IMAGE DATA ANALYSIS (6CFU)

MODULE OF

REMOTE SENSING

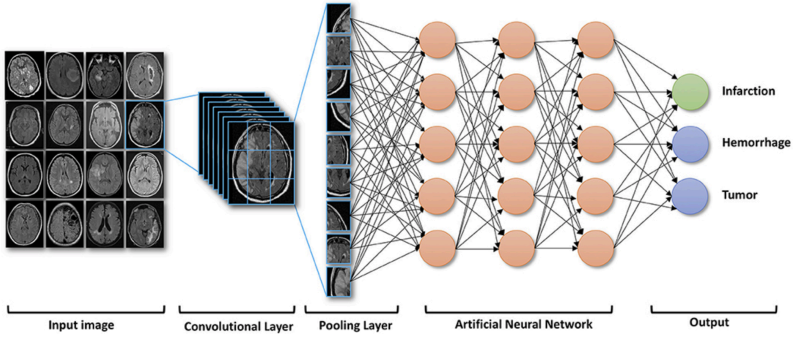
(9 CFU)

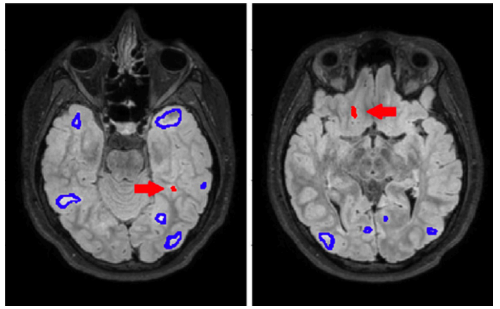
A.Y. 2022/23

MASTER OF SCIENCE IN COMMUNICATION TECHNOLOGIES AND MULTIMEDIA
MASTER OF SCIENCE IN COMPUTER SCIENCE, LM INGEGNERIA INFORMATICA

PROF. ALBERTO SIGNORONI – DR. MATTIA SAVARDI

AI AND DEEP LEARNING FOR MEDICAL IMAGE ANALYSIS





Contents

- Intro - Main issues/challenges
- Data driven approach
- Evaluation metrics
- Applications and Case studies
  - Medical Image Classification
  - Medical Image Segmentation and Detection
  - Higher-dimensionality data
- Interpretability, Fairness, Ethics

2. Design and Deployment of Data-driven solutions for Medical Image Analysis

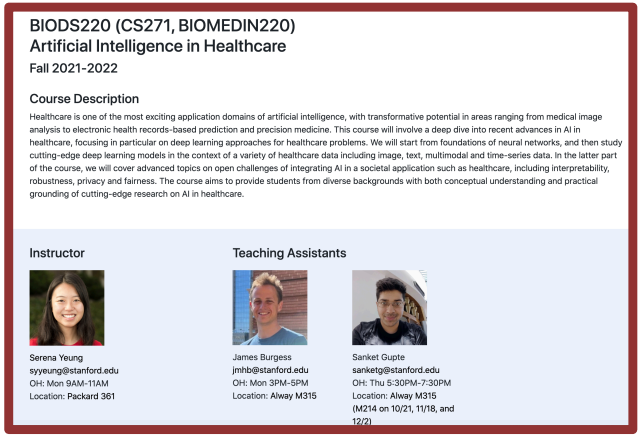
Alberto Signoroni[1,2], **Mattia Savardi**[1]

[1] Department of Information Engineering, University of Brescia - Italy
[2] Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia - Italy

Notes: Chapter 2 of CNIT TECHNICAL REPORT
Vol.8 ICT for Health, Isbn 9788894982541

Fair use disclaimer and Credits

- These slides are for personal use within the course of Image Data Analysis 2021-22 University of Brescia
- Many slides in this lecture come from:
  - AI in Healthcare, Stanford (by Serena Yeung) – https://biods220.stanford.edu


BIODS220 (CS271, BIOMEDIN220)
Artificial Intelligence in Healthcare
Fall 2021-2022

Course Description

Healthcare is one of the most exciting application domains of artificial intelligence, with transformative potential in areas ranging from medical image analysis to electronic health records-based prediction and precision medicine. This course will involve a deep dive into recent advances in AI in healthcare, focusing in particular on deep learning approaches for healthcare problems. We will start from foundations of neural networks, and then study cutting-edge deep learning models in the context of a variety of healthcare data including image, text, multimodal and time-series data. In the latter part of the course, we will cover advanced topics on open challenges of integrating AI in a societal application such as healthcare, including interpretability, robustness, privacy and fairness. The course aims to provide students from diverse backgrounds with both conceptual understanding and practical grounding of cutting-edge research on AI in healthcare.

Instructor

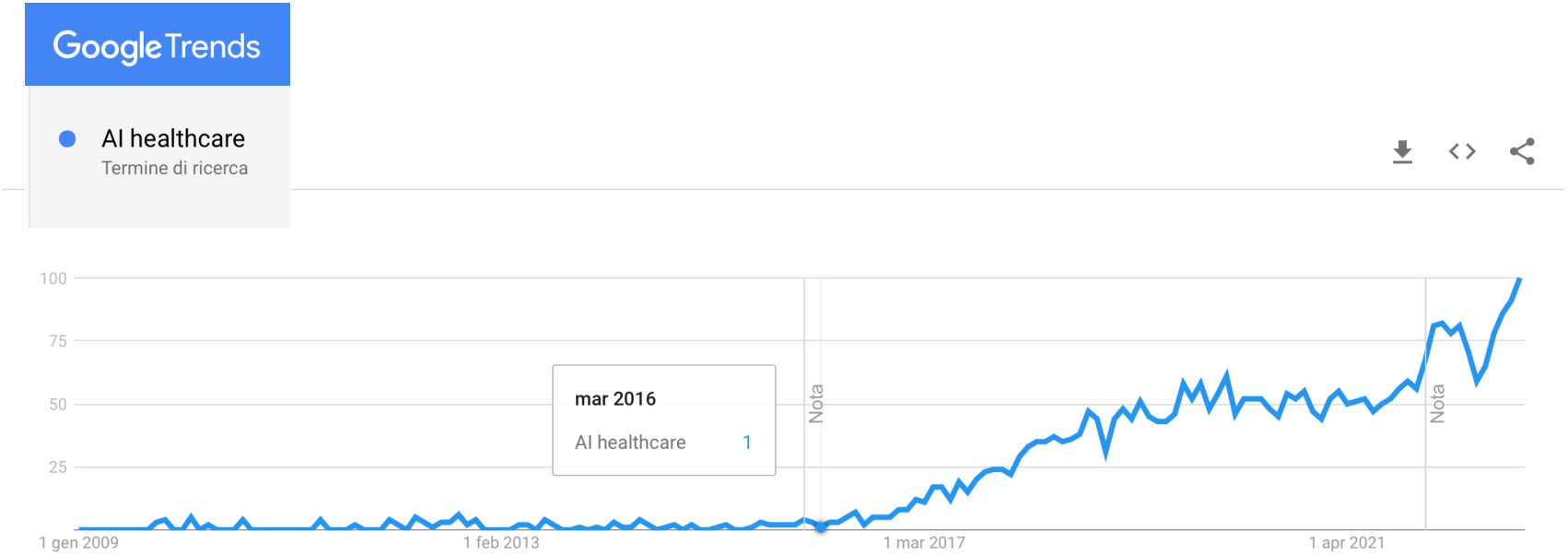
Serena Yeung
syeung@stanford.edu
OH: Mon 9AM-11AM
Location: Packard 361

Teaching Assistants

James Burgess
jmhb@stanford.edu
OH: Mon 3PM-5PM
Location: Alway M315

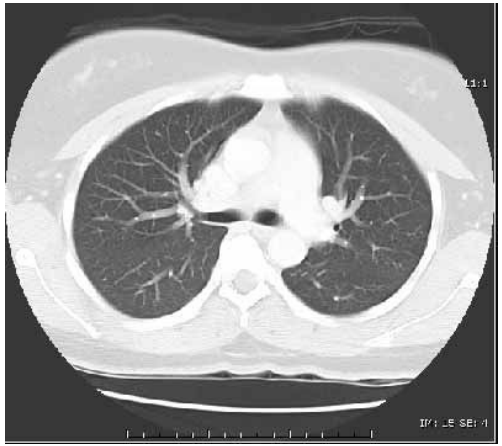
Sanket Gupte
sanketg@stanford.edu
OH: Thu 5:30PM-7:30PM
Location: Alway M315
(M214 on 10/21, 11/18, and 12/2)

AI in healthcare: a rapidly exploding field

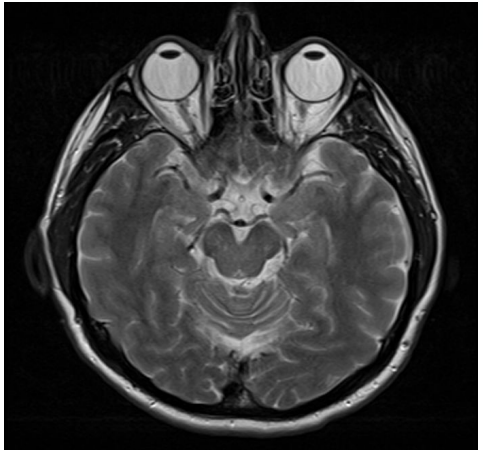


Deep learning for healthcare: the rise of medical data

X-rays (invented 1895).



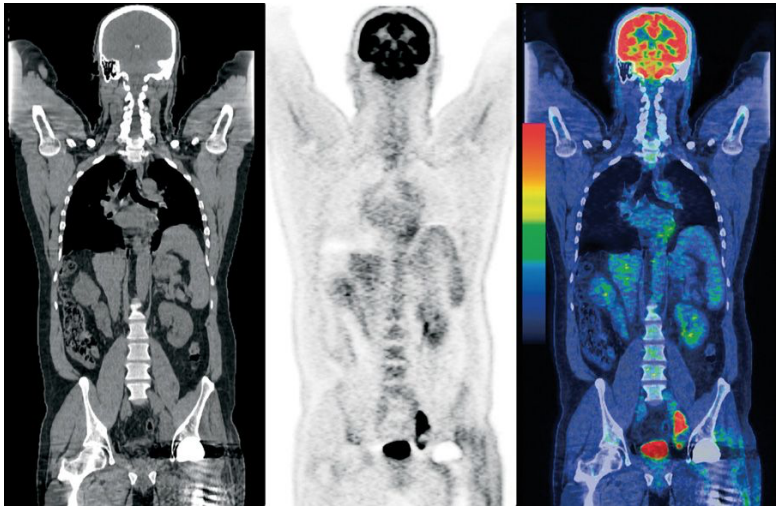
CT (invented 1972).



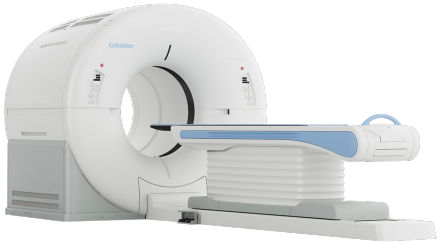
MRI (invented 1977).

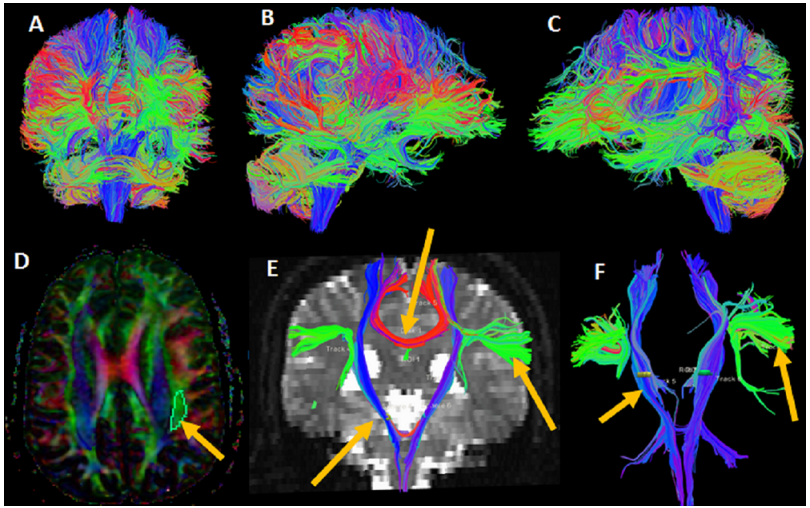
What are other examples of medical image data?

Deep learning for healthcare: the rise of medical data



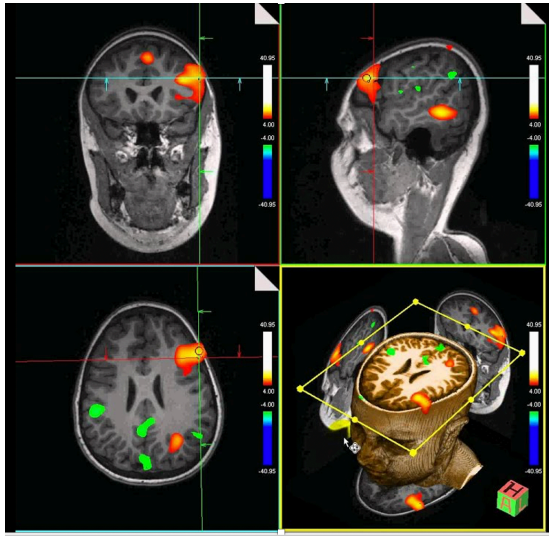
PET/CT (conceived 1991, commercialized 2001: today there is no PET without CT)



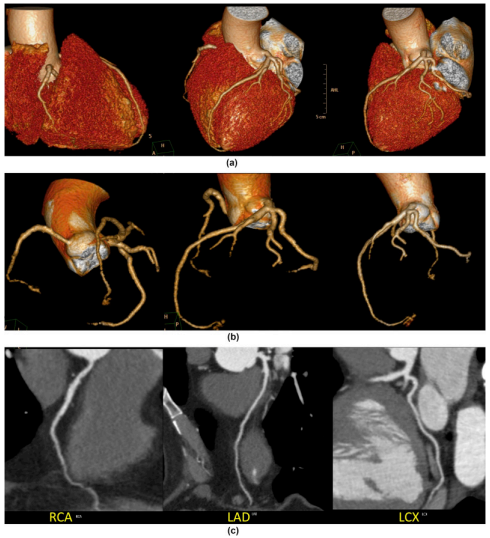


Diffusion Tensor Imaging DTI and Fiber Tractography in MRI (conceived 1994, see https://doi.org/10.1016/j.nec.2010.12.004)

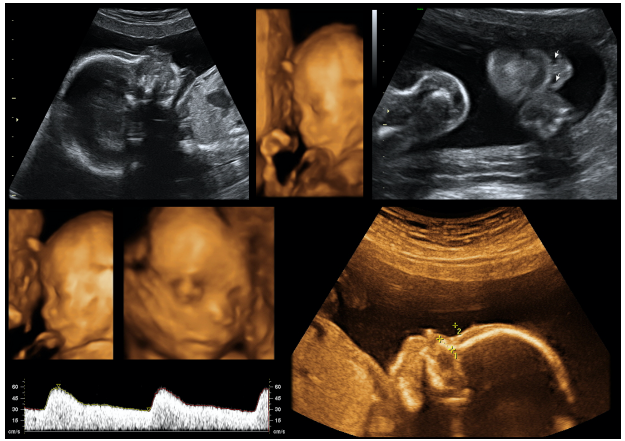
Deep learning for healthcare: the rise of medical data



Functional MRI, fMRI



ECG-Gated CT Angiography

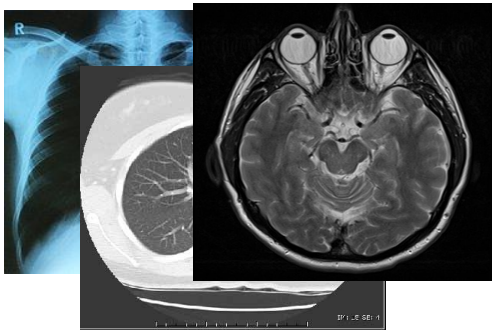


Ultrasound Imaging

… and MANY others!!

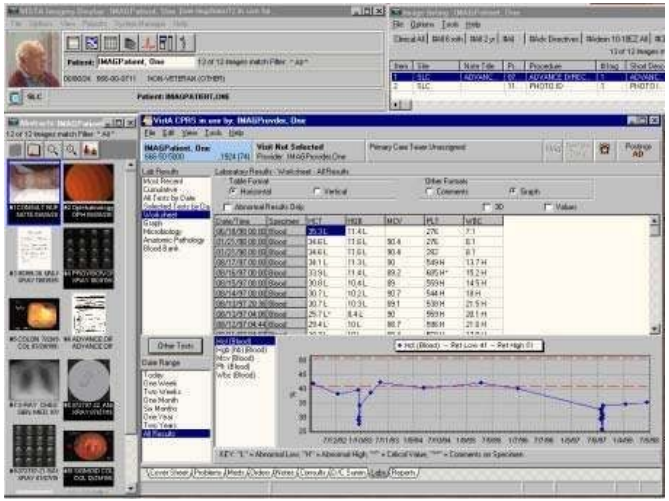
What are other examples of medical image data?

Electronic health records - making patient data available

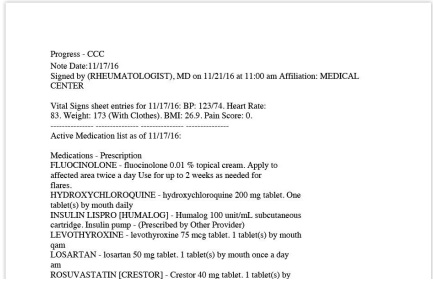


Imaging data

Patient measurements



1960s: invention
1980s: increased effort
2009: 51% adoption, HITECH Act
2017: 98% adoption (US)



Progress - CCC
Note Date:11/17/16
Signed by (RHEUMATOLOGIST), MD on 11/21/16 at 11:00 am Affiliation: MEDICAL CENTER

Vital Signs sheet entries for 11/17/16: BP: 123/74. Heart Rate: 83. Weight: 173 (With Clothes). BMI: 26.9. Pain Score: 0.

