Introduction to AI for Medical Imaging

DEPARTMENT OF Radiology
UNIVERSITY OF WISCONSIN SCHOOL OF MEDICINE AND PUBLIC HEALTH

August 23rd, 2021

Alan B McMillan, PhD
Dept. of Radiology
University of Wisconsin, Madison

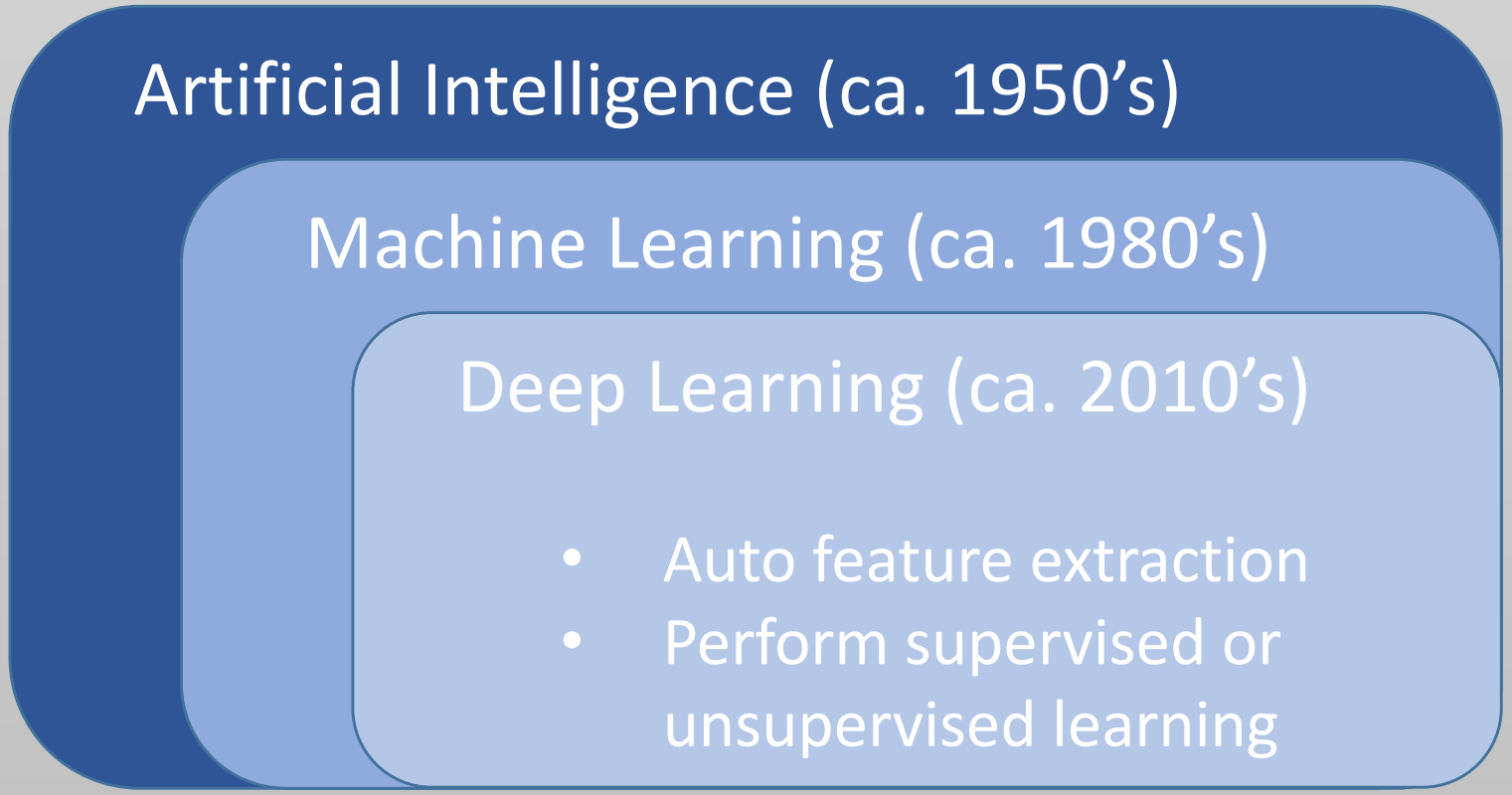
Overview

1. The Paradigm Shift of Solving Problems with Machine and Deep Learning
2. The Steps to Build a Machine Learning Solution
3. How Do We Evaluate a Model

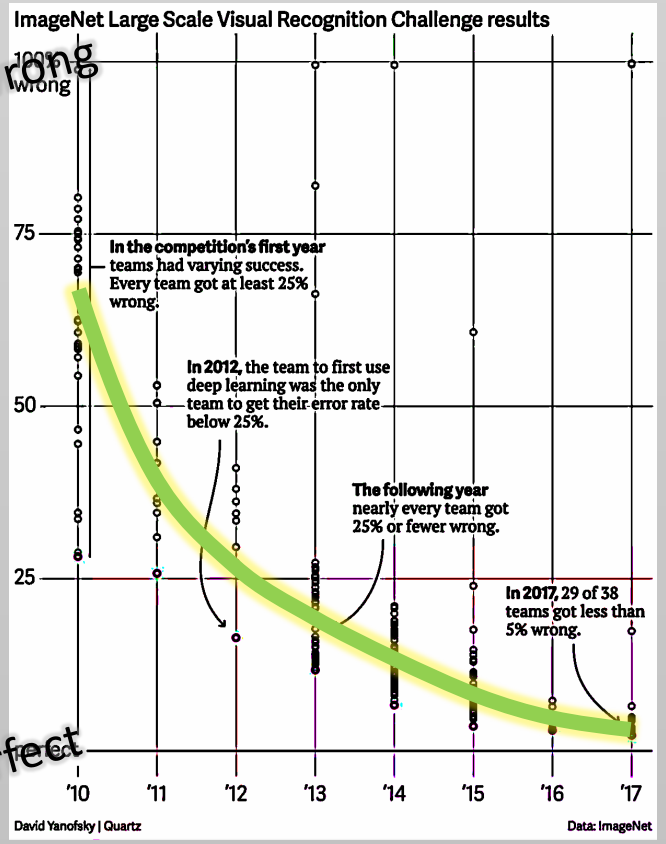
Overview

1. **The Paradigm Shift of Solving Problems with Machine and Deep Learning**
2. The Steps to Build a Machine Learning Solution
3. How Do We Evaluate a Model

AI vs. ML vs. DL


Artificial Intelligence (ca. 1950's)
Machine Learning (ca. 1980's)
Deep Learning (ca. 2010's)
• Auto feature extraction
• Perform supervised or unsupervised learning

ImageNet: Example of Revolution in Capability



from qz.com

Dataset and competition

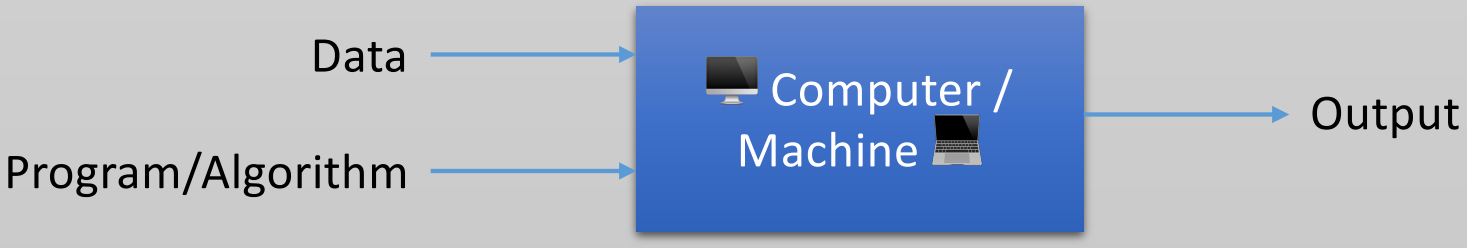
1000 objects categories, localize them in images

DL introduced in 2012

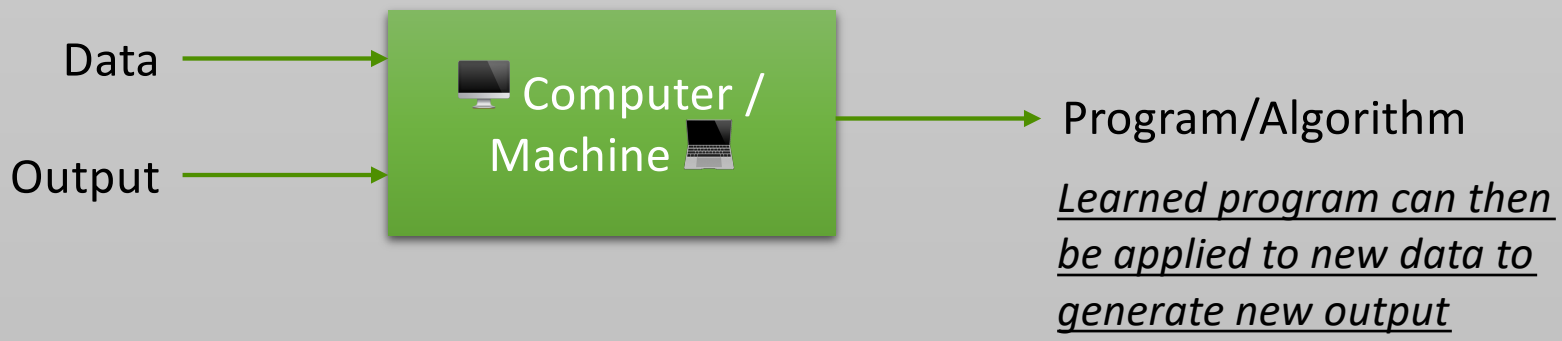
The Machine Learning Paradigm

The ingredients are similar. But the quality, quantity, and curation of the data becomes a primary focus in machine learning papers.

