlib.



Citizen science

Appropriate when:

- Need to scale (≥ 1000 s of labels)
- Task can be completed by trained non-experts
- Data is not sensitive
- You are broke?* * not sure this is a recipe for citizen science success
- You have time/staff to recruit volunteers and/or gamify task
 - Or some group of people is already trained and super into the task
- Project is long term
- Want methodology to be open
- Studying human efforts are part of the science

The Christmas Bird Count

Longest-running citizen survey in the world (since 1900)

Annual bird census conducted by volunteer birdwatchers



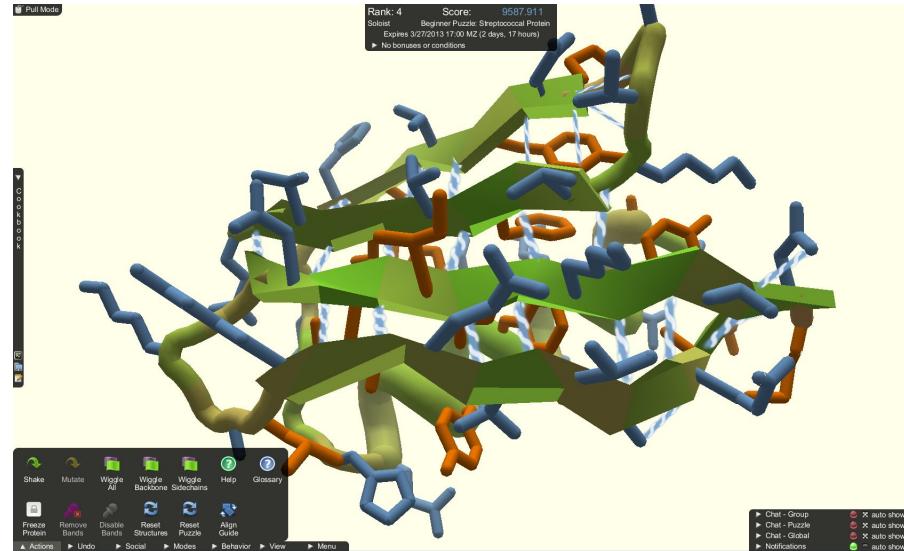
FoldIt

Starting with amino acid sequence, human citizen scientists try to fold proteins as perfectly as possible

Goals:

- Once humans get good at folding known proteins, fold ones with unknown structure
- Learn from human folding algorithms
- Ask humans to design new proteins

Started 2008; by 2010, recruited 200,000+ players

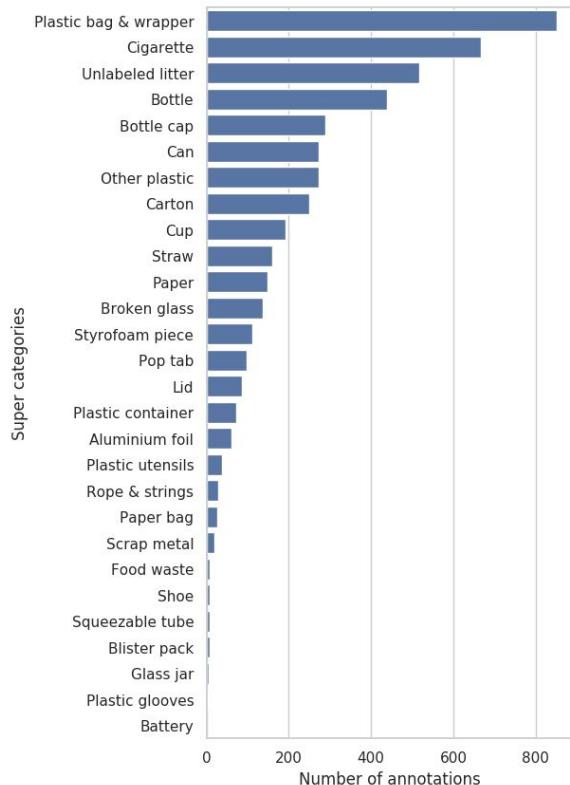


Trash Annotations in Context (TACO) Dataset

Started ~2019?

Users submit photos and annotations

Now: 1500 images with 4784 annotations



NASA Coral Mapping

NeMO-Net: “a video game in which players identify and classify corals using these 3D images while virtually traveling the ocean on their own research vessel, the Nautilus”



Summary

The quantity and quality of labels can often make or break your model's performance more than the model architecture or algorithm itself.

Options for generating labels range from DIY to massive crowdsourcing efforts.

Main considerations:

1. Label quality
2. Scalability
3. Pricing and incentives
4. Tooling
5. Security

Thanks, that's a wrap for today!

Return tomorrow for:

- Dimensionality reduction
- Variational autoencoders
- Representation learning
- Weakly supervised learning
- Semi-supervised learning
- Self-supervised learning