Active Medication list as of 11/17/16:

Medications - Prescription
FLUOCINOLONE - fluocinolone 0.01 % topical cream. Apply to affected area twice a day Use for up to 2 weeks as needed for flares.
HYDROXYCHLOROQUINE - hydroxychloroquine 200 mg tablet. One tablet(s) by mouth daily
INSULIN LISPRO [HUMALOG] - Humalog 100 unit/mL subcutaneous cartridge. Insulin pump - (Prescribed by Other Provider)
LEVOTHYROXINE - levothyroxine 75 mcg tablet. 1 tablet(s) by mouth qam
LOSARTAN - losartan 50 mg tablet. 1 tablet(s) by mouth once a day am
ROSUVASTATIN [CRESTOR] - Crestor 40 mg tablet. 1 tablet(s) by

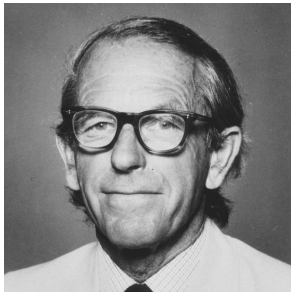
Clinical notes

Test	Value	Reference
Hemoglobin	8.0	8.5–11.0 mmol/L
C-reactive protein	279	<5 mg/L
Red blood cell count	3.86	4.3–6.0 × 10^{12}/L
White blood cell count	27.1	4.0–10.0 × 10^{9}/L
Thrombocytes	462	150–400 × 10^{9}/L
Glucose	12.9	4.0–7.8 mmol/L
Sodium	127	135–145 mmol/L
Potassium	4.2	3.5–5.0 mmol/L
Creatinine	40	50–110 µmol/L
Estimated glomerular filtration rate	>90	>60 ml/min
Ureum	3.2	2.5–7.5 mmol/L
Lactate dehydrogenase	166	<250 U/L
Aspartaat aminotransferase	14	<40 U/L
Alanine aminotransferase	13	<50 U/L
Alkaline phosphatase	127	<120 U/L
Gamma-glutamyl transferase	96	<50 U/L

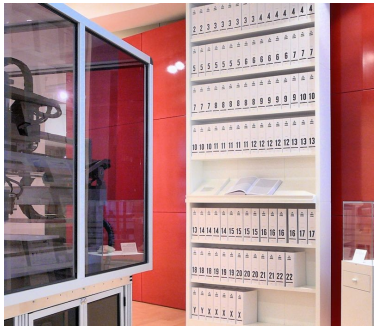
Lab results

Genomics data

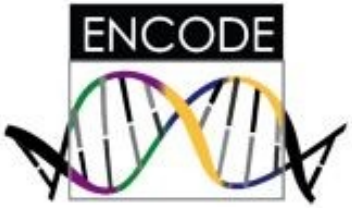
1953 - Watson and Crick discover double helix structures of DNA



1977 - Fred Sanger sequences first full genome of a virus



1990 - 2003: Human Genome Project sequences full human genome



2003: ENCODE project launched to identify and characterize genes in human genome

1000 Genomes Project: 2008 - 2015

UK100,000 Genomes Project: 2012 - 2018

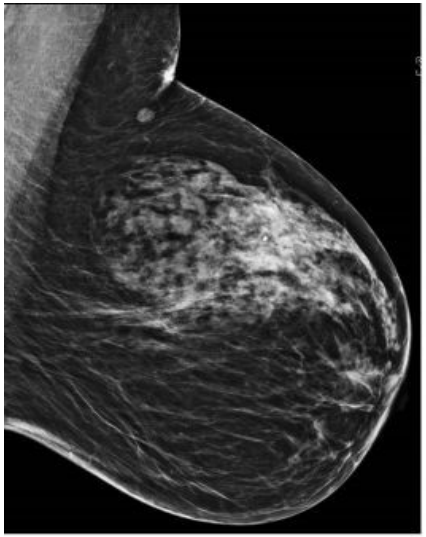
Wearables and other sensor data

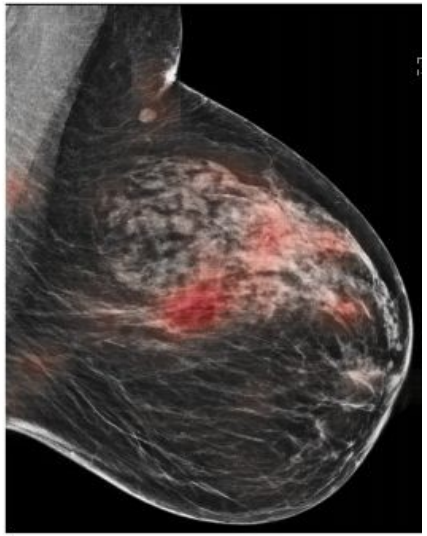
First iPhone: 2007

Fitbit: 2009

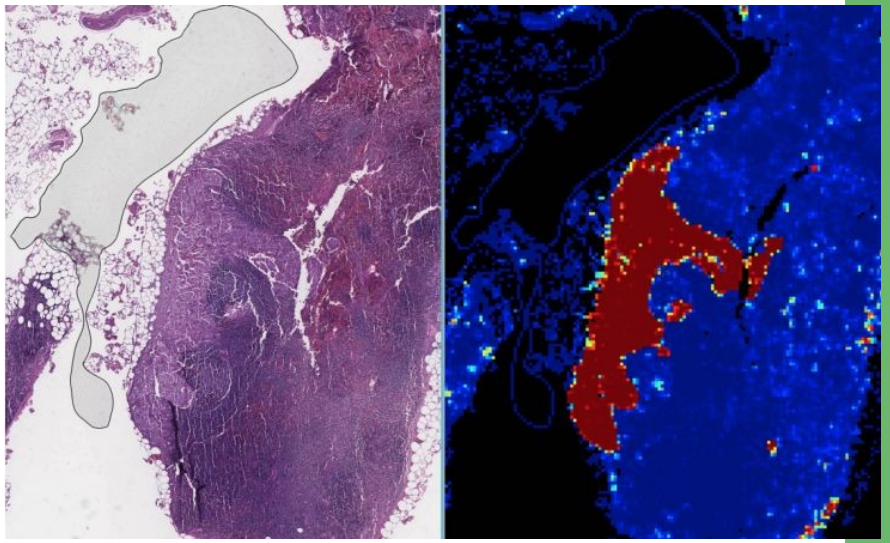
Apple Watch: 2014

AI in healthcare: biomedical image interpretation



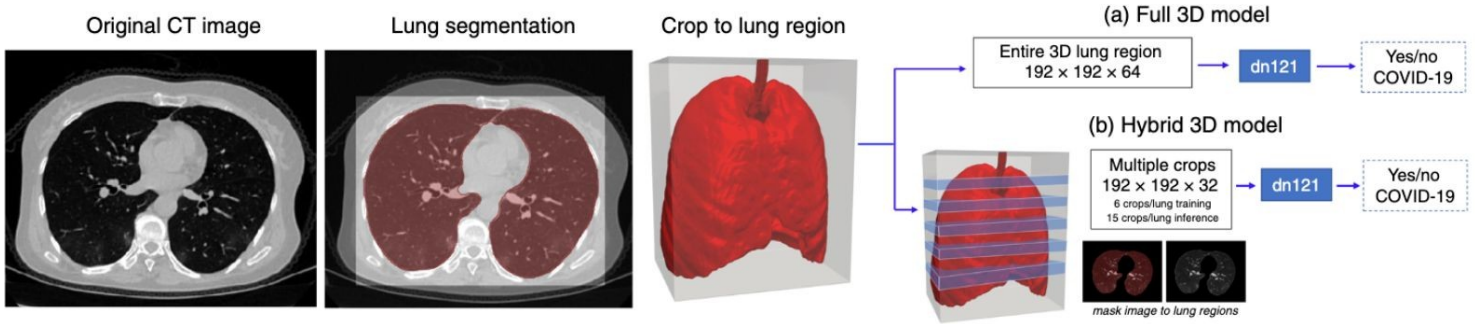


Wu et al. 2019

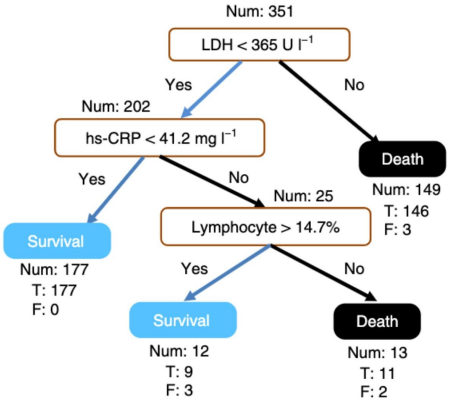


Liu et al. 2017

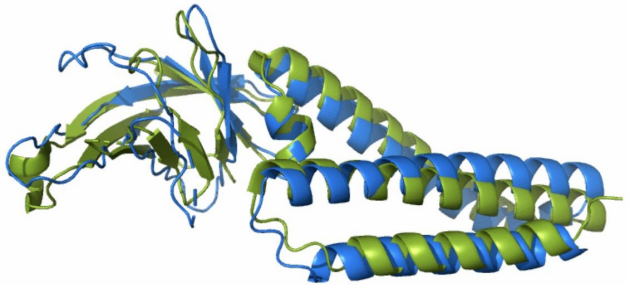
AI in healthcare: recent applications for COVID-19



Harmon et al. 2020



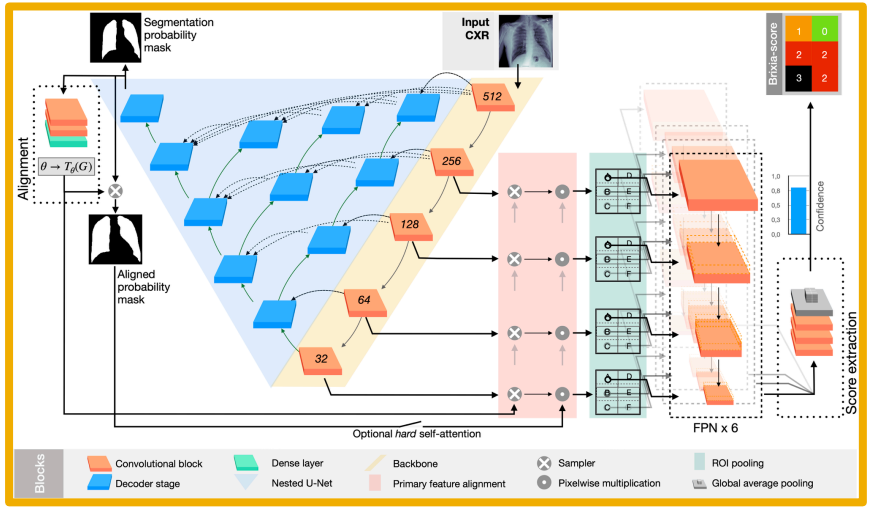
Yan et al. 2020

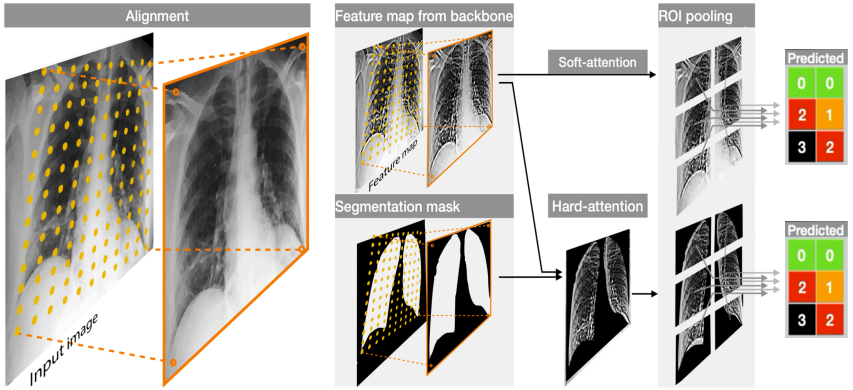


Jumper et al. 2020

AI in healthcare: recent applications for COVID-19

- BrixIA-Dataset
  - ~5000 Chest X-rays fro COVID-19 subjects (Mar-Apr 2020)
- BrixIA-Net
  - ~21 Milion parameters
- Public release of code and dataset
  - https://brixia.github.io
- Scientific paper
  - https://doi.org/10.1016/j.media.2021.102046
  - Medical Image Analysis (IF 8.5)
- Media
  - RAI TG1 Medicina





THE PROMISE IS GREAT…
BUT MANY OPEN CHALLENGES IN DEPLOYMENT AS WELL

La radiologia medica (2020) 125:517–521
https://doi.org/10.1007/s11547-020-01135-9

EDITORIAL

Artificial intelligence: Who is responsible for the diagnosis?

Emanuele Neri[1] · Francesca Coppola[2] · Vittorio Miele[3] · Corrado Bibbolino[4] · Roberto Grassi[5]

Insights into Imaging (2018) 9:745–753
https://doi.org/10.1007/s13244-018-0645-y

REVIEW

Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States

Filippo Pesapane[1] · Caterina Volonté[2] · Marina Codari[3] · Francesco Sardanelli[3,4]

nature machine intelligence

ANALYSIS

https://doi.org/10.1038/s42256-021-00307-0

Check for updates

OPEN

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts[1,2]✉, Derek Driggs[1], Matthew Thorpe[3], Julian Gilbey[1], Michael Yeung[4], Stephan Ursprung[4,5], Angelica I. Aviles-Rivero[1], Christian Etmann[1], Cathal McCague[4,5], Lucian Beer[4], Jonathan R. Weir-McCall[4,6], Zhongzhao Teng[4], Effrossyni Gkrania-Klotsas[7], AIX-COVNET*, James H. F. Rudd[8,36], Evis Sala[4,5,36] and Carola-Bibiane Schönlieb[1,36]

Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

Data driven approach to Medical Image Analysis

- A "checklist"

Insights into Imaging

EDUCATIONAL REVIEW Open Access

Deep learning workflow in radiology: a primer

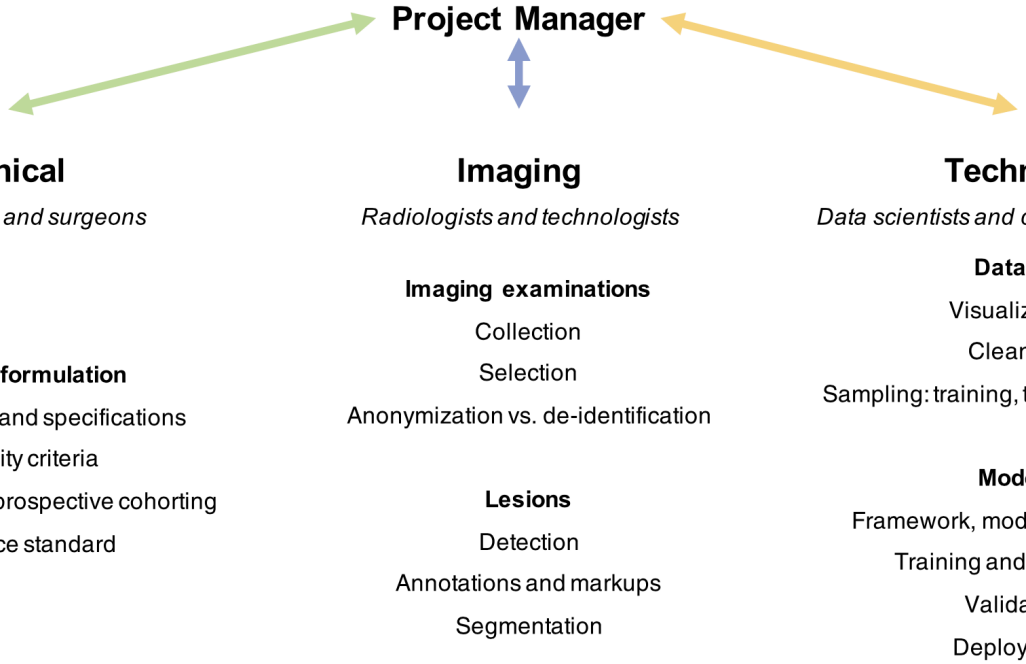
Emmanuel Montagnon[1], Milena Cerny[1], Alexandre Cadrin-Chênevert[2], Vincent Hamilton[1], Thomas Derennes[1], André Ilinca[1], Franck Vandenbroucke-Menu[1,4], Simon Turcotte[1,4], Samuel Kadoury[5] and An Tang[1,3]*

Montagnon et al. Insights into Imaging (2020) 11:22
https://doi.org/10.1186/s13244-019-0832-5

Scope	❑ Define scope of project: detection, segmentation, classification, monitoring, prediction or prognosis.
Team building	❑ Project manager (e.g, physician, data scientist) ❑ Clinical expertise (e.g., surgeon or hepatologist) ❑ Imaging expertise (e.g., radiologist) ❑ Technical expertise (e.g., data scientist)
Ethics	❑ Obtain IRB approval
Cohorting	❑ Selection process (e.g., by target population vs. database) ❑ Definition of eligibility criteria ❑ Identification of data source
Data	*De-identification* ❑ Data anonymization vs. pseudonymization *Collection and curation* ❑ Data collection ❑ Data exploration and quality control ❑ Labeling = markup and annotations ❑ Reference standard (synonyms: ground truth or gold standard) *Sampling* ❑ Creation of training, validation and test datasets ❑ Alternative: cross-validation
Model	❑ Defining performance metrics ❑ Selection of model (convolutional, recurrent, fully connected) and librairies ❑ Running the experiment followed by hyperparameters fine tuning ❑ Testing: assessing performance on separate test dataset
Hardware	❑ Determine best configuration based on model architecture and memory requirements ❑ Local (CPUs vs. GPUs) vs. cloud computing (GPUs vs. TPUs)
Regulatory	❑ Market research to inform decision to commercialize ❑ Quality management system ❑ Compliance with local regulatory jurisdictions
Clinical adoption	❑ Integration in distribution platform ❑ Clinical validation of performance ❑ Deployment in clinical practice

Team building…

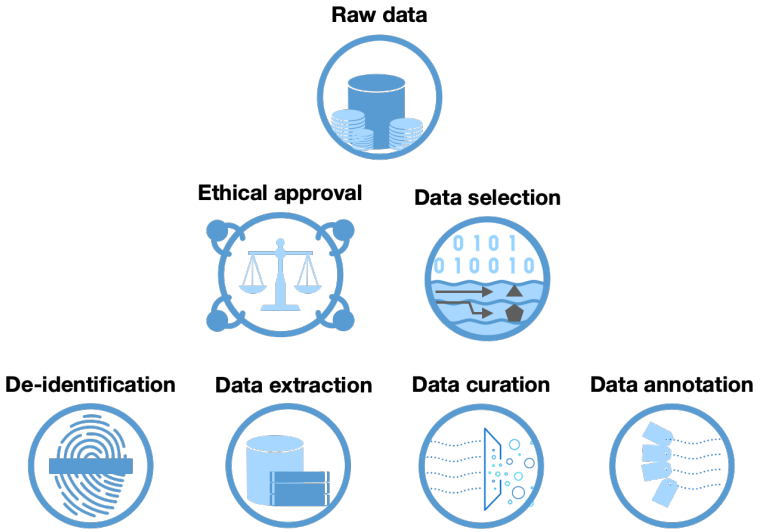
- …where people from different fields and levels of expertise are gathered to share their knowledge and collaborate on a joint project