Conventional Paradigm

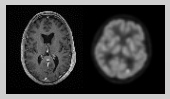


Machine Learning Paradigm (Supervised)



Supervised vs. Unsupervised Learning

Machine Learning Paradigm (Supervised)



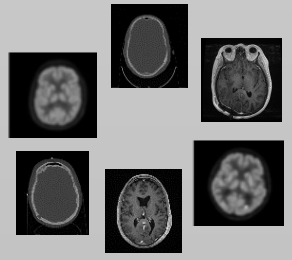
Data

Output

Computer / Machine

Program/Algorithm

Machine Learning Paradigm (Unsupervised)

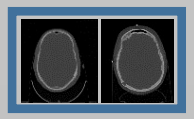


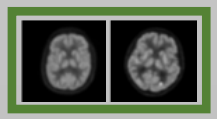
Data

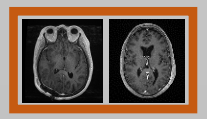
Computer / Machine

Program/Algorithm

Output







Machine Learning vs Deep Learning

Machine Learning

- Boosting / Adaptive Boosting
- Decision tree
- K-means
- K-nearest neighbor (KNN)
- Logistic Regression
- Random forest (RF)
- Support vector machine (SVM)
- Etc…

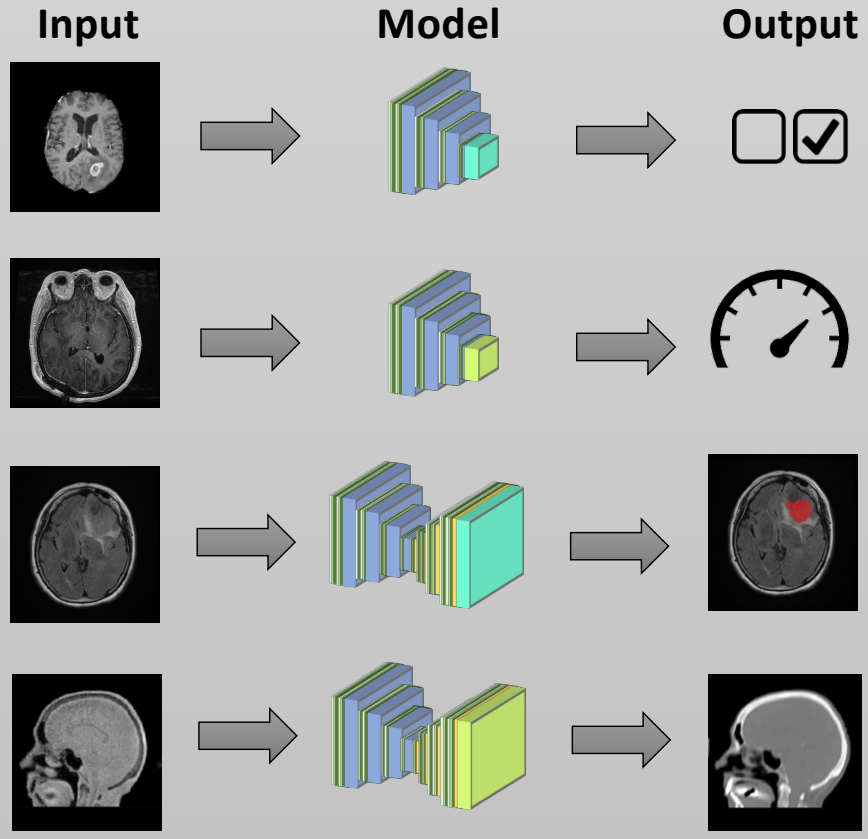
Deep Learning

- Convolution neural network (CNN)
- Deep belief network (DBN)
- Generative adversarial network (GAN)
- Long / Short Term Memory (LTSM)
- Recurrent neural network (RNN)
- Etc…

No time to explain them all! But training/evaluation approach is generally similar.

Data! Data! Data!

Structure is Specific to Application



Classification

- Determine category
- Binary or polynary

Regression

- Convert input to a continuous measure or test score

Segmentation

- Reduce manual labor
- Delineate tumors or structures

Image to Image Synthesis

- Modality Transfer
- Image Reconstruction

Inputs can be images and/or other data: disease status, test scores, demographic data, etc.

Common Classification Architectures

- Many different possibilities
- Top performers:
  - VGG
  - ResNet
  - Inception
  - EfficientNet *more recent

You should first start with existing architectures! Most often, the application is more novel than the network.

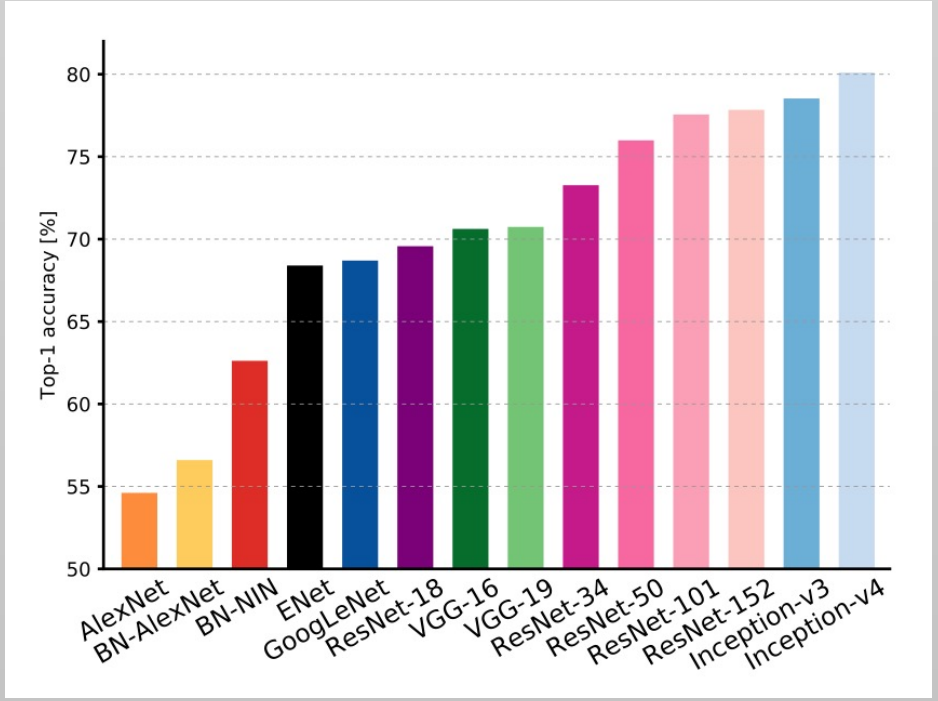
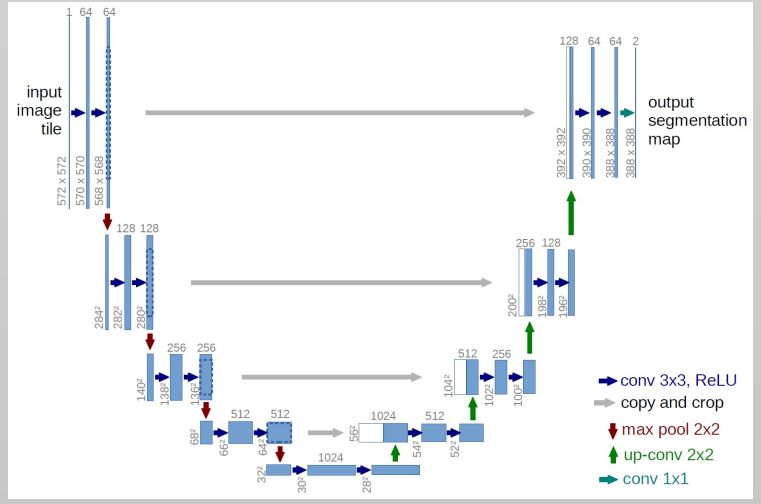


Fig. 1 from Canziani et al. (2017). arXiv:1605.07678

Segmentation Architectures

- Incredible success of **Unet** architecture
  - Contracting (encoder)
  - Expanding (decoder)



https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/

U-net: Convolutional networks for biomedical image segmentation

O Ronneberger, P Fischer, T Brox - International Conference on Medical …, 2015 - Springer

There is large consent that successful training of deep networks requires many thousand annotated training samples. In this paper, we present a network and training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. We show that such a network can be trained end-to-end from very few images and outperforms the prior best method (a sliding …

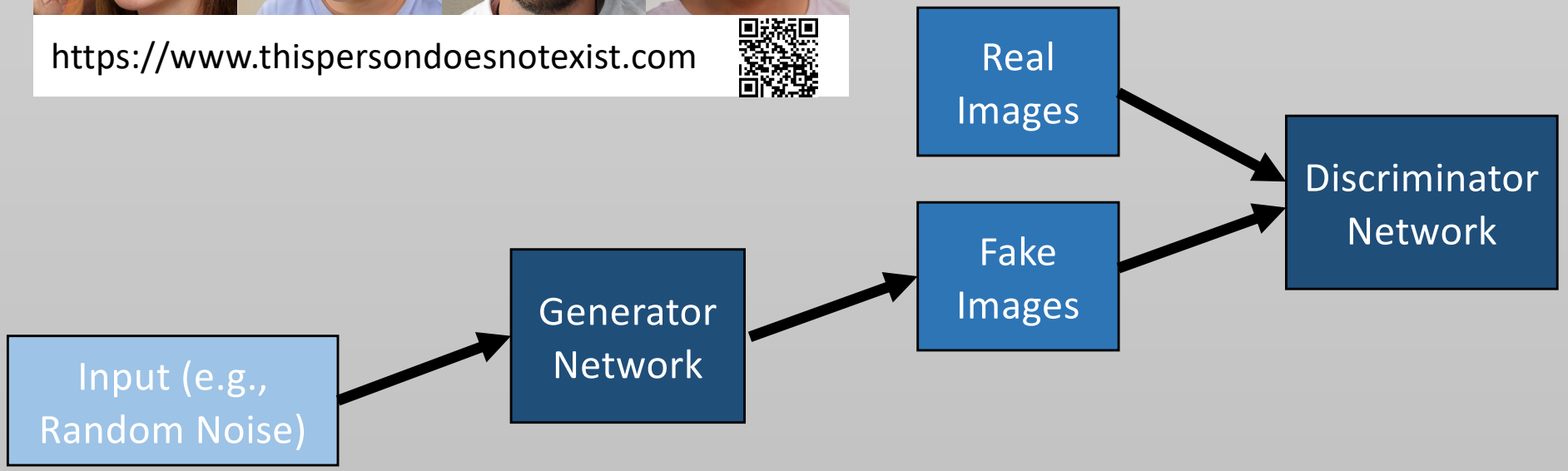
☆ Cited by 25515 Related articles All 21 versions

Nearly ubiquitous in image segmentation and image synthesis applications

GANs – Potential for Image Synthesis

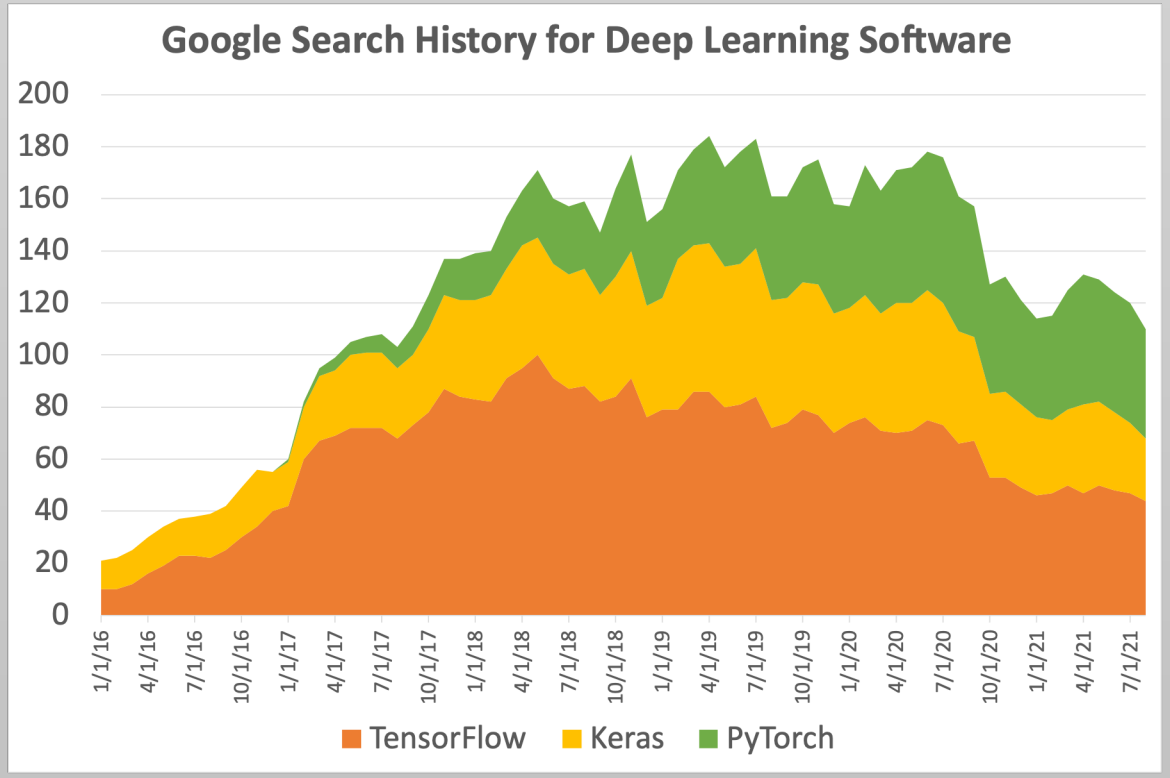
https://www.thispersondoesnotexist.com

Training loss is based on the discriminator, so the output images are often sharper and more realistic.



What software do we use to do DL?

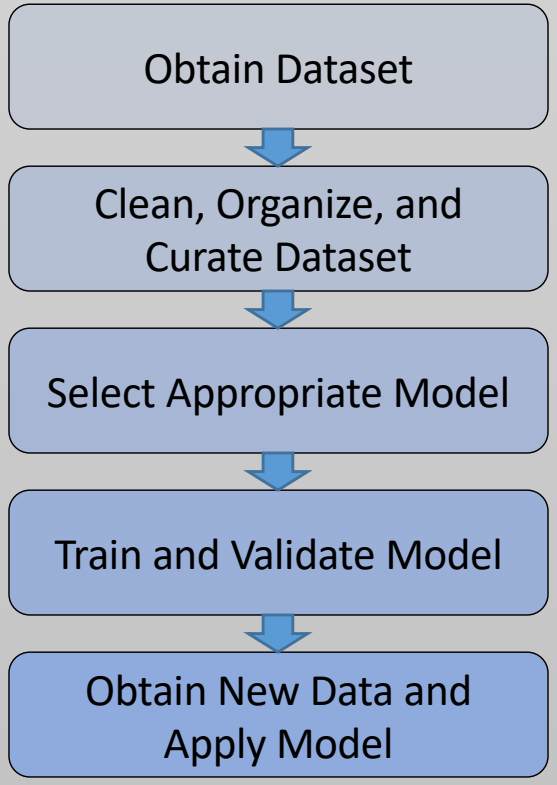
- Two major deep learning software packages:
  - Tensorflow/Keras – we will use this
  - PyTorch



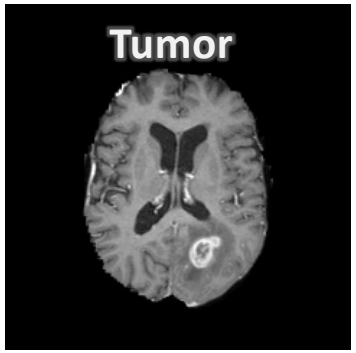
Overview

1. The Paradigm Shift of Solving Problems with Machine and Deep Learning
2. **The Steps to Build a Machine Learning Solution**
3. How Do We Evaluate a Model

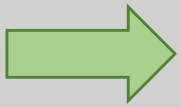
Steps needed to implement ML & DL



Let's make a system to detect tumors on T1 brain images:



Steps needed to implement ML & DL (1)

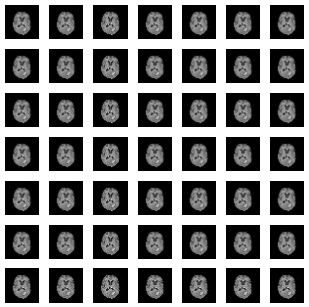


- Obtain Dataset
- Clean, Organize, and Curate Dataset
- Select Appropriate Model
- Train and Validate Model
- Obtain New Data and Apply Model

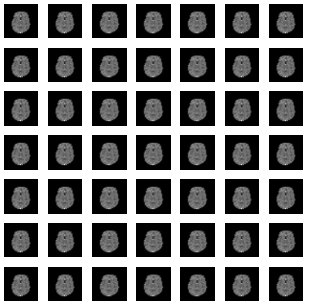
Examples of input and desired output.
Sufficiently large to represent the diversity in your population

For our brain tumor detection system, we need many examples!