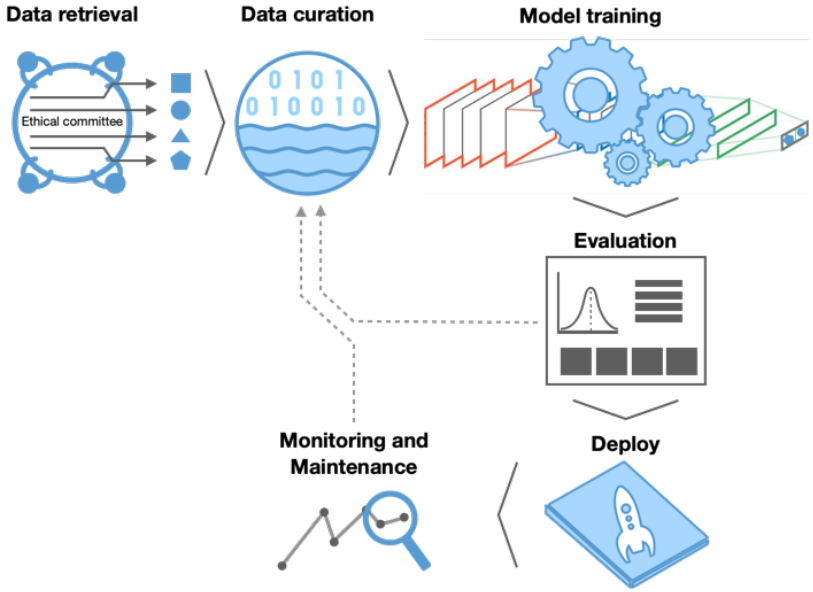
Montagnon E, Cerny M, Cadrin-Chênevert A, et al (2020) **Deep learning workflow in radiology: a primer.** Insights Imaging 11:22

Data driven approach in Medical Image Analysis

- Data preparation/curation/annotation



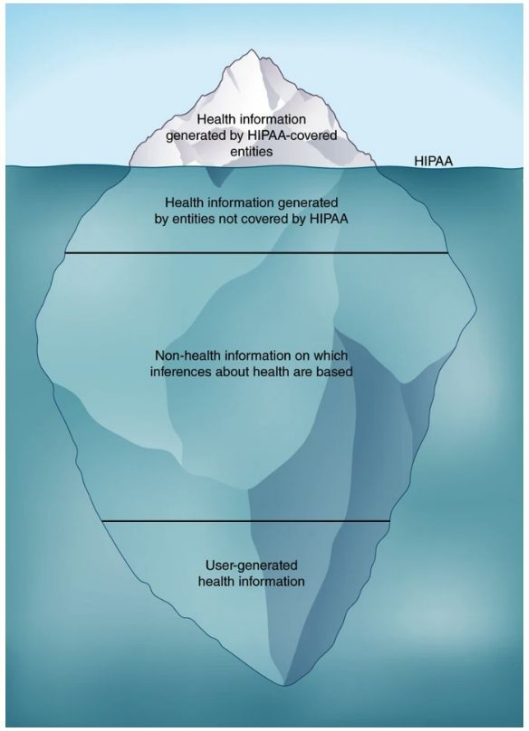
- The global workflow (feedback loops)



Design and Deployment of Data-driven solutions for Medical Image Analysis
Alberto Signoroni, Mattia Savardi
CNIT Technical reports, n.8 ch.2, December 2021 (in press)

Privacy and security

(patient) Privacy laws:
EU GDPR
US HIPAA



Price et al. 2019

Federated learning can reduce privacy concerns in multicentre studies

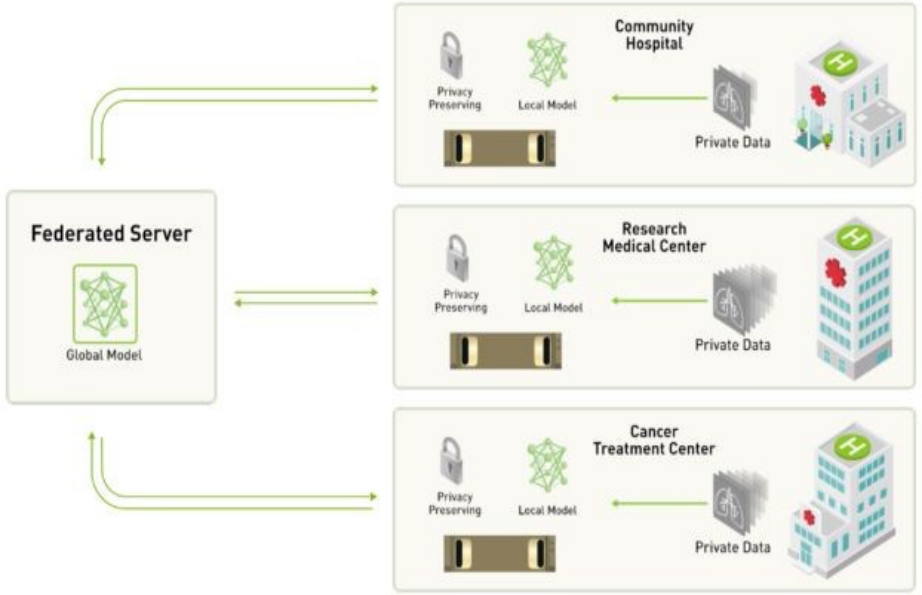
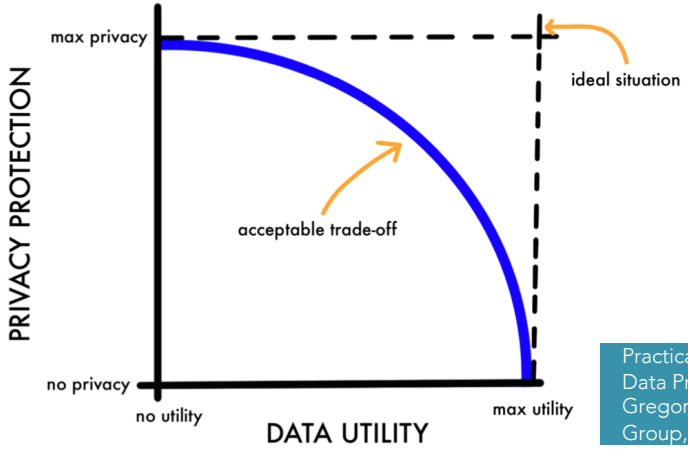


Figure: https://news.developer.nvidia.com/first-privacy-preserving-federated-learning-system/

Data de-identification, anonymization, and pseudonymization

- Three important concept/procedures to protect patient identity/privacy while collecting/sharing data
  1. **De-identification** refers to the masking patient-related information from individual records in order to minimize the risks of identification and breach of privacy;
  2. **Anonymization**, a subtype of de-identification, refers to the irreversible removal of patient-related information from individual records. It is the preferred approach for sharing of medical data.
  3. **Pseudonymization**, a subtype of de-identification, refers to the substitution of patient-related information with artificial values in a way that the original data can only be revealed with a secret key



Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification Gregory Nelson Conference: SAS Global Users Group, 2015

Montagnon E, Cerny M, Cadrin-Chênevert A, et al (2020) **Deep learning workflow in radiology: a primer.** Insights Imaging 11:22

Data Curation

☐ Data exploration and quality control

- → assessing general qualitative (e.g., through visualization) or quantitative properties (e.g., through statistics) of the initial raw dataset, in order to exhibit specific features, global trends, or outliers

☐ Data labeling (markup and annotation)

- After selection of appropriate images, **data labeling** may require delineating lesions, either through bounding boxes (for detection) or pixel-wise contours (segmentation masks) accompanied by annotations on the type of lesions and their location
- **Markup** refers to "graphic symbols placed on an image to depict an annotation," whereas annotation refers to explanatory or descriptive information regarding the meaning of an image that is generated by a human observer

☐ Reference standard

- a.k.a. "ground truth," represents the knowledge that the model is expected to learn;
- depending on the project, the choice of reference standard may include
  - (1) histopathology,
  - (2) follow-up examinations,
  - (3) alternative imaging modality (e.g. MRI or PET while targeting CT or X-ray), or
  - (4) clinical outcomes (e.g., time to tumor recurrence, disease-specific survival).

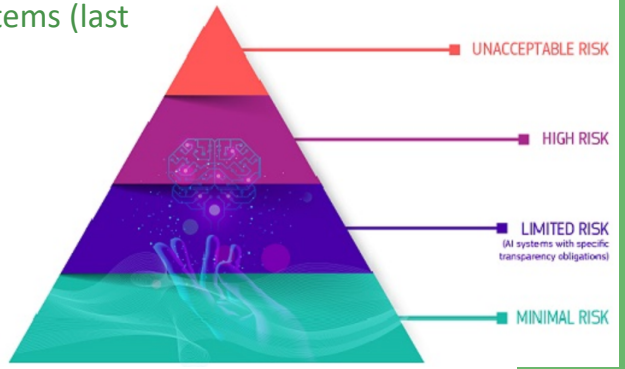
Large amount of medical data & data curation

- A large amount of medical data are available with moderate ease in access and retrieval
  - data can be clinical data (biobank), images, and related metadata (DICOM), human annotations (radiology reports), and machine-generated features
  - rarely curated → major bottleneck in attempting to learn any AI model
- **curation is the most time-consuming step in an AI project, but is critical to any model training**
- this step still requires human knowledge and supervision to achieve high-quality datasets
- exacerbated in uncommom diseases (a limited number of human readers with expertise): **unsupervised learning** → *generative adversarial networks* and *variational autoencoders* show great promise, as discriminative features are learned without explicit labelling

Montagnon E, Cerny M, Cadrin-Chênevert A, et al (2020) **Deep learning workflow in radiology: a primer.** Insights Imaging 11:22

Regulatory perspective

- Since the earliest days of computing, FDA (US Food and Drug Administration) had been regulating Computer Aided Diagnosis systems that rely on machine learning and pattern-recognition techniques;
- the shift to deep learning now poses new regulatory challenges and requires new guidance for submissions seeking approval
  - DL methods evolve over time as more data are processed and learned from,
  - which implications of a lifelong learning of these adaptive systems?
  - a periodic testing could potentially ensure that learning and its associated prediction performance?
- European legal framework for AI to address fundamental rights and safety risks specific to the AI systems
  - EU rules to address liability issues related to new technologies, including AI systems (last quarter 2021-first quarter 2022);
  - Safety components of products (e.g. AI application in robot-assisted surgery)



Hosny A, Parmar C, Quackenbush J, et al (2018) **Artificial intelligence in radiology**. Nat Rev Cancer 18:500–510.

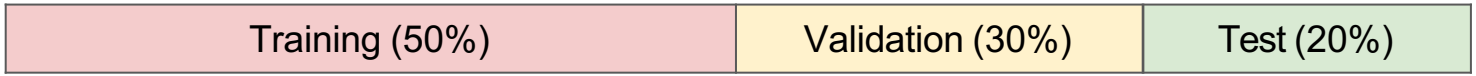
Perspectives on AI – radiologist interplay

- AI is unlike human intelligence: excelling in one task does not necessarily imply excellence in others
  - almost all advances in AI fall under the narrow AI (a.k.a. **weak AI**) category, where AI is trained for one task and one task only. Such advances lack higher-level, top-down knowledge of contexts and fail to make associations the way a human brain does.
- AI is unlikely to replace radiologists within the near/even distant future: the roles of radiologists will expand as they become more connected to technology and have access to better tools.
- As different forms of AI exceed human performance, we expect it to evolve into a valuable educational resource.
- Human operators will not only oversee outcomes but also seek to interpret the reasoning behind them — as a means of validation and as a way to potentially discover hidden information that might have been overlooked;
- → open questions include the ambiguity of who controls AI and is ultimately responsible for its actions, the nature of the interface between AI and health care and whether implementation of a regulatory policy too soon will cripple AI application efforts.

Hosny A, Parmar C, Quackenbush J, et al (2018) **Artificial intelligence in radiology**. Nat Rev Cancer 18:500–510.

DATA CONSIDERATIONS FOR IMAGE CLASSIFICATION MODELS

Training, validation, and test sets



Held-out evaluation set for selecting best hyperparameters during training

Do not use until final evaluation

Other splits e.g. 60/20/20 also popular.
Balance sufficient data for training vs. informative performance estimate on validation / testing.

Maximizing training data for the final model

Once hyperparameters are selected using the validation set, common to merge training and validation sets into a larger “trainval” set to train a final model using the hyperparameters.

This is OK, since we can use non-test data however we want during model development!

K-fold cross validation: for small datasets

Sometimes we have small labeled datasets in healthcare… in this case K-fold cross validation (which is more computationally expensive) may be worthwhile.

Fold 1	Fold 2	Fold 3	Fold 4	Test
Fold 1	Fold 2	Fold 3	Fold 4	Test
Fold 1	Fold 2	Fold 3	Fold 4	Test
Fold 1	Fold 2	Fold 3	Fold 4	Test

Train model K times with a different fold as the validation set each time; then average the validation set results. Allows more data to be used for each training of the model, while still using enough data to get accurate validation result.

Can also apply same concept to test-time evaluation.

How much data do you need for deep learning?

Examples per class of your dataset, in addition to transfer learning (take this with grain of salt, it really depends on the problem):

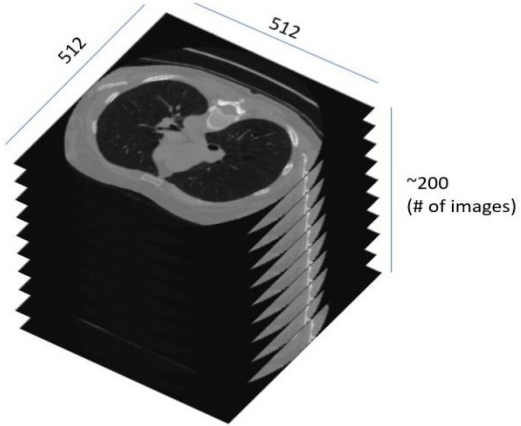
- Low dozens: generally too small to learn a meaningful model, using standard supervised deep learning
- High dozens to low hundreds: may see models with some predictive ability, unlikely to really wow or be “superhuman” though
- High hundreds to thousands: “happy regime” for deep learning

	very similar dataset	**very different dataset**
very little data	Use Linear Classifier on top layer features	You’re in trouble… Try linear classifier on different layer features
quite a lot of data	Finetune a few layers	Finetune a large number of layers

In general, deep learning is data hungry -- the more data the better

Almost always leverage transfer learning unless you have extremely different or huge (e.g. ImageNet-scale) dataset

What counts as a data example?



1 3D CT volume with 200 slices／ 200 data examples

5 surgery videos with thousands of frames each／ thousands of data examples

Guidelines for amount of training data refers to # of unique instances representative of diversity expected during testing / deployment. E.g. # of independent CT scans or surgery videos. Additional correlated data (e.g. different slices of the same tumor or different suturing instances within the same video) provide relatively less incremental value in comparison.

What if there are multiple possible sources of data?

E.g., some with noisier / less accurate labels than others, from different hospital sites, etc.

- Expected diversity of data during deployment should be reflected in both training and test sets
  - Need to see these during training to learn how to handle them
  - Need to see these during testing to accurately evaluate the model

- Want test set labels to be as accurate as possible

- Noisy labels is often still useful during training -- can provide useful signal in aggregate.
  Much larger amount, but noisy, data is often better than small but clean data.
  - "Weakly supervised learning" is a major area of research focused on learning with large amounts of noisy or imprecise labels

EVALUATING IMAGE CLASSIFICATION MODELS

Evaluation metrics

Confusion matrix

Ground Truth \ Prediction	0	1
0	TN	FP
1	FN	TP

Accuracy: (TP + TN) / total

Q: When might evaluating purely accuracy be problematic?

A: Imbalanced datasets.

Evaluation metrics

Confusion matrix

Ground Truth \ Prediction	0	1
0	TN	FP
1	FN	TP

We can trade-off different values of these metrics as we vary our classifier's score threshold to predict a positive

Accuracy: (TP + TN) / total

Sensitivity / Recall (true positive rate)**:**
TP / total positives → TP+FN

Specificity (true negative rate)**:**
TN / total negatives → TN+FP

Precision (positive predictive value)**:**
TP / total predicted positives → TP+FP

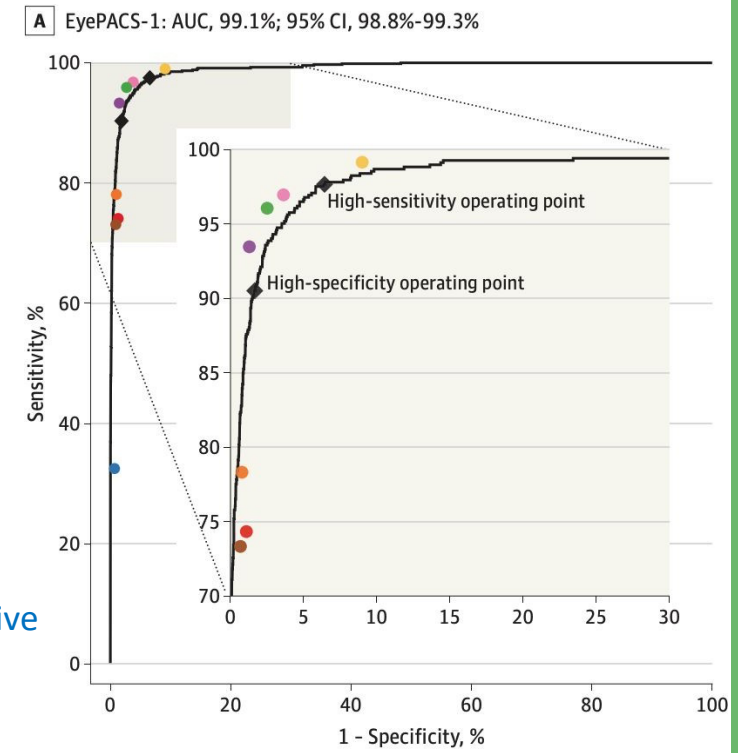
Negative predictive value:
TN / total predicted negatives

Q: As prediction threshold increases, how does that generally affect sensitivity? Specificity?