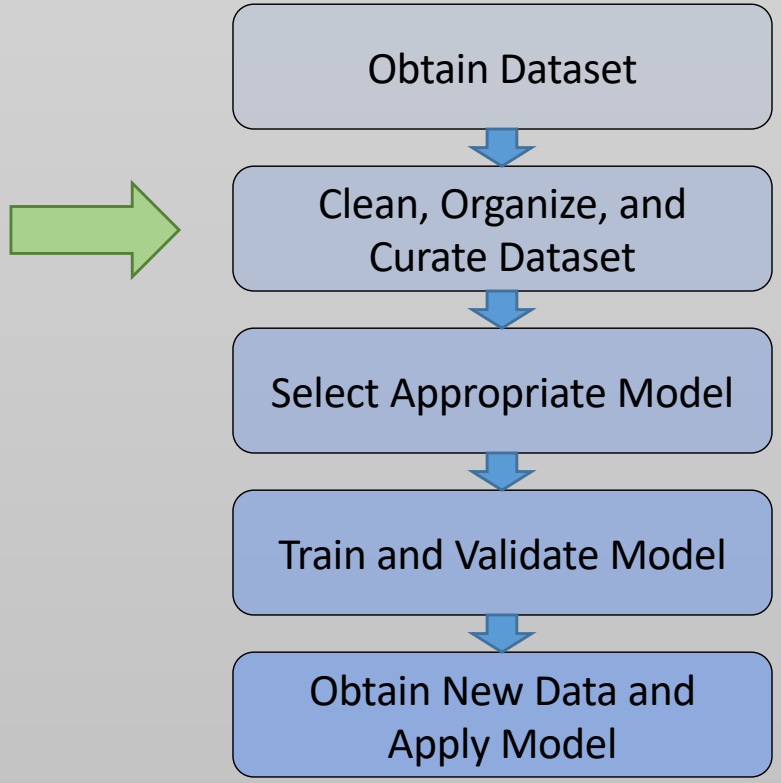
Tumor:



No Tumor:



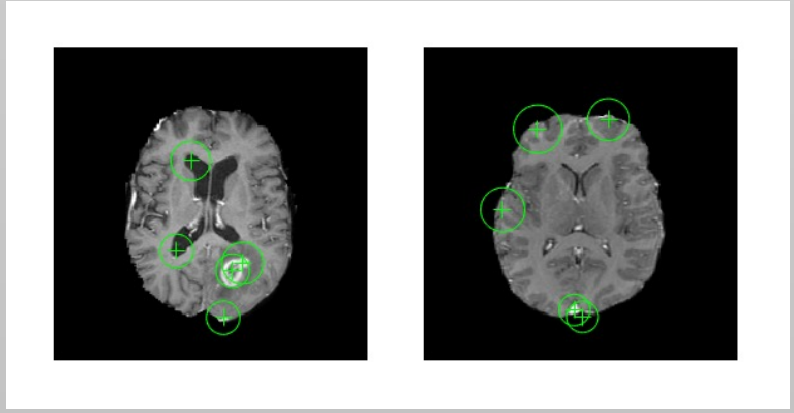
Steps needed to implement ML & DL (2)



Often the most time consuming step. Garbage in = Garbage out
For conventional ML, feature engineering occurs here

A Quick Aside About Feature Engineering

- Deep Learning methods do not require feature engineering
- What is feature engineering?
  - Provide the best representation of data to learn the problem
  - Reduce dimensionality of data
- Example: SURF Features – approx. local maxima



Determining the best features requires:

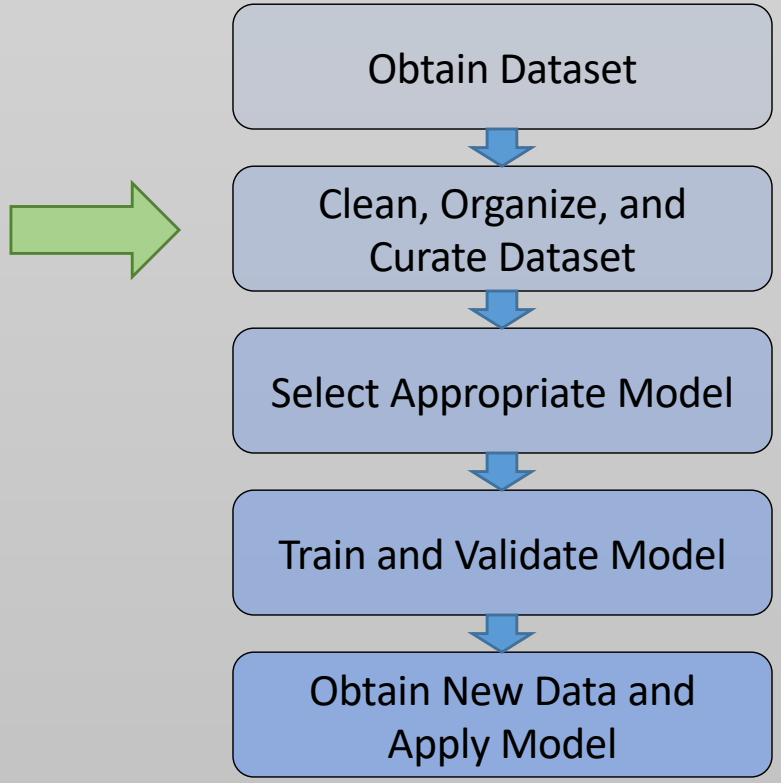
- Domain Knowledge
- Selection + Elimination
- Time / Luck

Recent Attempts to standardize Radiomics:

The Image Biomarker Standardization Initiative:
Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping

Zwanenburg, Vallières et al. (2020). Radiology.

Steps needed to implement ML & DL (2)

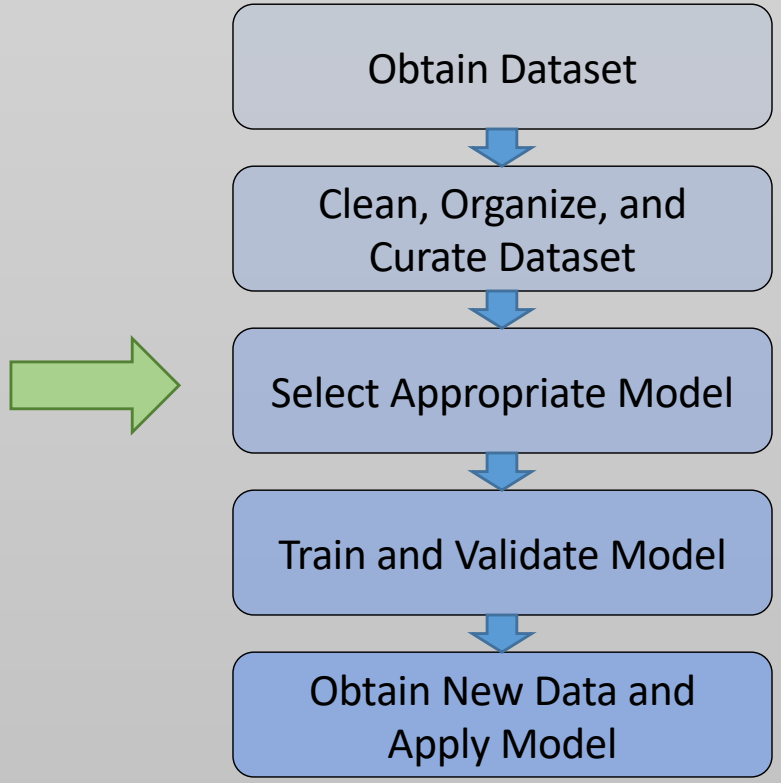


Often the most time-consuming step. Garbage in = Garbage out

We want our brain tumor detection system to learn from the correct answer only (true negatives and true positives)

- **Hematoma**
- **Infection**
- **Radiation necrosis**

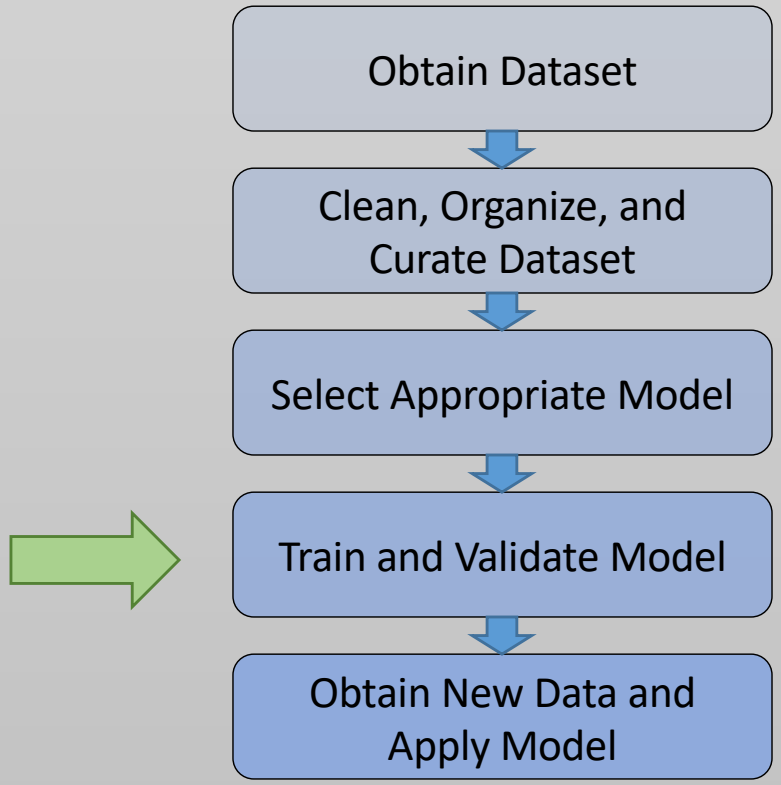
Steps needed to implement ML & DL (3)



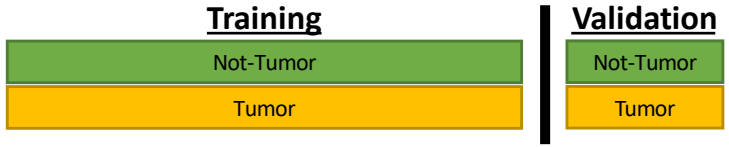
Sometimes an iterative process after training/validation.

Our model should be structured to deliver the answer to the question we are asking it: Tumor or Not?

Steps needed to implement ML & DL (4)



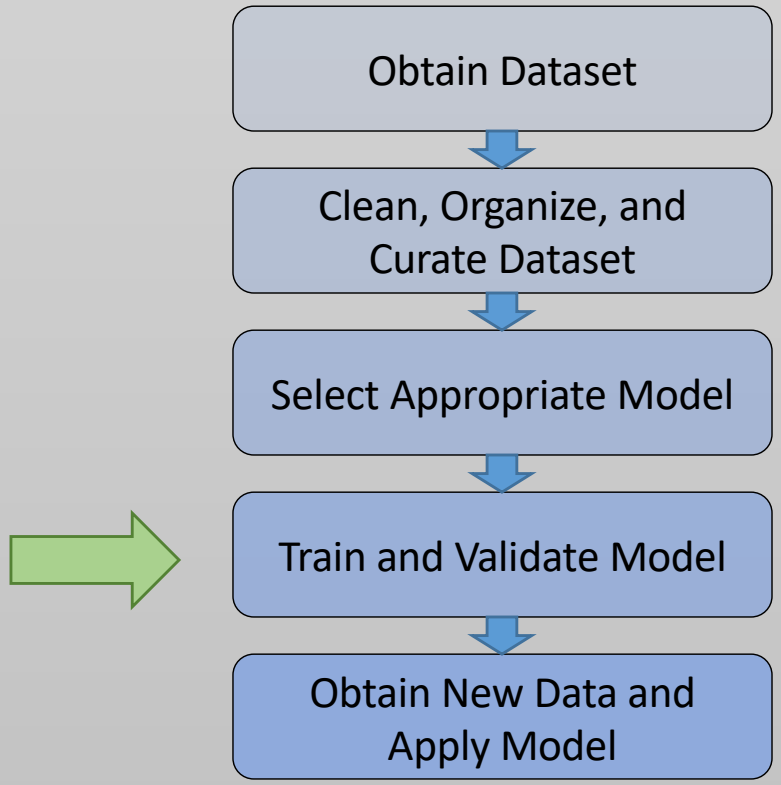
The learning algorithm requires hours (or days) to process. We reserve a portion of our data set to validate the training process (are we learning?)

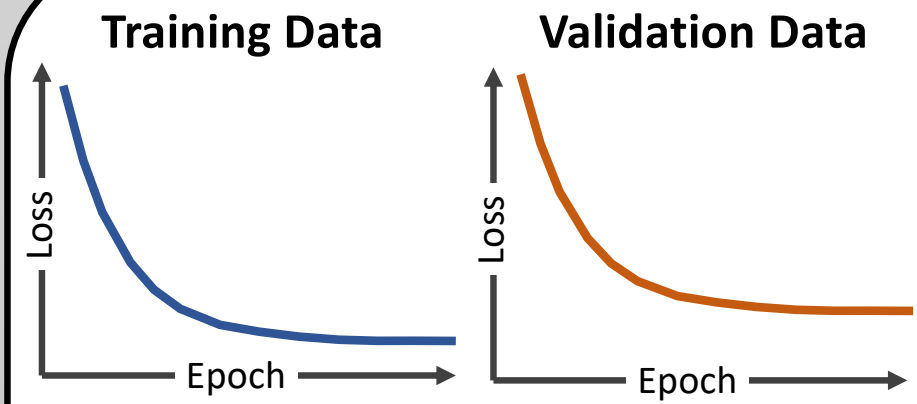


Need big compute to train (GPUs). Naïve validation data important

Often we want to shuffle Training and Validation sets and retrain from scratch to test robustness => Cross Validation

Steps needed to implement ML & DL (4)