A: Sensitivity goes down, specificity up

Evaluation metrics

- **Receiver Operating Characteristic (ROC) curve**:
  - Plots sensitivity and specificity (specifically, 1 - specificity) as prediction threshold is varied
  - Gives trade-off between sensitivity and specificity
  - Also report summary statistic AUC (area under the curve)



True positive rate (TPR)

False positive rate (FPR)

Evaluation metrics

- Sometimes also see **precision recall curve**
  - More informative when dataset is heavily imbalanced (sensitivity = true negative rate less meaningful in this case)

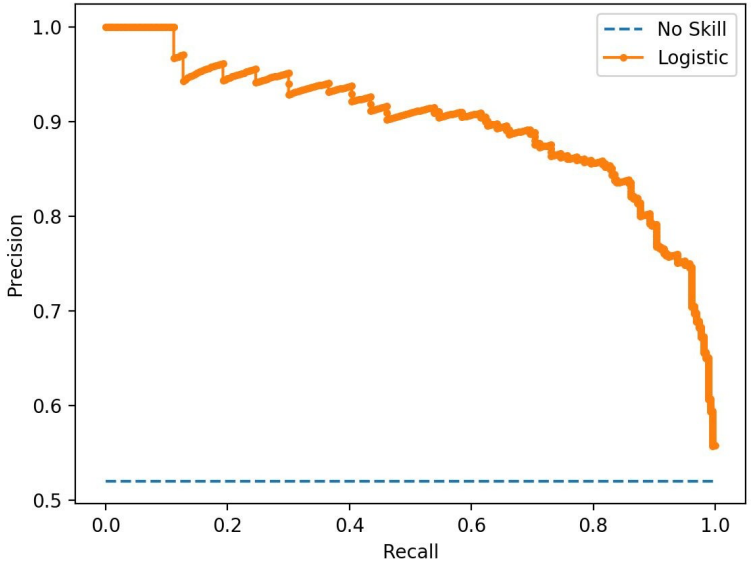


Figure credit: https://3qeqpr26caki16dnhd19sv6by6v-wpengine.netdna-ssl.com/wp-content/uploads/2018/08/Precision-Recall-Plot-for-a-No-Skill-Classifier-and-a-Logistic-Regression-Model4.png

Evaluation metrics

- Selecting optimal trade-off points
- Maximize **Youden's Index**
  - J = sensitivity + specificity - 1
  - Gives equal weight to optimizing true positives and true negatives
- Sometimes also see F-measure (or F1 score)
  - F1 = 2*(precision*recall) / (precision + recall)
  - Harmonic mean of precision and recall

But selected trade-off points could also depend on application

Also equal to distance above chance line for a balanced dataset: sensitivity - (1 - specificity) = sensitivity + specificity - 1

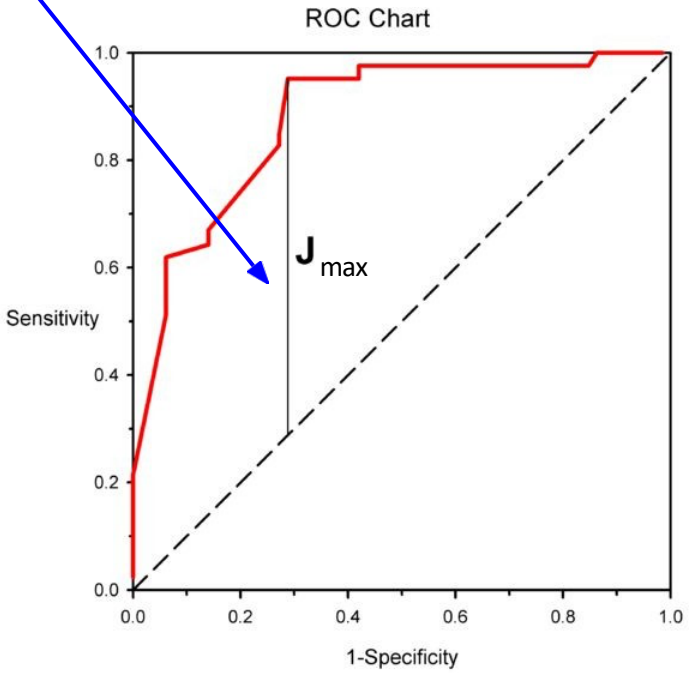


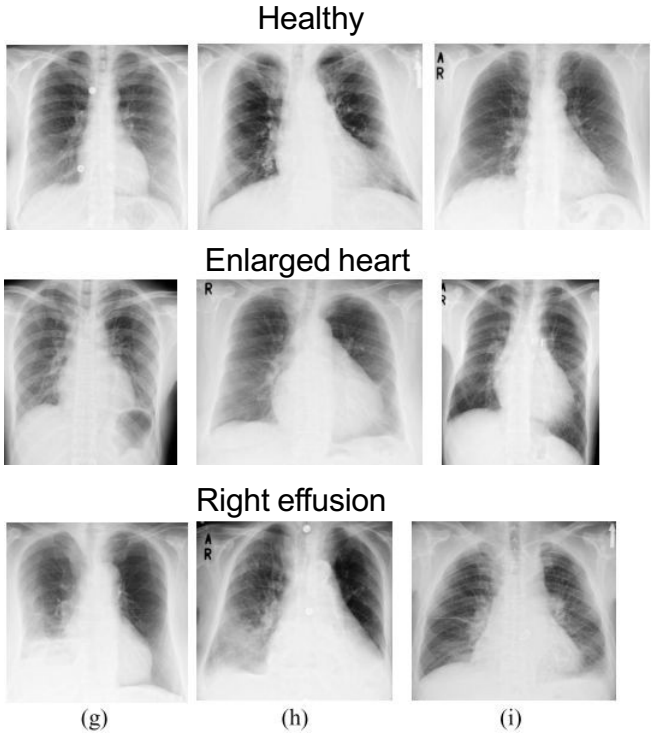
Figure credit: https://en.wikipedia.org/wiki/File:ROC_Curve_Youden_J.png

CASE STUDIES OF CNNS FOR MEDICAL IMAGING CLASSIFICATION

Early steps of deep learning in medical imaging: using ImageNet CNN features

Q: How might we approach this problem?

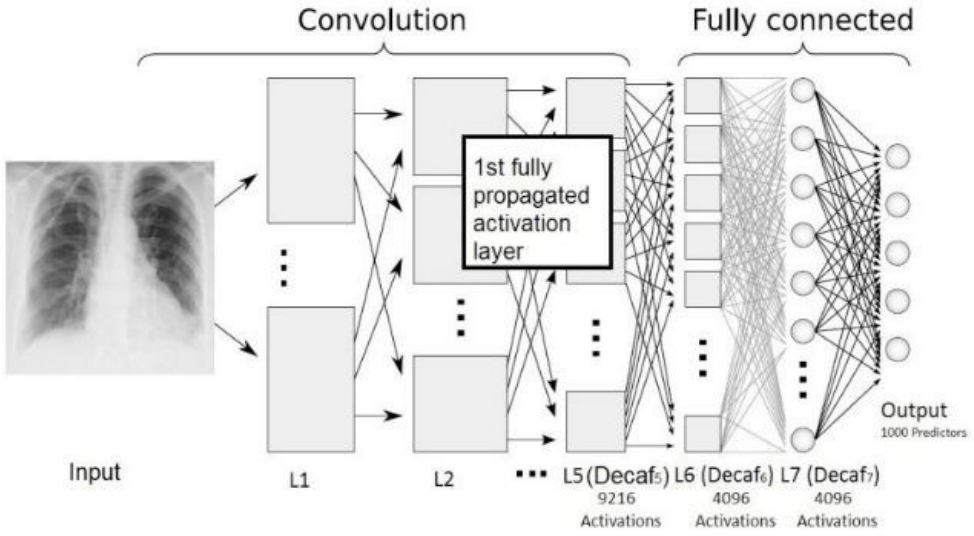
- Input: Chest **x-ray images**
- Output: Several binary classification tasks
  - Right pleural effusion or not
  - Enlarged heart or not
  - Healthy or abnormal
- Very small dataset: 93 frontal chest x-ray images



Bar et al. Deep learning with non-medical training used for chest pathology identification. SPIE, 2015.

Bar et al. 2015

- Did not train a deep learning model on the medical data
- Instead, extracted features from an AlexNet trained on ImageNet
  - 5th, 6th, and 7th layers
- Used extracted features with an SVM classifier
- Performed zero-mean unit-variance normalization of all features
- Evaluated combination with other hand-crafted image features



Bar et al. Deep learning with non-medical training used for chest pathology identification. SPIE, 2015.

Bar et al. 2015

Q: How might we interpret the AUC vs. CNN feature trends?

Table 1. Right Pleural Effusion Condition.

	Low Level		High Level	Deep			Fusion
	LBP	GIST	PiCoDes	Decaf L5	Decaf L6	Decaf L7	PiCoDes+Decaf L5
Sensitivity	0.71	0.79	0.79	0.93	0.86	0.86	**0.93**
Specificity	0.77	0.92	0.91	0.84	0.86	0.80	**0.84**
AUC	0.75	0.93	0.91	0.92	0.91	0.84	**0.93**

Table 2. Healthy vs. Pathology.

	Low Level		High Level	Deep			Fusion
	LBP	GIST	PiCoDes	Decaf L5	Decaf L6	Decaf L7	PiCoDes+Decaf L5
Sensitivity	0.65	0.68	0.59	0.73	0.89	0.76	**0.81**
Specificity	0.61	0.66	0.79	0.80	0.64	0.64	**0.79**
AUC	0.63	0.72	0.72	0.78	0.79	0.72	**0.79**

Table 3. Enlarged Heart Condition.

	Low Level		High Level	Deep			Fusion
	LBP	GIST	PiCoDes	Decaf L5	Decaf L6	Decaf L7	PiCoDes+Decaf L5
Sensitivity	0.75	0.79	0.79	0.88	0.79	0.79	**0.83**
Specificity	0.78	0.81	0.84	0.78	0.88	0.77	**0.84**
AUC	0.80	0.82	0.87	0.87	0.84	0.79	**0.89**

Bar et al. Deep learning with non-medical training used for chest pathology identification. SPIE, 2015.

Gulshan et al. 2016

- **Task**: Binary classification of referable diabetic retinopathy from **retinal fundus photographs**
- **Input**: Retinal fundus photographs
- **Output**: Binary classification of referable diabetic retinopathy (y in {0,1})
  - Defined as moderate and worse diabetic retinopathy, referable diabetic macular edema, or both

Gulshan, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 2016.

Gulshan et al. 2016

- **Dataset**:
  - 128,175 images, each graded by 3-7 ophthalmologists.
  - 54 total graders, each paid to grade between 20 to 62508 images.
- **Data preprocessing**:
  - Circular mask of each image was detected and rescaled to be 299 pixels wide
- **Model**:
  - Inception-v3 CNN, with ImageNet pre-training
  - Multiple BCE losses corresponding to different binary prediction problems, which were then used for final determination of referable diabetic retinopathy

Gulshan, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 2016.

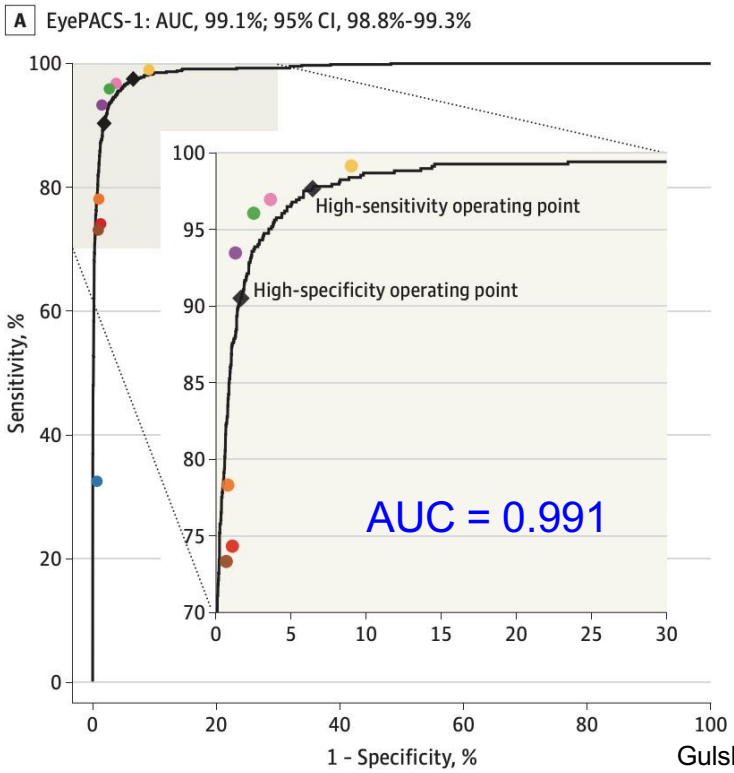
Gulshan et al. 2016

Graders provided finer-grained labels which were then consolidated into (easier) binary prediction problems

- **Dataset**:
  - 128,175 images, each graded by 3-7 ophthalmologists.
  - 54 total graders, each paid to grade between 20 to 62508 images.
- **Data preprocessing**:
  - Circular mask of each image was detected and rescaled to be 299 pixels wide
- **Model**:
  - Inception-v3 CNN, with ImageNet pre-training
  - Multiple BCE losses corresponding to different binary prediction problems, which were then used for final determination of referable diabetic retinopathy

Gulshan, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 2016.

Gulshan et al. 2016



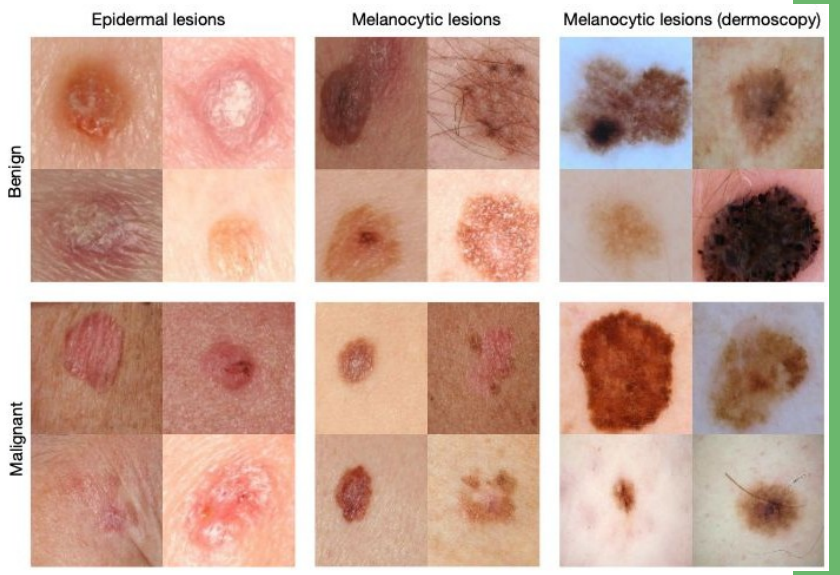
Looked at different operating points

- High-specificity point approximated ophthalmologist specificity for comparison. Should also use high-specificity to make decisions about high-risk actions.
- High-sensitivity point should be used for screening applications.

Gulshan, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 2016.

Esteva et al. 2017

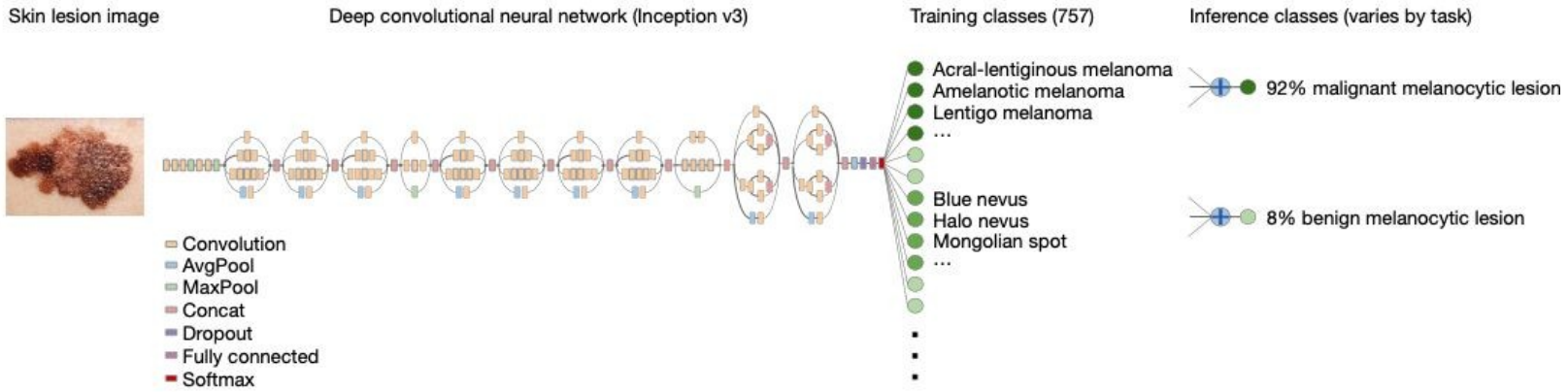
- Two binary classification tasks: malignant vs. benign lesions of epidermal or melanocytic origin
- Inception-v3 (GoogLeNet) CNN with ImageNet pre-training
- Fine-tuned on dataset of 129,450 lesions (from several sources) comprising 2,032 diseases
- Evaluated model vs. 21 or more dermatologists in various settings



Esteva*, Kuprel*, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017.

Esteva et al. 2017

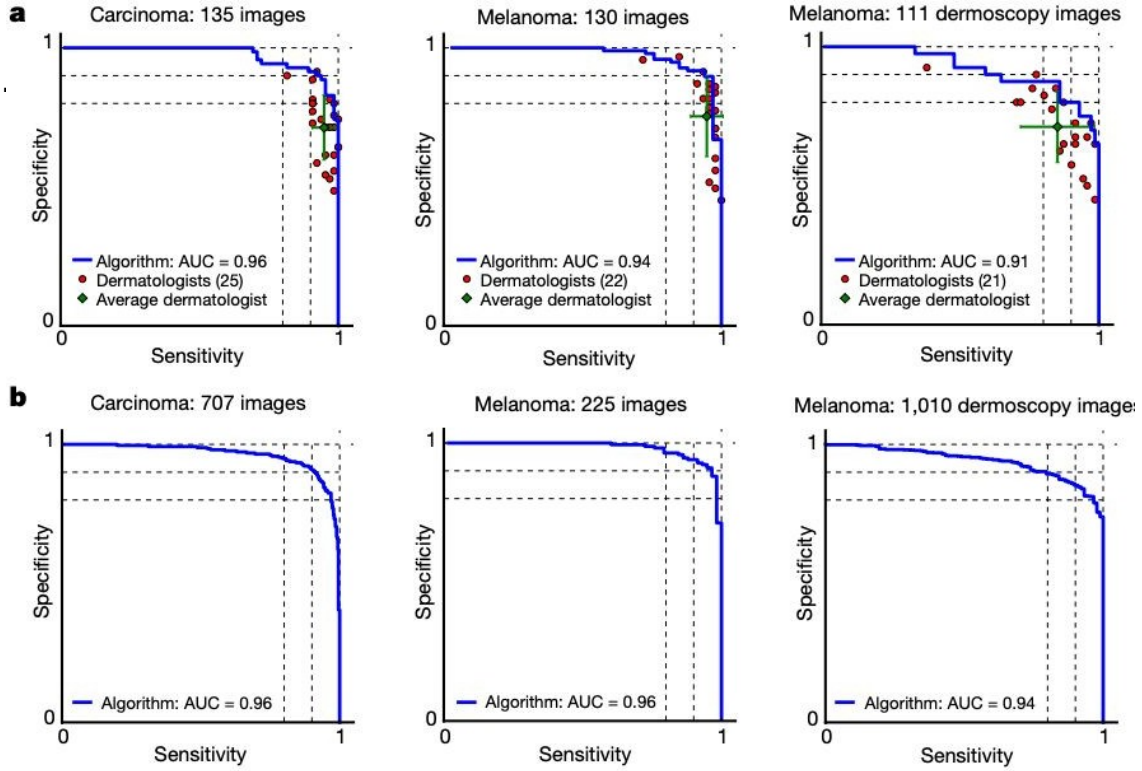
- Train on finer-grained classification (757 classes) but perform binary classification at inference time by summing probabilities of fine-grained sub-classes
- The stronger fine-grained supervision during the training stage improves inference performance!



Esteva*, Kuprel*, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017.

Esteva et al. 2017

- Evaluation of algorithm vs. dermatologists



Esteva*, Kuprel*, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017.