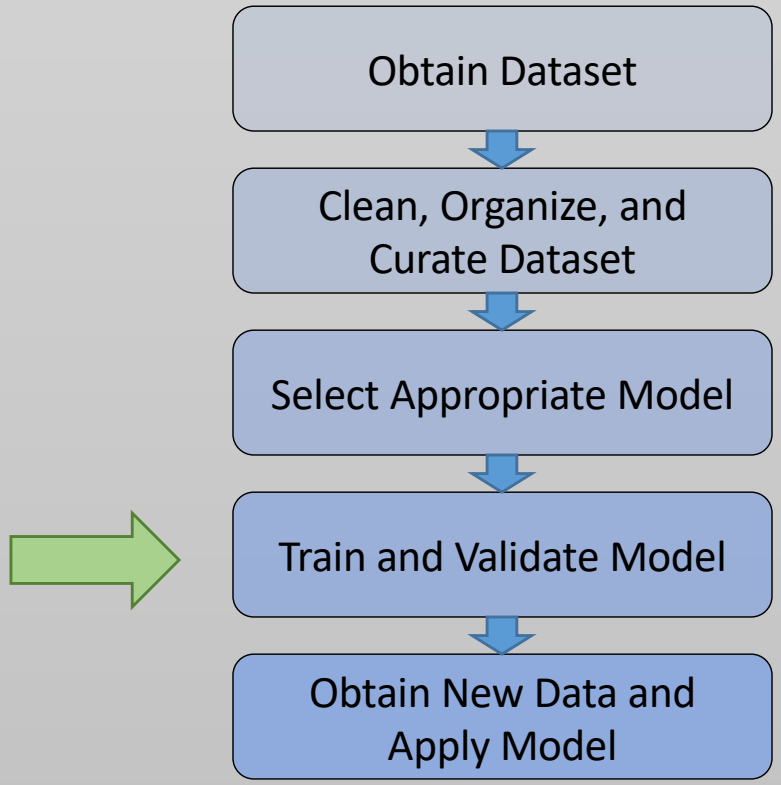


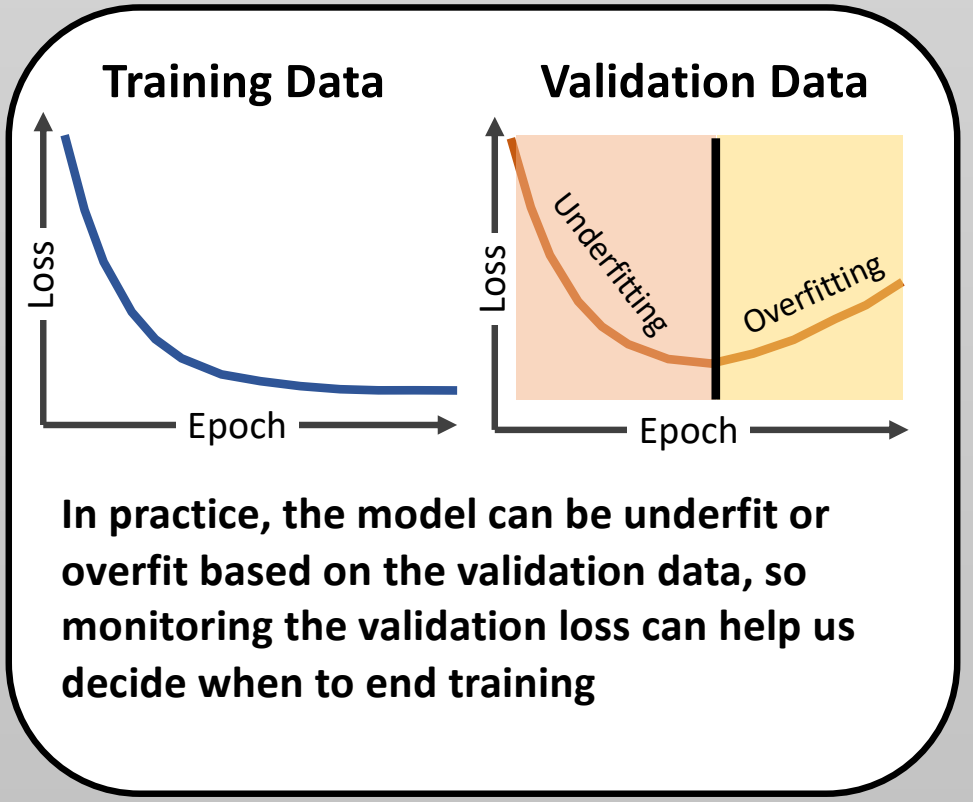


During training, the loss function should decrease at each epoch. Validation data allows us to guess how well the model might work at this point if we stop training

The loss function needs to match the problem at hand.

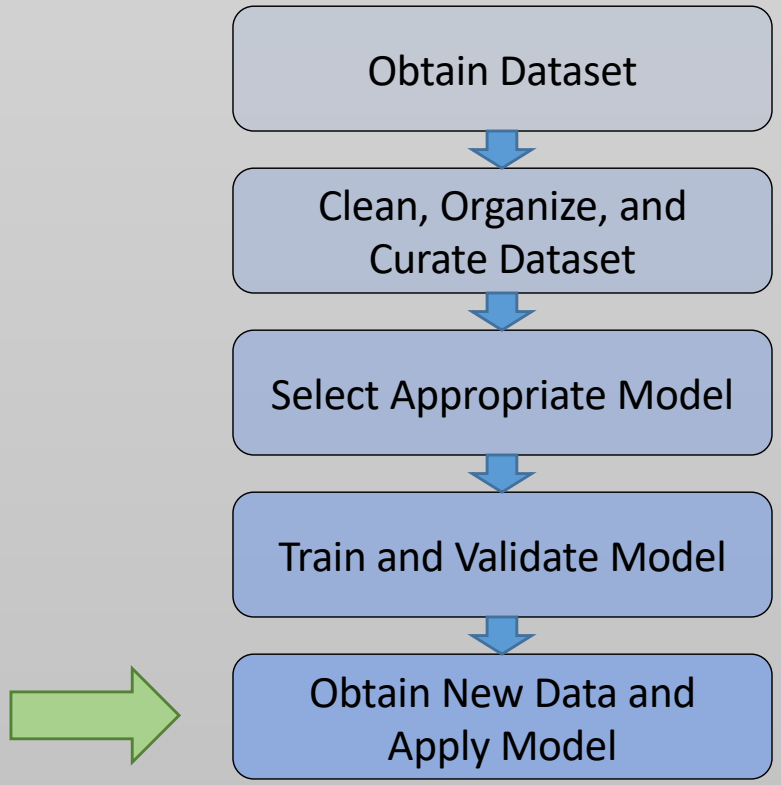
Steps needed to implement ML & DL (4)





In practice, the model can be underfit or overfit based on the validation data, so monitoring the validation loss can help us decide when to end training

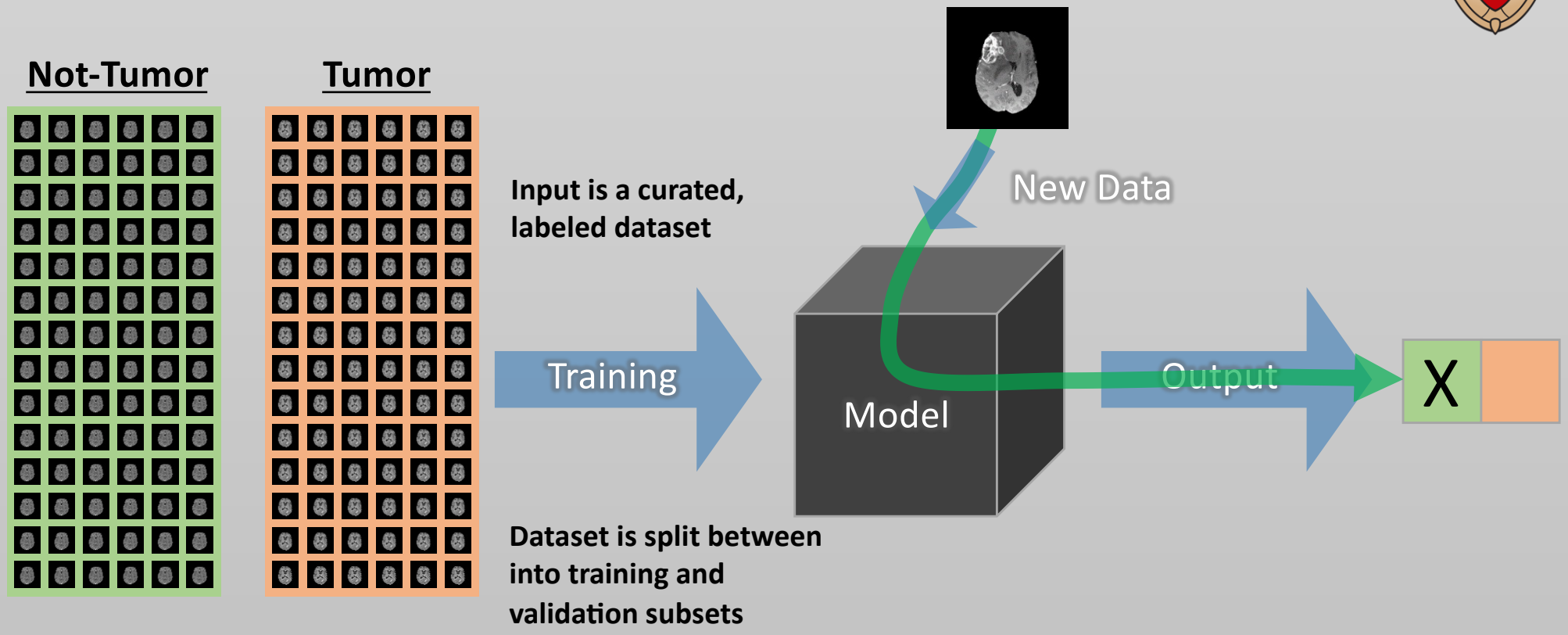
Steps needed to implement ML & DL (5)



If successful, our model can make accurate inferences on new data

New input data can be rapidly assessed.

Example: Brain tumor detection system

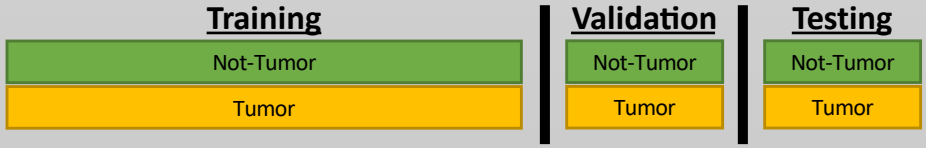


Overview

1. The Paradigm Shift of Solving Problems with Machine and Deep Learning
2. The Steps to Build a Machine Learning Solution
3. **How Do We Evaluate an AI Model**

Evaluating Performance

- Once trained, the model should be evaluated on a testing dataset that also includes ground-truth:



  - Cross-validation is a good strategy for small datasets
- Several performance metrics should be evaluated, depending on problem, for example:

Classification
Accuracy
Sensitivity
Specificity
AUC
TP, FP, FN, TN, etc.

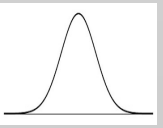
Regression
Mean Absolute Error
Mean Square Error
%-Difference

Segmentation
Dice coefficient
Intersection over union
TP, FP, FN, TN, etc.

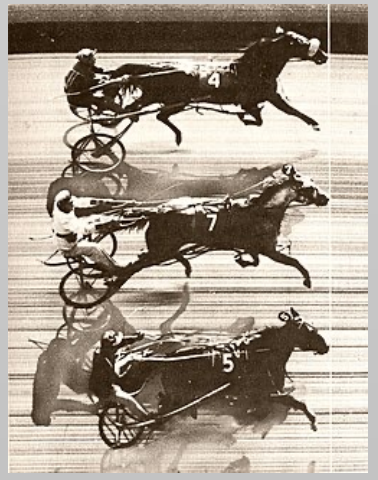
Synthesis
Mean Absolute Error
Mean Square Error
%-Difference
Peak SNR
Structural Similarity
Subjective interpretation

Evaluating Performance

- Model performance should be reported as a confidence interval or mean ± standard deviation.

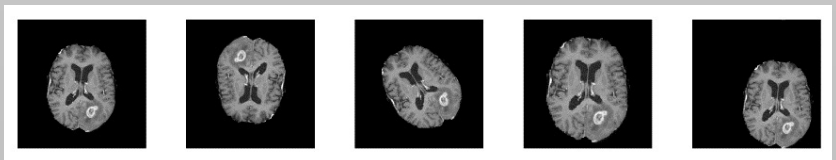


- Statistical tests should be used to demonstrate meaningful differences between developed approaches, e.g.:



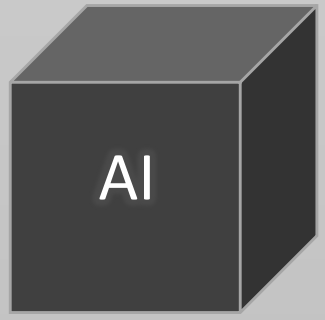
Generalizability and Bias

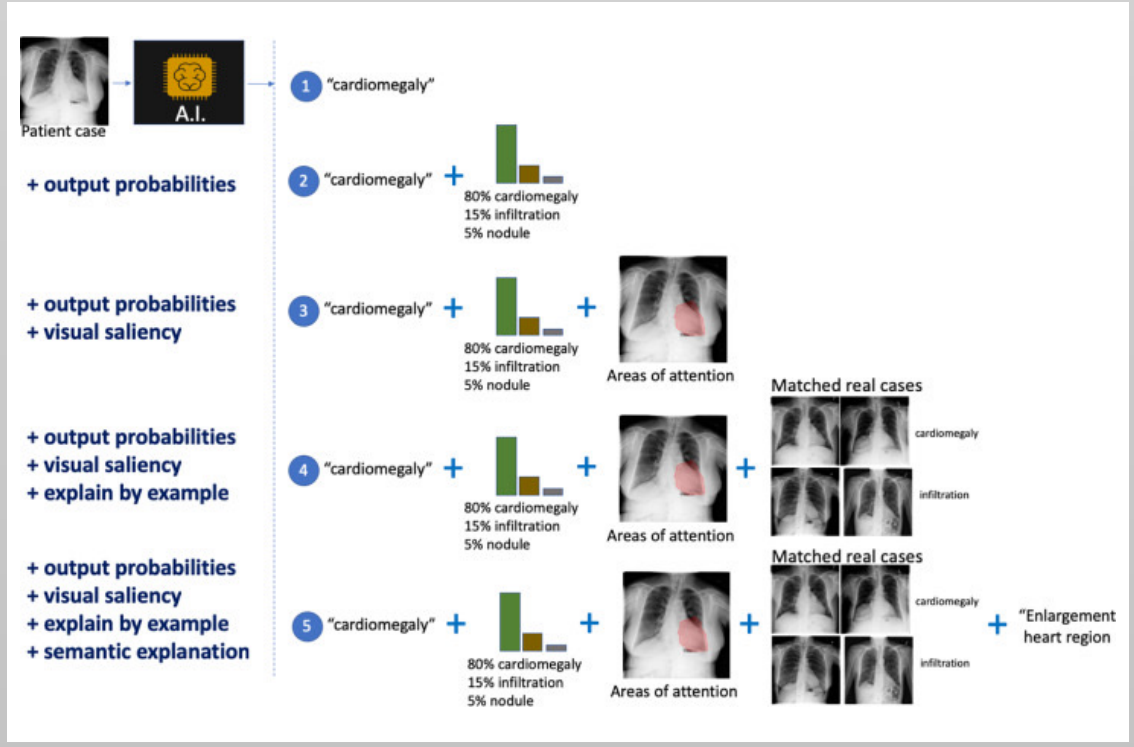
- ML and DL will always perform best on the data it was trained on
- Is it Generalizable? Is the training data biased?
  - Does model work on data outside of the training/validation/testing cohort?
  - Technical dependence? Specific scanner, PSD, vendor, artifact-free, etc.
  - Biased training data? Gender, ethnic, disease, anatomical abnormalities, etc.
- What can we do about these issues?
  - Enlarge dataset
    - Naturally, by getting more data, e.g., multi-institutional federated learning
    - Artificially, by synthesizing more data, e.g., doing data augmentation
  - Be mindful about potential biases



Interpretability

- Deep learning is often considered a black box
  - However, there is a great deal of recent focus on interpretable AI

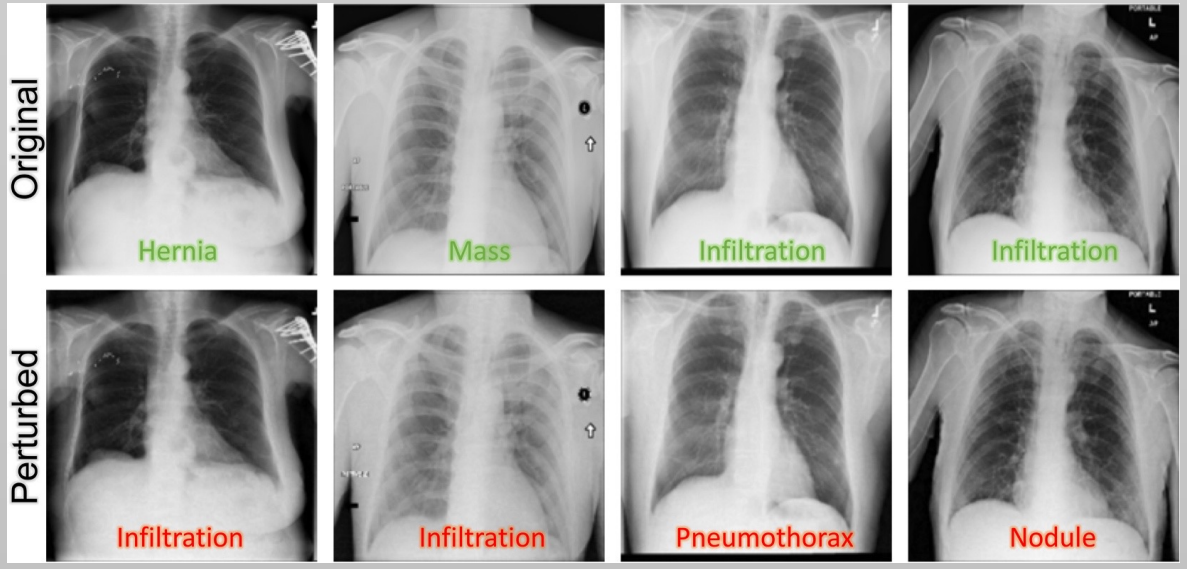




Different approaches for deep learning model interpretation.
Fig. 5 from Reyes et al. (2020). Radiology: AI.