Lakhani and Sundaram 2017

- Binary classification of pulmonary tuberculosis from x-rays
- Four de-identified datasets
- 1007 chest x-rays (68% train, 17.1% validation, 14.9% test)
- Tried training CNNs from scratch as well as fine-tuning from ImageNet

AUC Test Dataset

Parameter	Untrained	Pretrained	Untrained with Augmentation*	Pretrained with Augmentation*
AlexNet	0.90 (0.84, 0.95)	0.98 (0.95, 1.00)	0.95 (0.90, 0.98)	0.98 (0.94, 0.99)
GoogLeNet	0.88 (0.81, 0.92)	0.97 (0.93, 0.99)	0.94 (0.89, 0.97)	0.98 (0.94, 1.00)
Ensemble				0.99 (0.96, 1.00)

Note.—Data in parentheses are 95% confidence interval.

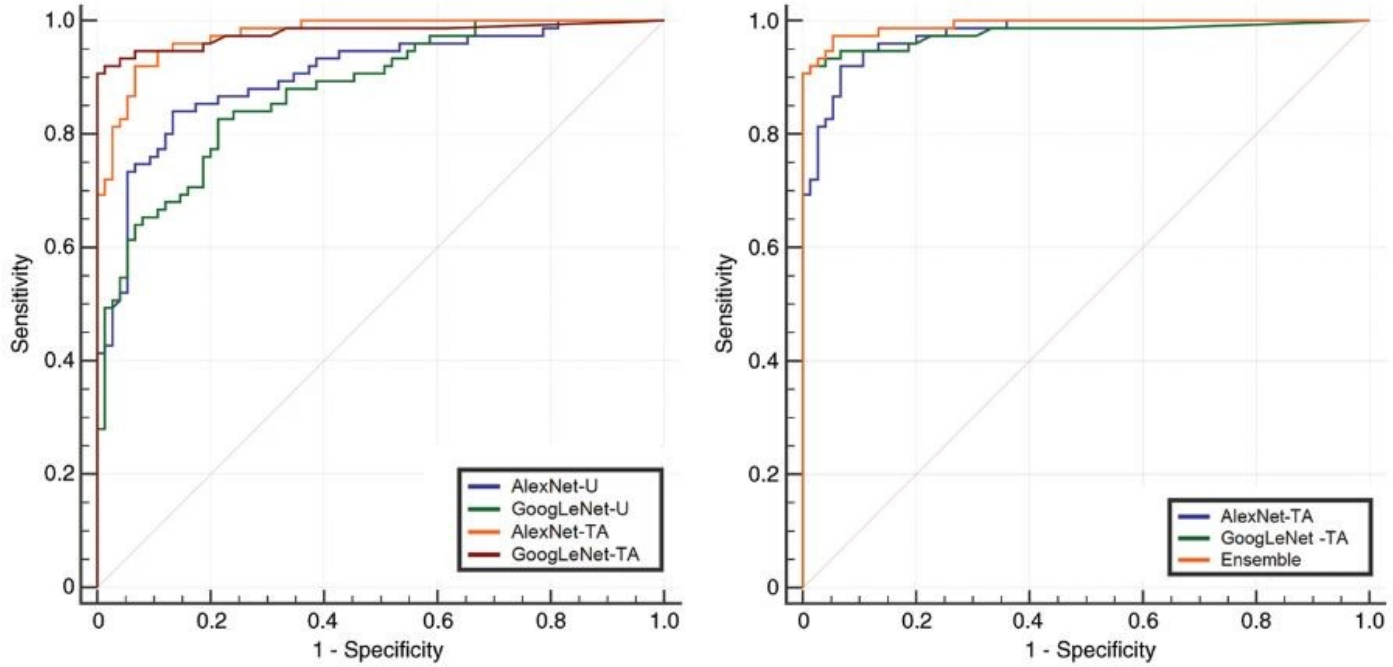
* Additional augmentation of 90, 180, 270 rotations, and Contrast Limited Adaptive Histogram Equalization processing.

All training images were resized to 256x256 and underwent base data augmentation of random 227x227 cropping and mirror images. Additional data augmentation experiments in results table.

Often resize to match input size of pre-trained networks. Also fine approach to making high-res dataset easier to work with!

Lakhani and Sundaram. Deep learning at chest radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology, 2017.

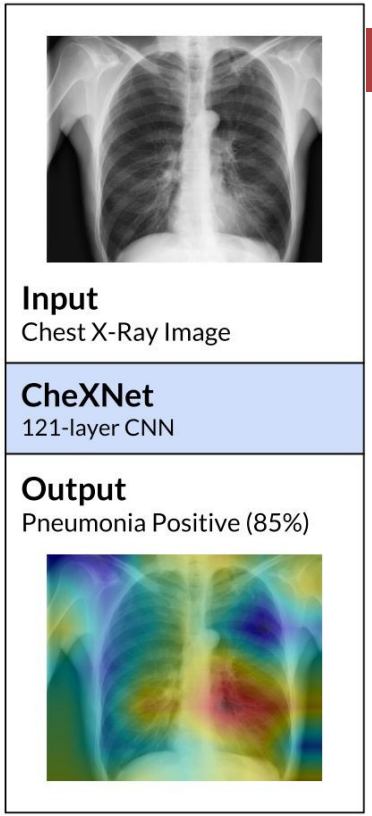
Lakhani and Sundaram 2017



Lakhani and Sundaram. Deep learning at chest radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology, 2017.

Rajpurkar et al. 2017

- Binary classification of pneumonia presence in chest X-rays
- Used ChestX-ray14 dataset with over 100,000 frontal X-ray images with 14 diseases
- 121-layer DenseNet CNN
- Compared algorithm performance with 4 radiologists
- Also applied algorithm to other diseases to surpass previous state-of-the-art on ChestX-ray14



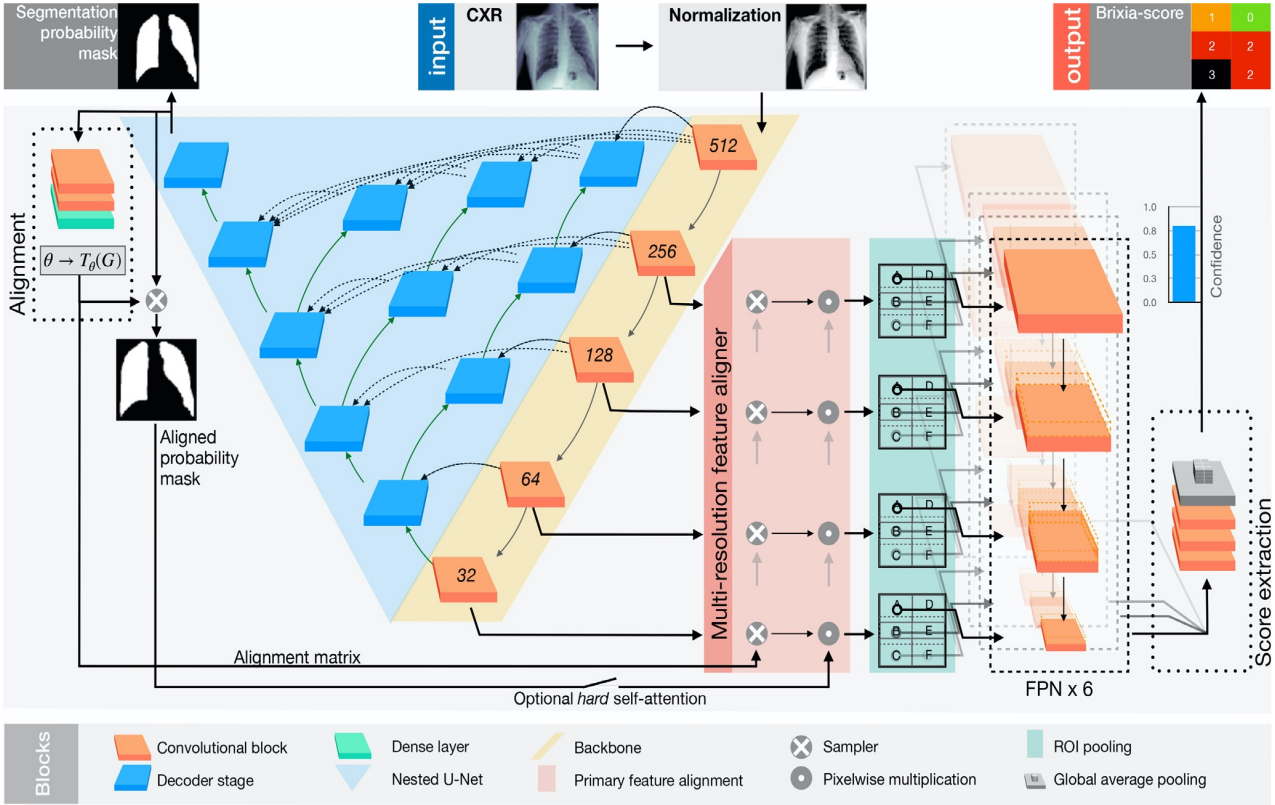
Rajpurkar et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017.

Signoroni, Savardi et al. 2021

- BS-Net for Severity estimation of COVID-19 pneumonia
- https://brixia.github.io

«Cognitive» workflow (multi-network)

1. Segmentation
2. Normalization (alignment)
3. Feature extraction (multi-scale, multi-region)
4. Score estimation (6-fold discrete regression)



Hackathon COVID CXR at Expo 2020, Dubai, UAE

Explainable AI for COVID-19 prognosis from early Chest X-ray and clinical data in the context of the COVID-CXR international hackathon

Edoardo Coppola

Damiano Ferrari

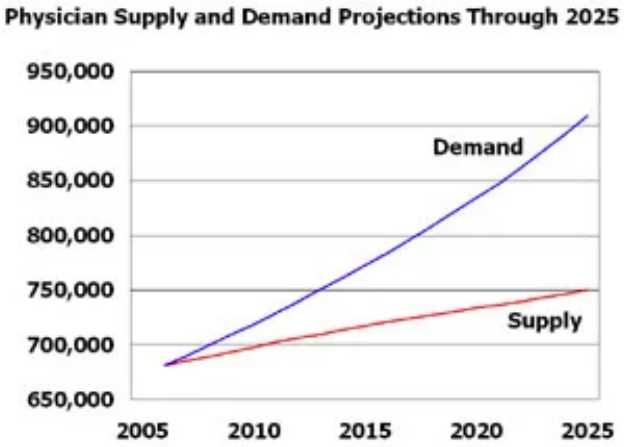
Dr. Savardi

Prof. Signoroni

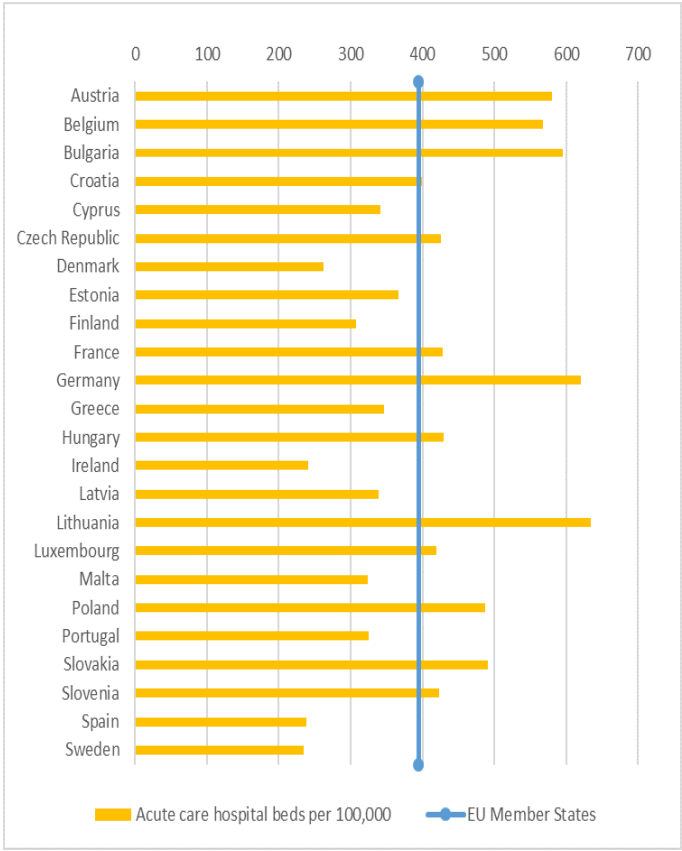
- https://github.com/ferraridamiano/covidcxr-hackathon

Introduction

- Shortage of physicians worldwide since 2005
  - As time goes by it is only going to be worse
- Limitation of (non) intensive care unit (ICU) beds
  - Lots of studies show the number of ICU beds is correlated with the number of deaths
- Year after year the costs in health care sector are increasing
  - So is longevity, which means the projected costs will boom up
- **Any resource in hospitals and clinic (personnel, machines, etc.) is limited and thus needs to be allocated optimally**

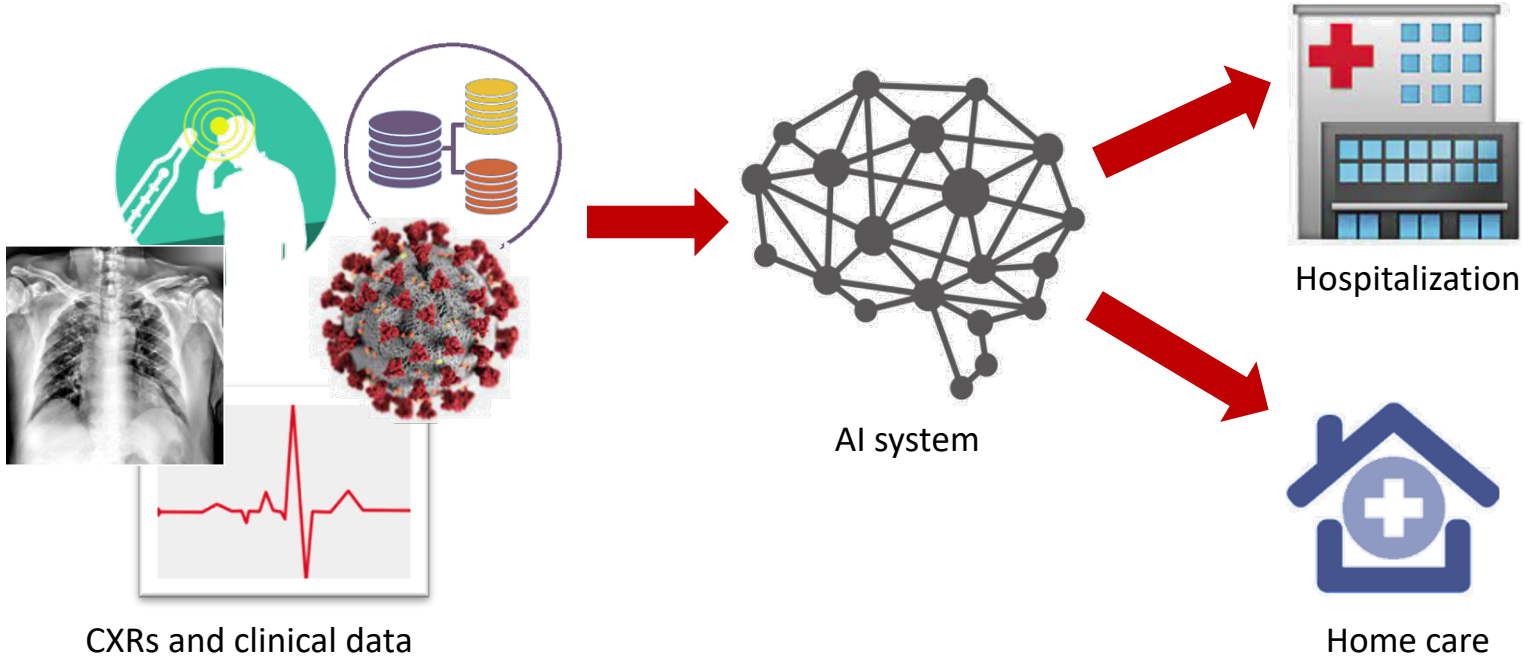


ICU beds per 100k inhabitants



Objectives of the project

- To develop an AI system able to distinguish COVID-19 patients with a mild pneumonia, and hence safely treatable at home, from those with a severe pneumonia requiring hospitalization
- Make the AI system predictions explainable/comprehensible to physicians as mush as possible



Project phases

- **Data analysis**
  - CXR images (quality, rotation, inversion, shape)
  - Clinical parameters (feature type, full or partial availability, feature importance)

- **Pre-processing CXRs and computing the Brixia-score**

- **Pre-processing the clinical data**

- **Building a pipeline of models to deal with (1) chest radiographs and (2) clinical parameters**
  - Testing different types of machine learning model to find which one shows the best balanced accuracy

- **Explaining the full model predictions**
  - Feature importance
  - SHAP values to find which features push the result to which class
  - Explainability maps of the lungs

Data analysis

- Complete dataset comprising almost **1600 records**, each of them with the associated chest radiograph and clinical parameters
  - **1100 records composed the training set**, while **the remaining were unlabelled** and were the instances on which we were called to make predictions in order to be evaluated and ranked

- The **label** describing the pneumonia severity was **categorical**: **mild or severe**
  - Which means this was a **binary classification task**
  - The **classes distribution was almost balanced**

- Clinical data was described by almost **40 features**
  - Both **numerical and categorical**
  - Partially or fully unavailable due to **missing values** (as is often the case with health care data)

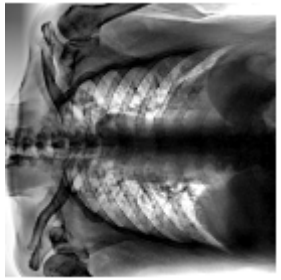
- The CXR images were **dark**, **shadowed**, **didn't share a common shape** and were in png format with a 12-bit color depth
  - Some of them were even **rotated** or **inverted**

Pre-processing CXRs images (1)

- **To deal with chest radiographs**, we decided to use **BS-net**: a multi-module deep neural network trained over 5000 CXR images to predict a **6-digit score** (**the Brixia-score**) **describing the compromise of the lungs ideally subdivided into as many parts.**
  - To use its modules, the radiographs must satisfy certain conditions like shape, pixel brightness range, pixel brigthness distribution etc.

- In order to obtain images like (2), we performed some image processing techniques on images like (1). The techniques performed were: **resizing** to 512x512, **normalization** to 0-1 range, **adaptive histogram equalization** (CLAHE, clip=0.01), **median filtering** (kernel size=3) and clipping outside the 2nd and 98th percentile

1) Original radiograph



2) Processed radiograph

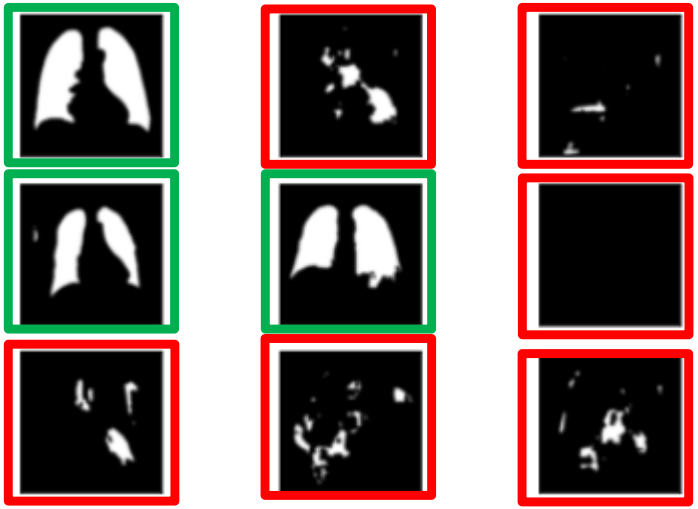
- **Thanks to these operations, dark and shadowed images got brighter and the lungs became much more visible**

- **Nevertheless, many images (like the one shown in this slide) were inverted and/or rotated by multiples of 90 degrees**

Pre-processing CXRs images (2)

- To solve the problem of inversion of the processed CXRs, we used the segmentation module of BS-net to get a segmentation map of the lungs alone.