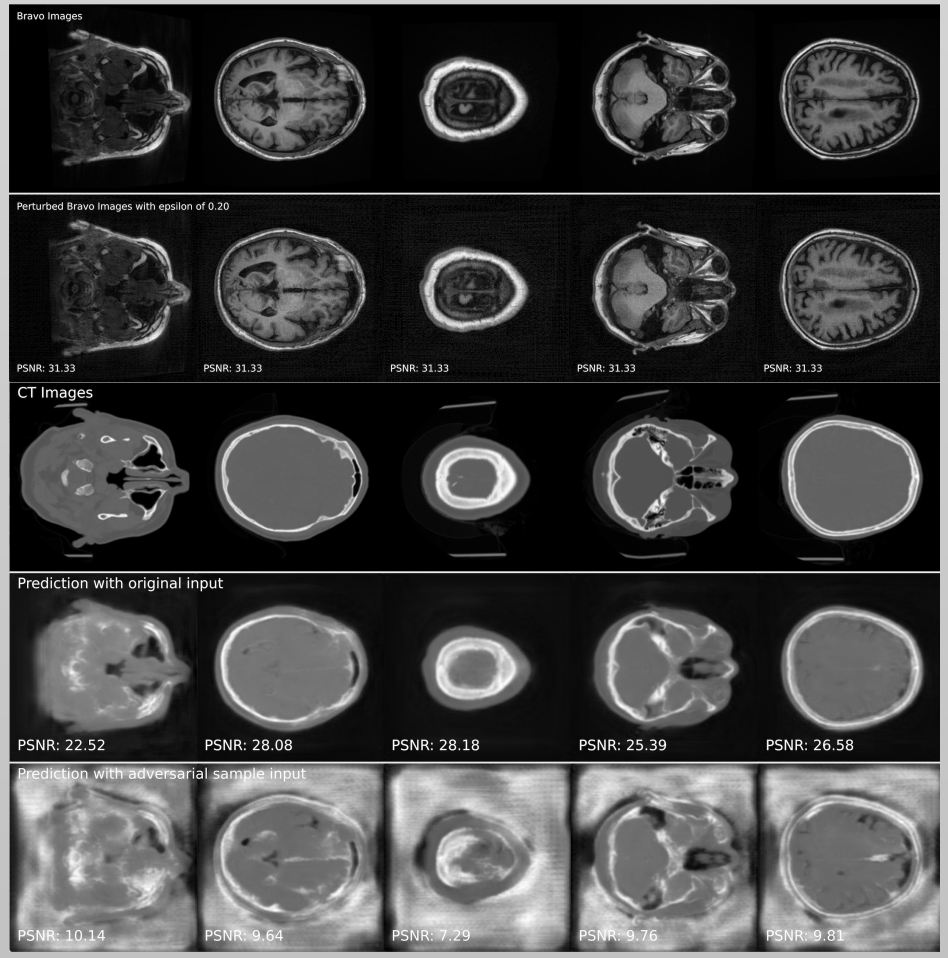
Robustness

- Models can be very sensitive to minor alterations in input
  - Example of correct classifications being made 100% wrong by a minor change to the inputs:



Robustness

- Similar effects for an MR -> CT synthesis approach:



Original MRI

Perturbed MRI

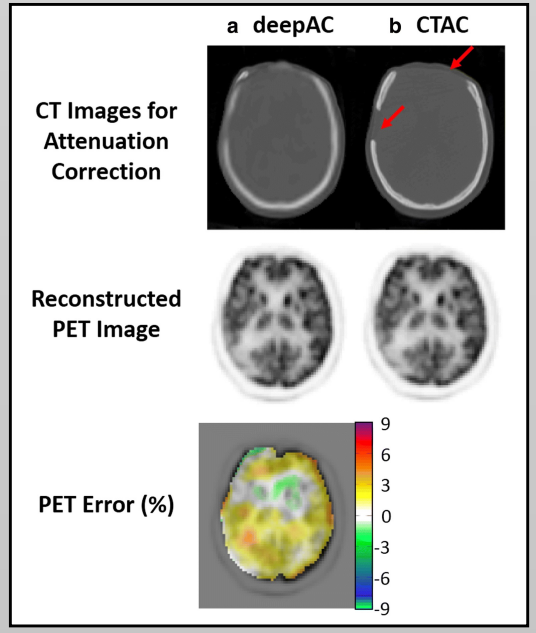
Target CT

Synthesized CT

Synthesized CT
From perturbed MRI

Robustness

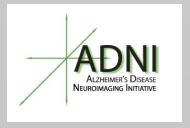
- What can be done to improve robustness?
  - Out of distribution detection
    - Utilize an approach to determine whether the input matches the training data
      - E.g., a classifier network that runs on the input data
  - Adversarial training
    - Train the model with more extreme ranges of input data
      - Can reduce overall performance
  - Embrace unusual results
    - Detecting failure in edge cases can be useful to help understand where a model may fail in the field



Deep learning model exposed to patient with skull abnormality.
Fig. 6 from Liu et al. (2019). EJNMMI Physics. DOI:

Reproducibility through sharing data and code

- Most ML and DL frameworks are open source
- Many authors have shared source code to implementations
- The availability of large databases and example code will help translate technology and foster reproducible research:



Reporting Guidelines for AI Research

- With the rapid proliferation of AI, several reporting guidelines have emerged to ensure high quality research:
- **CLAIM** – minimum reporting for medical imaging AI
- **CONSORT-AI** – minimum reporting for randomized controlled trials
- **DECIDE-AI** – minimum reporting for decision support systems
- **MI-CLAIM** – minimum reporting for reporting medical AI algorithms
- **MINIMAR** – minimum reporting for reporting medical AI algorithms
- **PRIME** – minimum reporting for cardiovascular imaging-related studies
- **SPIRIT-AI** – minimum reporting for interventional trials
- **STARD-AI** – minimum reporting for diagnostic accuracy studies

You should look at these reporting guidelines BEFORE starting your project!

Checklist for Artificial Intelligence in Medical Imaging (CLAIM)

Mongan et al. (2020). Radiology:AI

Checklist for Artificial Intelligence in Medical Imaging (CLAIM)

Section/Topic	No.	Item
TITLE or ABSTRACT		
	1	Identification as a study of AI methodology, specifying the category of technology used (eg, deep learning)
ABSTRACT		
	2	Structured summary of study design, methods, results, and conclusions
INTRODUCTION		
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach
	4	Study objectives and hypotheses
METHODS		
Study Design	5	Prospective or retrospective study
	6	Study goal, such as model creation, exploratory study, feasibility study, noninferiority trial
Data	7	Data sources
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (eg, symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)
	9	Data preprocessing steps
	10	Selection of data subsets, if applicable
	11	Definitions of data elements, with references to common data elements
	12	De-identification methods
	13	How missing data were handled
Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication
	15	Rationale for choosing the reference standard (if alternatives exist)
	16	Source of ground truth annotations; qualifications and preparation of annotators
	17	Annotation tools
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies
Data Partitions	19	Intended sample size and how it was determined
	20	How data were assigned to partitions; specify proportions
	21	Level at which partitions are disjoint (eg, image, study, patient, institution)
Model	22	Detailed description of model, including inputs, outputs, all intermediate layers and connections
	23	Software libraries, frameworks, and packages
	24	Initialization of model parameters (eg, randomization, transfer learning)
Training	25	Details of training approach, including data augmentation, hyperparameters, number of models trained
	26	Method of selecting the final model
	27	Ensembling techniques, if applicable
Evaluation	28	Metrics of model performance
	29	Statistical measures of significance and uncertainty (eg, confidence intervals)
	30	Robustness or sensitivity analysis
	31	Methods for explainability or interpretability (eg, saliency maps) and how they were validated
	32	Validation or testing on external data
RESULTS		
Data	33	Flow of participants or cases, using a diagram to indicate inclusion and exclusion
	34	Demographic and clinical characteristics of cases in each partition
Model performance	35	Performance metrics for optimal model(s) on all data partitions
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)
	37	Failure analysis of incorrectly classified cases
DISCUSSION		
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability
	39	Implications for practice, including the intended use and/or clinical role
OTHER INFORMATION		
	40	Registration number and name of registry
	41	Where the full study protocol can be accessed
	42	Sources of funding and other support; role of funders

Summary

1. The Paradigm Shift of Solving Problems with Machine and Deep Learning
2. The Steps to Build a Machine Learning Solution
3. How Do We Evaluate a Model

Thanks!

@alan_b_mcmillan
amcmillan@uwhealth.org
go.wisc.edu/mimrtl

Torres-Velázquez M, Chen W, Li X, McMillan AB. (2021). Application and Construction of Deep Learning Networks in Medical Imaging. IEEE TRPMS, 5(2) 137-159.

Funding acknowledgement:

NIH NIBIB R01EB026708

NIH NLM R01LM013151