- If an image was inverted, we couldn't obtain a good segmentation map of the lungs out of BS-net
  - Hence, check for the existence of two regions of a given size in the segmentation map was a good idea to see if an image was inverted or not

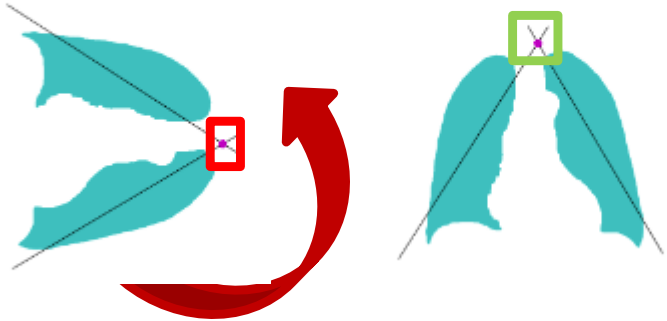


Examples of good (green) and bad (red) segmentation maps of the lungs. The bad ones come from inverted images

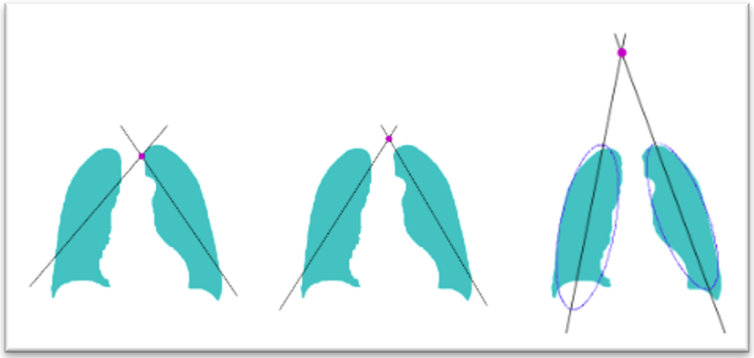
Pre-processing CXRs images (3)

- To solve the problem of rotation we fitted two lines with the pixels of the two regions found earlier and calculated their intersection point. If this point lied in the upper part of the image, then the chest radiograph was correctly oriented

- We tested three different techniques to fit these lines: least square method (LSM), RANSAC and ellipse fitting
  - The latter showed to be the best in terms of number CXR images correctly identified as not oriented in the right direction



The CXR not correctly oriented (left) is found thanks to the erroneous position of the intersection point. A 90° counter clockwise rotation is then performed

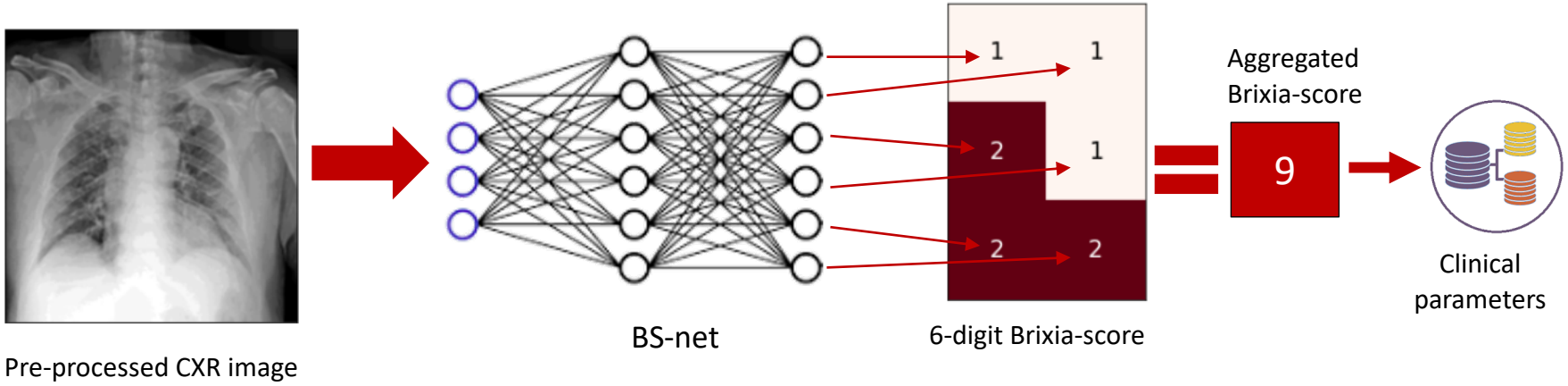


Segmentation maps of the lungs from the same CXR image, each of them with the fitted lines through LSM (left), RANSAC (centre) and ellipse fitting (right)

Computing the Brixia-score

- As said before, the Brixia-score is a 6-digit score, each digit ranging from 0 (no lung abnormalities) to 3 (interstitial and dominant alveolar infiltrates), that describe the lungs compromise of 6 different parts (three for each lung)

- Once pre-processed all the CXR images, we were able to compute the corresponding 6-digit Brixia-score for each of them and to use its sum as additional clinical parameter
  - Since the score measures the lung compromise, it can be extremely helpful to identify the right classification



Pre-processing clinical data

- From the very beginning we tried to **put ourselves in the same conditions of the test set**
  - We dropped the features in the training set that were completely unavailable in the test set

- We handled **missing values in categorical features through a constant filling strategy**
- We handled **missing values in numerical features through a KNN imputation[1]** (neighbourhood size = 10)

- In order to **both identify significant features for the explainability** of the AI system and to **reduce the dimensionality** of the dataset, we conducted **several model-based feature extractions[2] varying both the feature importance threshold and the maximum number of attributes to keep**

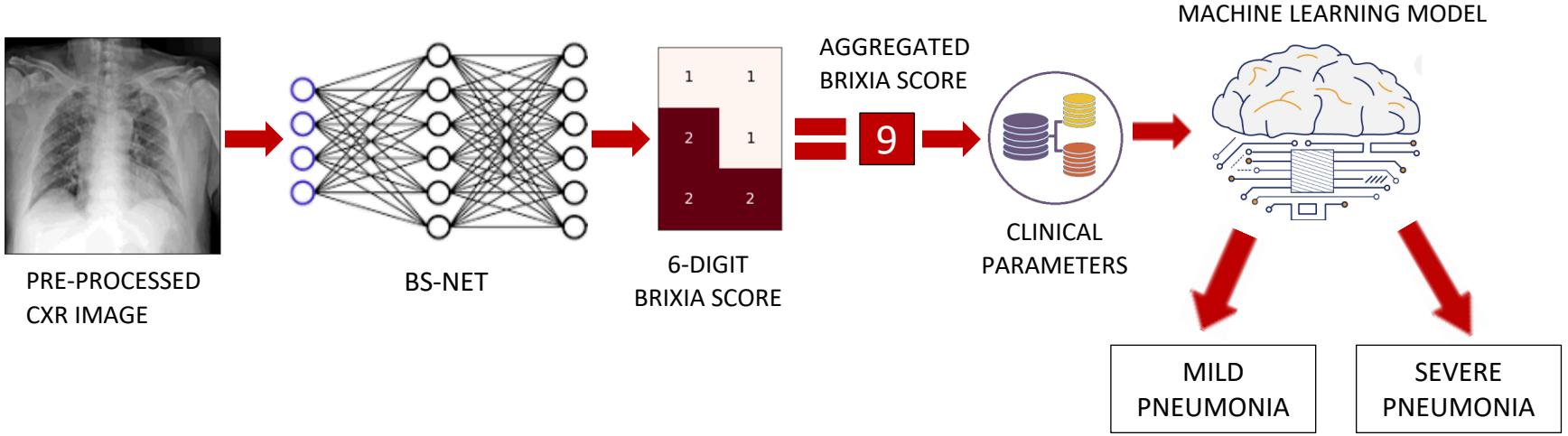
1 – KNN is a clustering algorithm that gathers points that are ''sufficiently close to each other'' according to some distance measure. However, **if a point has some unknown coordinates**, **KNN can impute** them **considering the corresponding known coordinates of his neighbours** (the necessary **neighbourhood computation is based on the known dimensions**).

2 – A **model-based feature extraction** is basically a **sensitivity analysis** in which the **features are dropped one at a time** and the **accuracy of a trained model is computed and compared to the accuracy obtained by the same model considering all the features.** If the **difference exceeds a given threshold**, then the **dropped feature is important** to the model in order to make the predictions

Building a pipeline of models to deal with chest radiographs and clinical parameters

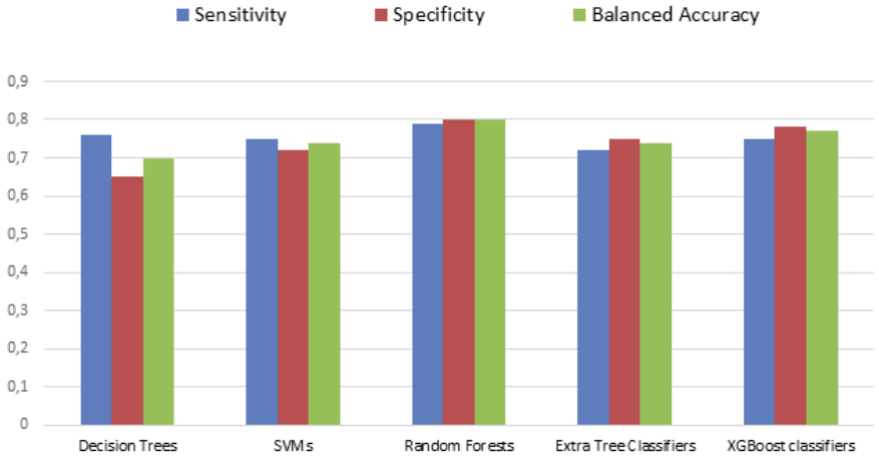
- We realized a pipeline like the one shown below and in which the machine learning model has been decided after some tests in order to find the best classifier for this task

- The model we considered were:
  - Decision Trees (params …)
  - Support Vector Machines (SVM) (params …)
  - Extra Trees (params …)
  - XGBoost (params …)
  - Random Forest (params …)



How to choose the best model

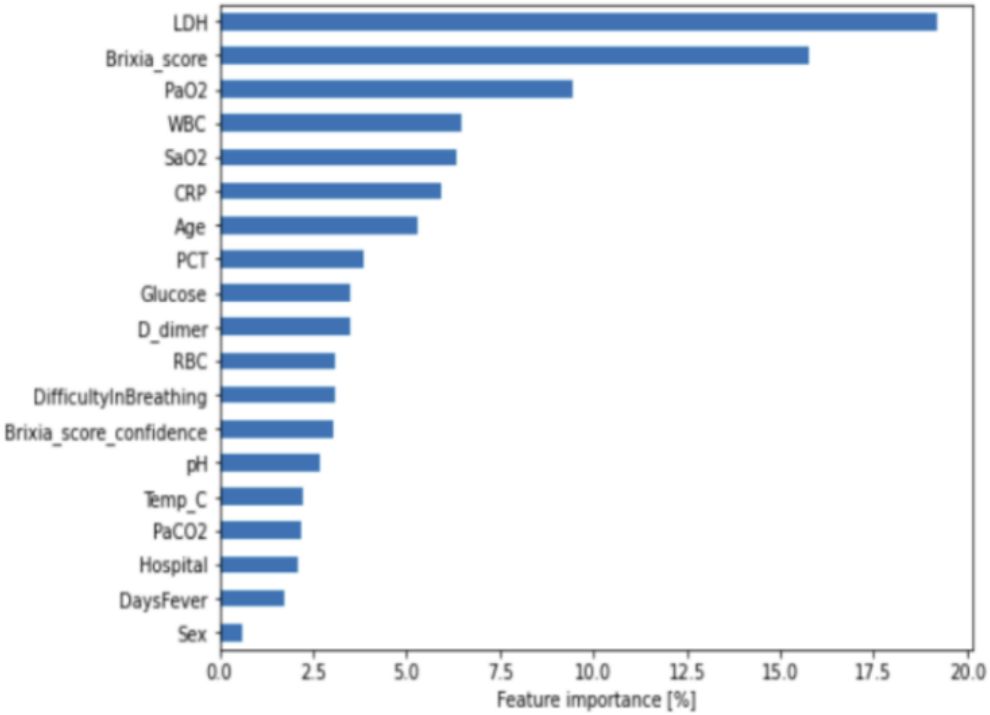
- First of all, we split our training set into two parts: **one for training (75%) and one for validation (25%)**
- **Each model was tuned through a grid search cross-validation to find the best hyper parameters** and then trained
- Hence, each model was evaluated over the validation set according to **sensitivity**, **specificity** and **balanced accuracy**. Their performance are shown in the chart below.



- As could be noted looking at the chart above, the **Random Forest proved to be the best classifier for this task**. **Before its employment over the test set** for the team evaluation, **it was retrained over the entire labelled dataset ensuring us a balanced accuracy of 74.4% (2° best result)**

Explaining the predictions – the feature importance to the model

The **first type of explanation** given by our AI system is the **importance that our model gives to the features**

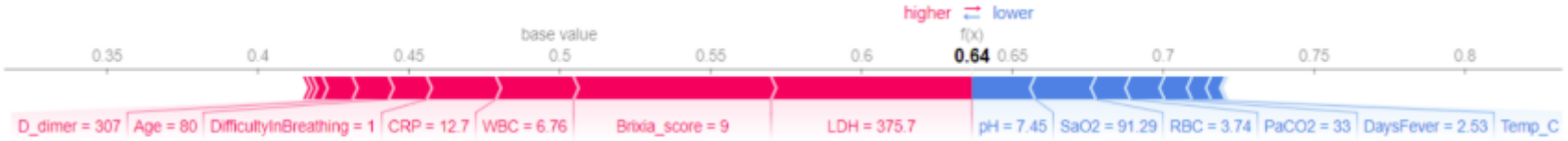


Feature importance to our Random Forest. The **Brixia-score, the LDH (cholesterol) and the arterial pressure are considered the most important parameters**. Conversely, the sex and the days of fever are so important to our classifier.

The feature importance technique, whose computation is quite similar to the model-based feature extraction shown earlier, **does not explain the inference mechanism of our Random Forest** (ensemble models are among the most difficult models to explain). However, **it acts as a justification: it is a set of reasons which support the (supposed) correct prediction of a model**

Explaining the predictions – SHAP values to show the features impact on a single prediction

The second type of explanation is given by **SHAP (Shapley Addictive exPlanations) values**, an explanation method based on cooperative game theory that **aims to justify a single prediction**

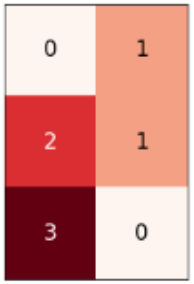


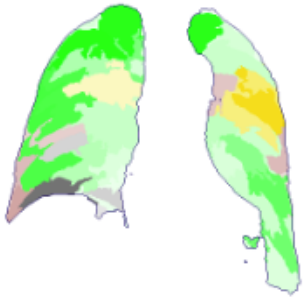
There exist several types of explanation plots or charts based on SHAP values. **The one above, which justifies a single prediction, shows which features push the model to predict a *severe pneumonia (red)* against those which drag the model to predict a *mild pneumonia (blue).***

In this particular case, **the features refer to a patient with a *severe pneumonia*** and since the prediction *f(x) > 0.5*, **the predicted class is correct**. The **red features are those which pushed the model to predict a *severe pneumonia*** (**Brixia score** = 9, age = 80, **LDH** = 375.7 etc.), while **the blue ones are those which dragged the model to a *mild pneumonia***

Explaining the predictions – Explainability maps of the lungs

- The last type of explanation is **purely visua**l and is an **explainability map of the lungs produced by BS-net.**
- An **explainability map is an image showing the lungs subdivided into super-pixels** (i.e. regions that share a similar intensity and pattern), **each with a different colour and, given a colour, a different intensity**
  - The **6 regions into which the lungs are subdivided can contain more than one super-pixel and a super-pixel can be shared between different regions**
- **The darker the colour, the greater the impact** of a super-pixel **on the score** associated to the region containing that super-pixel (or part of it). For the same colour, **the greater the intensity, the greater the impact on the associated score**



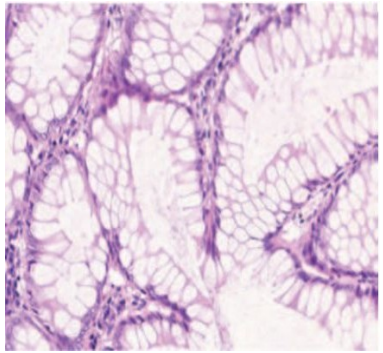


Let's see an example. Each lung is subdivided into 3 regions (upper, middle and lower) and each one has an associated severity from 0 to 3 (the 6-digit Brixia score). The grey lower regions of the left lung have a stronger impact (they are darker) on the score of class 3 (whose colour is black) than the adjacent parts (they are more tenuous). Similarly, the green upper part of the same lung has a heavier impact (it is brighter) on the score of class 0 (whose colour is green) than the adjacent yellow region (it is paler).

ADVANCED VISION MODELS: SEGMENTATION AND DETECTION

Richer visual recognition tasks: segmentation and detection

Classification

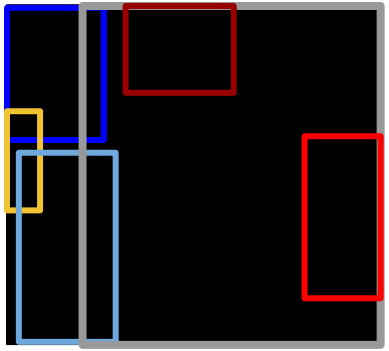


Output:
one category label for image (e.g., colorectal glands)

Semantic Segmentation

Output:
category label for each pixel in the image

Detection



Output:
Spatial bounding box for each **instance** of a category object in the image

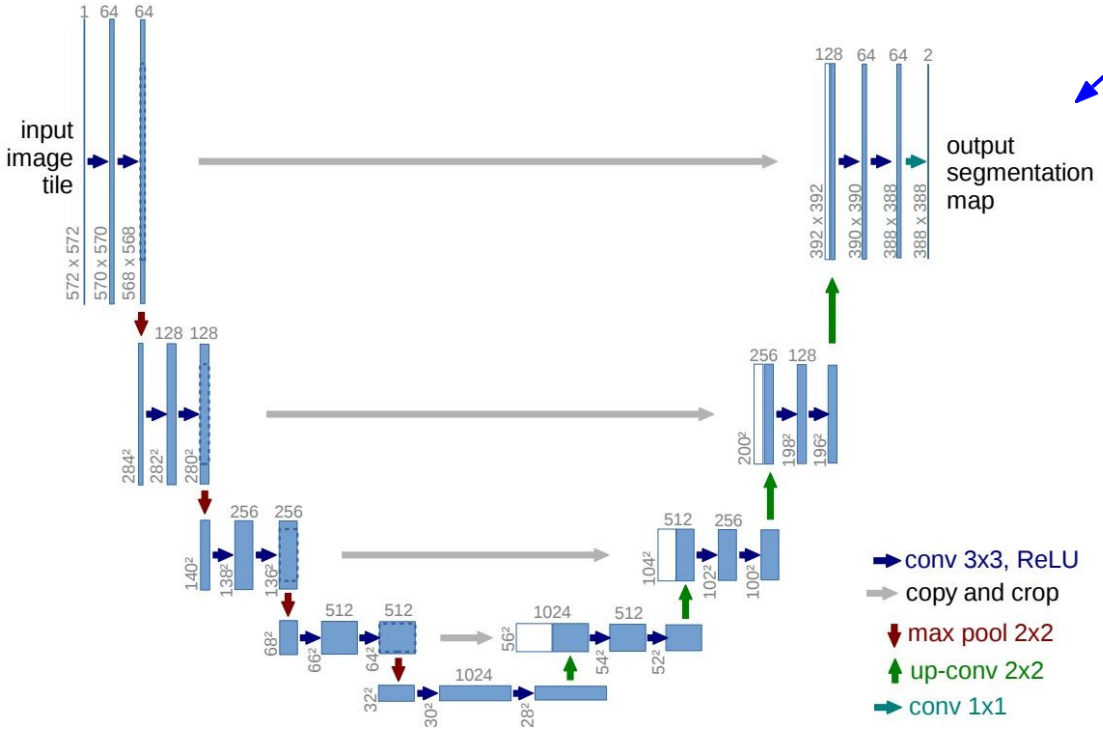
Instance Segmentation

Output:
Category label and instance label for each pixel in the image

Distinguishes between different instances of an object

Figures: Chen et al. 2016. https://arxiv.org/pdf/1604.02677.pdf

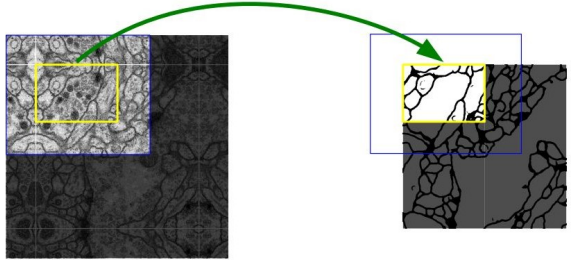
Semantic segmentation: U-Net



Ronneberger et al. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

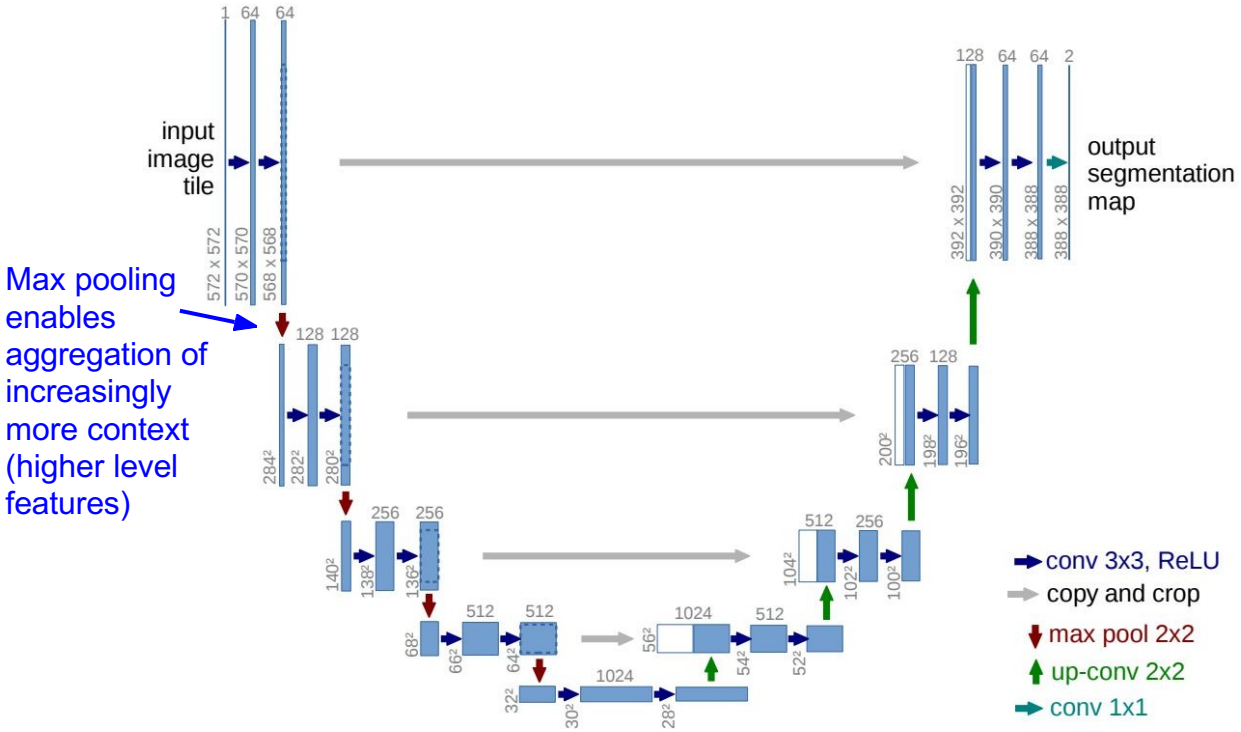
Output is an image mask: width x height x # classes

Output image size a little smaller than original, due to convolutional operations w/o padding



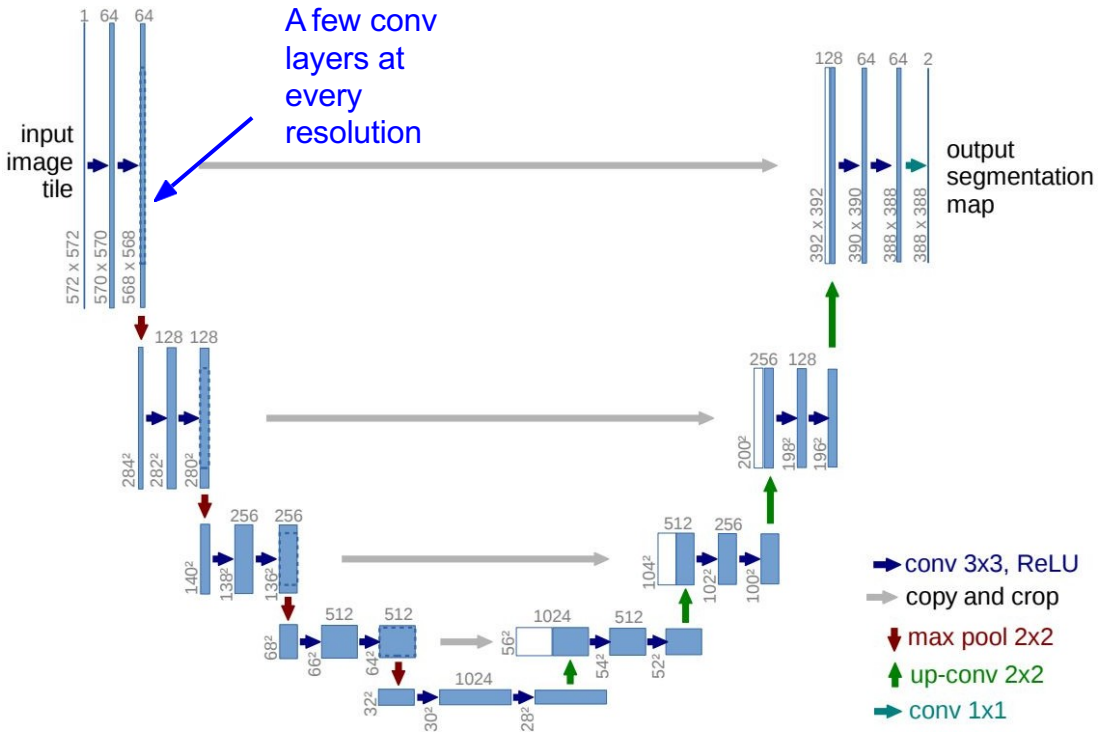
Gives more “true” context for reasoning over each image area. Can tile to make predictions for arbitrarily large images

Semantic segmentation: U-Net



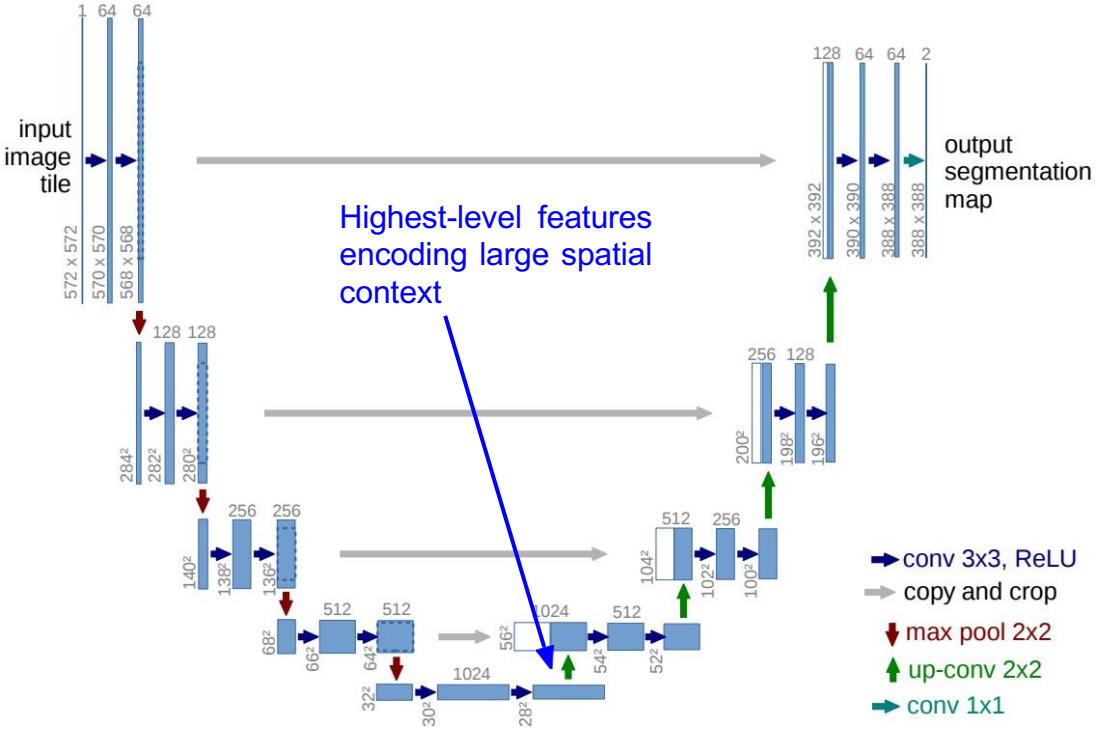
Ronneberger et al. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

Semantic segmentation: U-Net



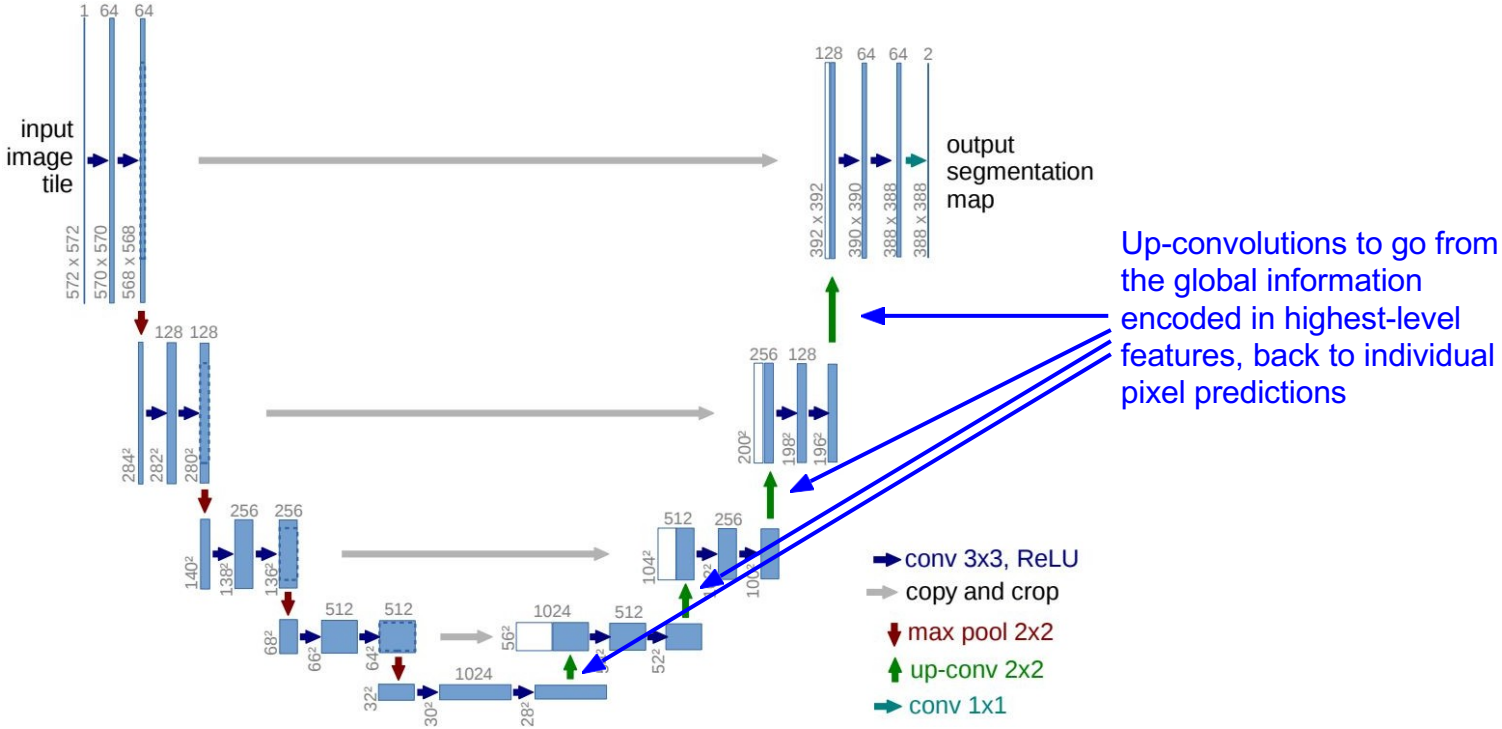
Ronneberger et al. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

Semantic segmentation: U-Net



Ronneberger et al. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

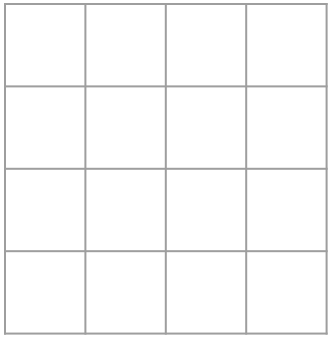
Semantic segmentation: U-Net



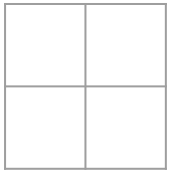
Ronneberger et al. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

Up-convolutions

Recall: Normal 3 x 3 convolution, stride 2 pad 1



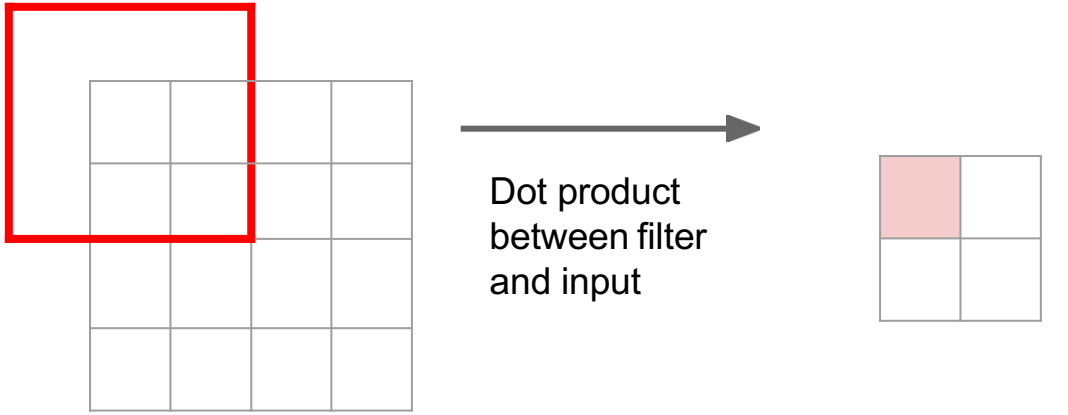
Input: 4 x 4



Output: 2 x 2

Up-convolutions

Recall: Normal 3 x 3 convolution, stride 2 pad 1

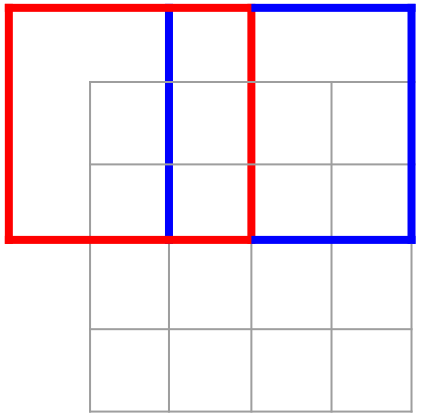


Input: 4 x 4

Output: 2 x 2

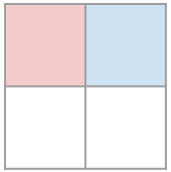
Up-convolutions

Recall: Normal 3 x 3 convolution, stride 2 pad 1



Input: 4 x 4

Dot product between filter and input



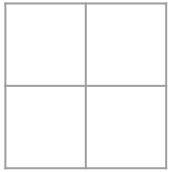
Output: 2 x 2

Filter moves 2 pixels in the input for every one pixel in the output

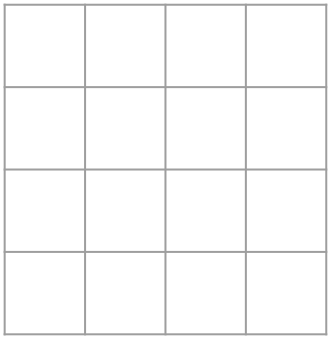
Stride gives ratio between movement in input and output

Up-convolutions

3 x 3 **transpose** convolution, stride 2 pad 1



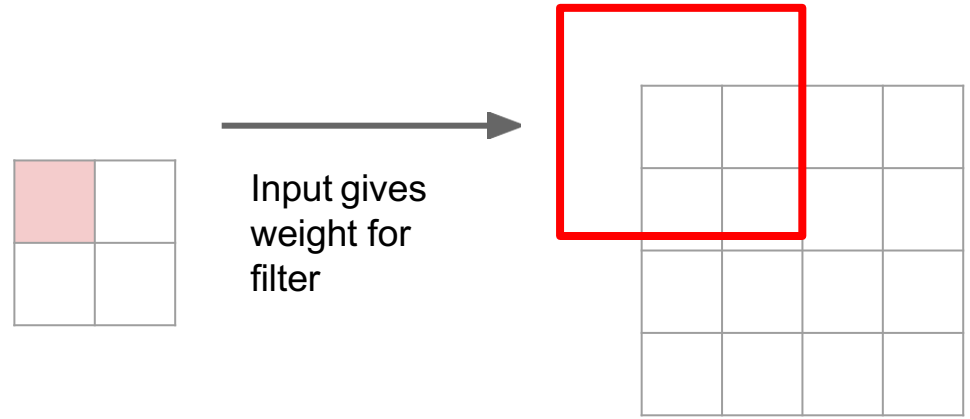
Input: 2 x 2



Output: 4 x 4

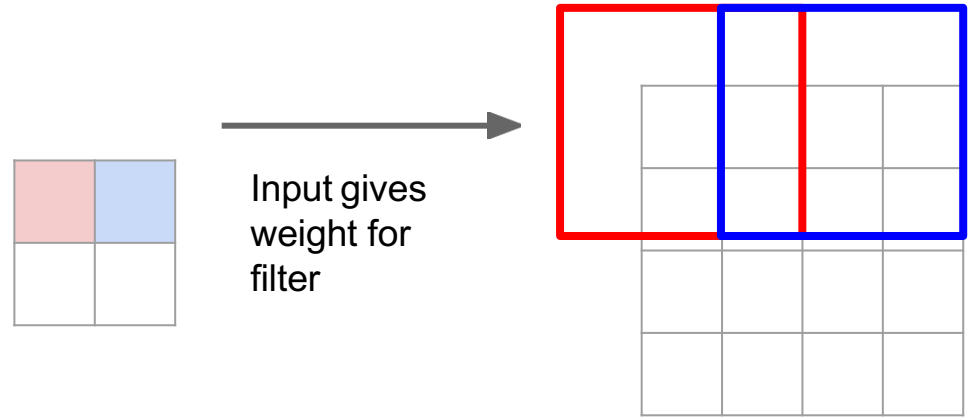
Up-convolutions

3 x 3 **up-convolution**, stride 2 pad 1



Up-convolutions

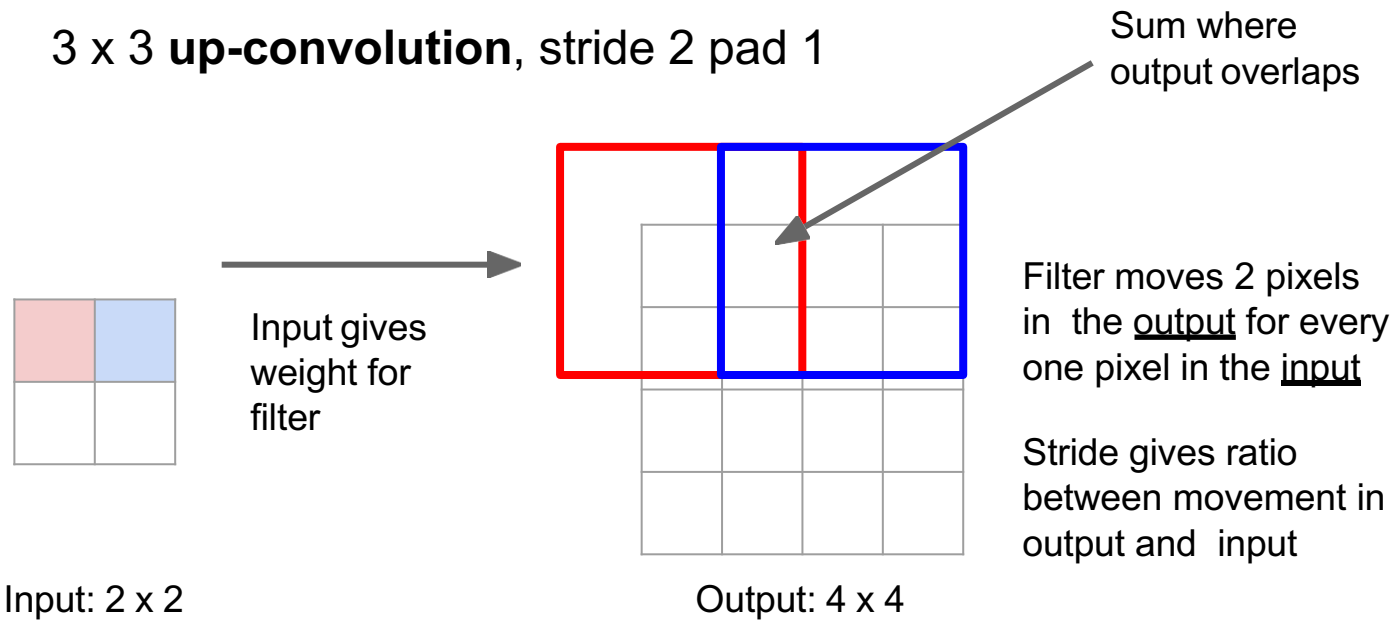
3 x 3 **up-convolution**, stride 2 pad 1



Filter moves 2 pixels in the output for every one pixel in the input

Stride gives ratio between movement in output and input

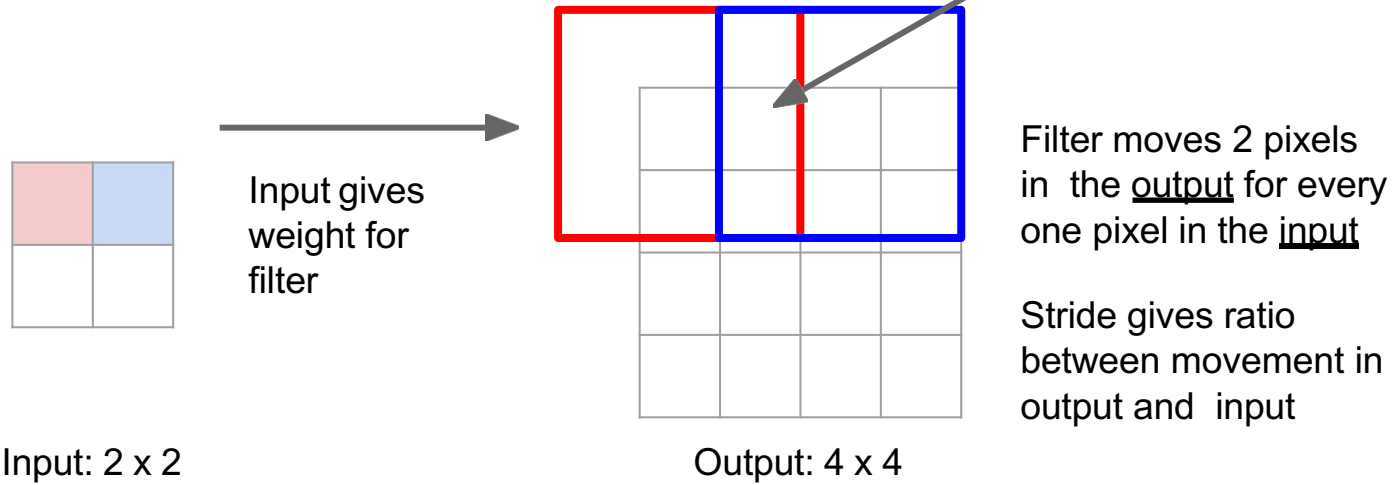
Up-convolutions



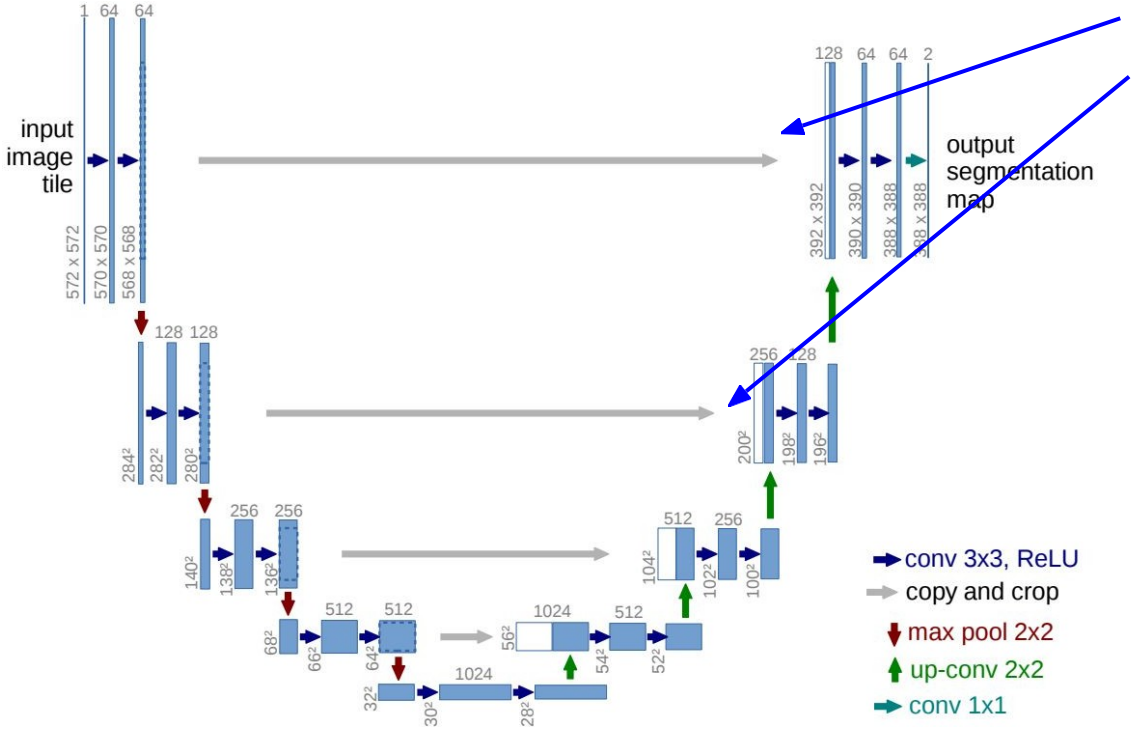
Up-convolutions

Other names:
-Transpose convolution
-Fractionally strided convolution
-Backward strided convolution

3 x 3 **up-convolution**, stride 2 pad 1



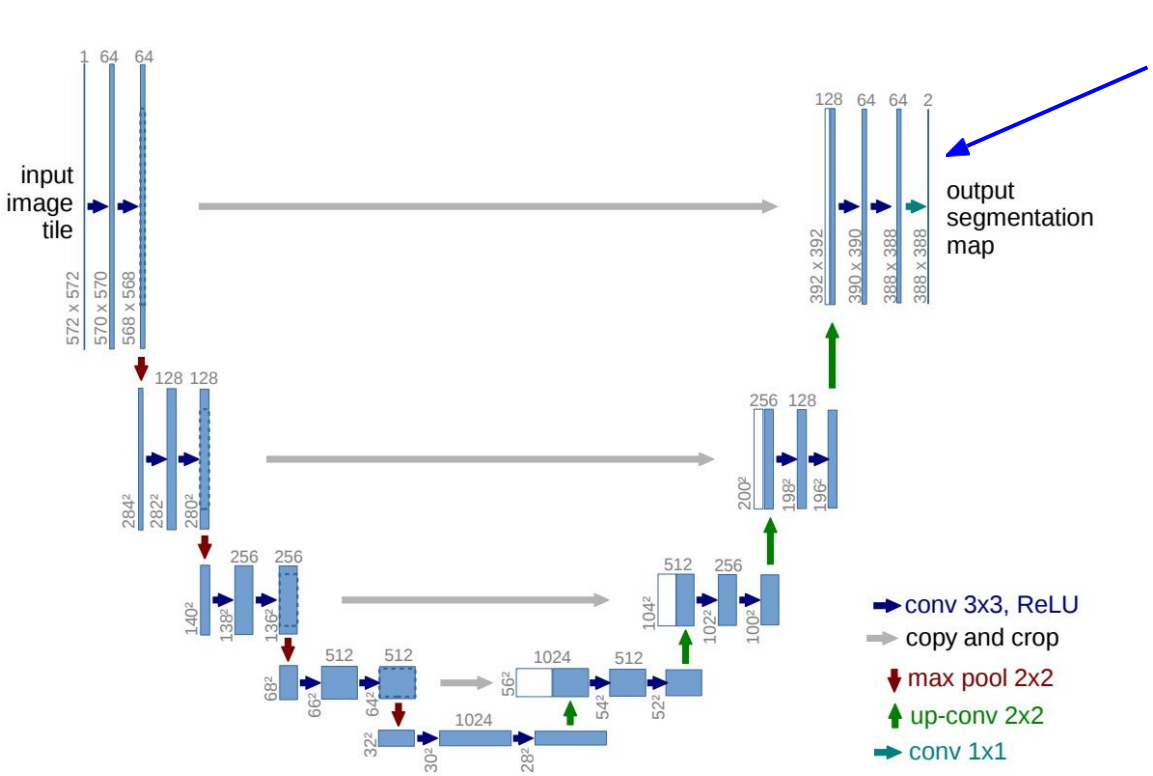
Semantic segmentation: U-Net



Concatenate with same-resolution feature map during downsampling process to combine high-level information with low-level (local) information

Ronneberger et al. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

Semantic segmentation: U-Net



Ronneberger et al. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

Semantic segmentation: IOU evaluation

Intersection over Union:

$$IoU = \frac{target \cap prediction}{target \cup prediction}$$

pixels included in both target and prediction maps

Total # pixels in the union of both masks

Can compute this over all masks in the evaluation set, or at individual mask and image levels to get finer-grained understanding of performance.

Also known as Jaccard Index

Semantic segmentation: Pixel Accuracy evaluation

$$\text{Pixel Accuracy (PA)} = \frac{\#\text{ correctly classified pixels}}{\#\text{ total pixels}}$$

TP + TN

Total pixels in image

Q: What is a potential problem with this?

A: Think about what happens when there is class imbalance.

Semantic segmentation: Dice coefficient evaluation

2 * intersection

$$\text{Dice Coefficient} = \frac{2 * (\text{target} \cap \text{prediction})}{\#\ \text{target mask pixels} + \#\ \text{prediction mask pixels}}$$

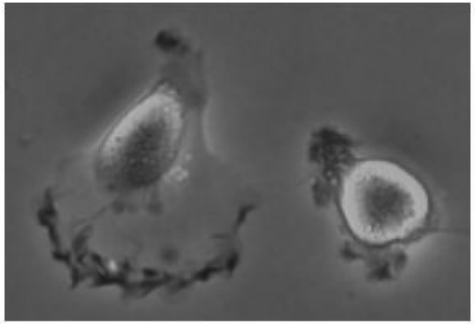
Sum of target mask size
+ prediction mask size

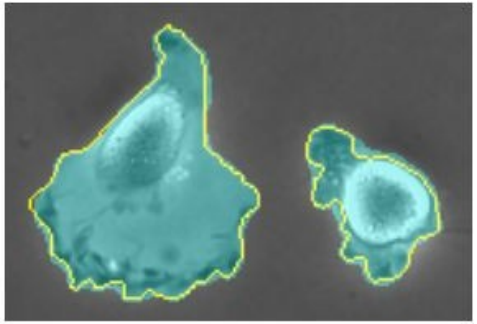
Very similar to IOU / Jaccard, can derive one from the other

Semantic segmentation: summary of evaluation metrics

- Most commonly use IOU / Jaccard or Dice Coefficient
- Sometimes will also see pixel accuracy
- If multi-class segmentation task, typically report all these metrics per-class, and then a mean over all classes

Semantic segmentation: U-Net cell segmentation





Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

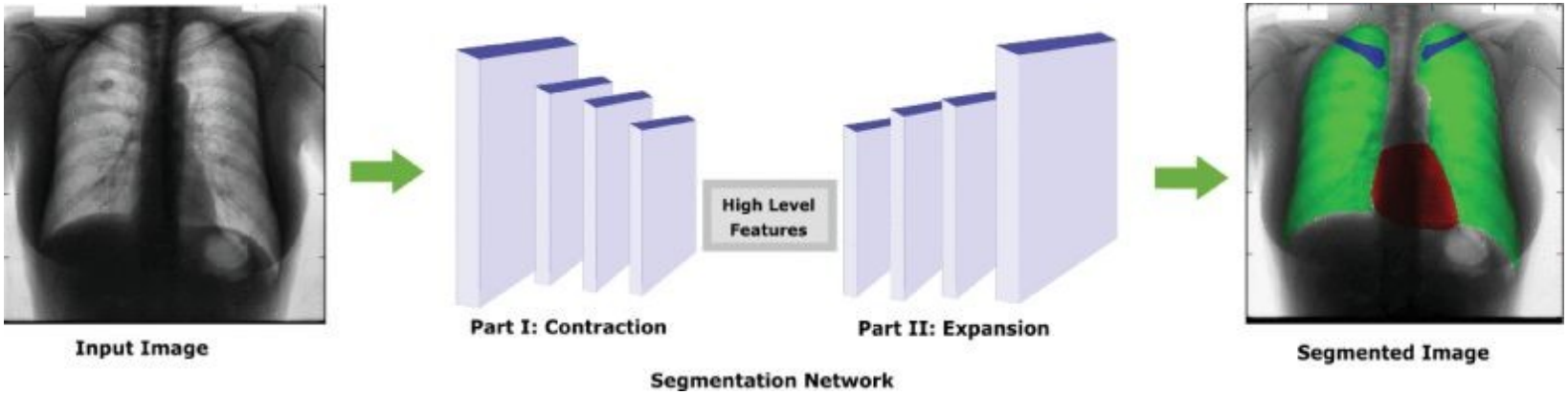
Very small dataset: 30 training images of size 512x512, in the ISBI 2012 Electron Microscopy (EM) segmentation challenge. Used excessive data augmentation to compensate.

Ronneberger et al. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

Novikov et al. 2018

Q: What loss function would be appropriate here?

- Chest x-ray segmentation of lungs, clavicles, and heart
- JSRT dataset of 247 chest-xrays at 2048x2048 resolution. (But downsampled to 128x128 and 256x256!)
- Used a U-Net based segmentation network with a few modifications



Novikov et al. Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs. IEEE Trans. on Medical Imaging, 2018.

Novikov et al. 2018

Image ground truth class mask

$$L_{\text{dice}}(y, \hat{y}) = 1 - \frac{2\sum_{i,j} y_{i,j}\hat{y}_{i,j}}{\sum_{i,j} y_{i,j} + \sum_{i,j} \hat{y}_{i,j}}$$

Image pixel class probabilities

- Multi-class segmentation -> tried both a per-pixel softmax loss as well as a loss based on the Dice coefficient.

Note: this Dice loss is often useful to try!

- Class imbalance -> weight loss terms corresponding to each ground-truth class by inverse of class frequency: (# class pixels) / (total # pixels in data)

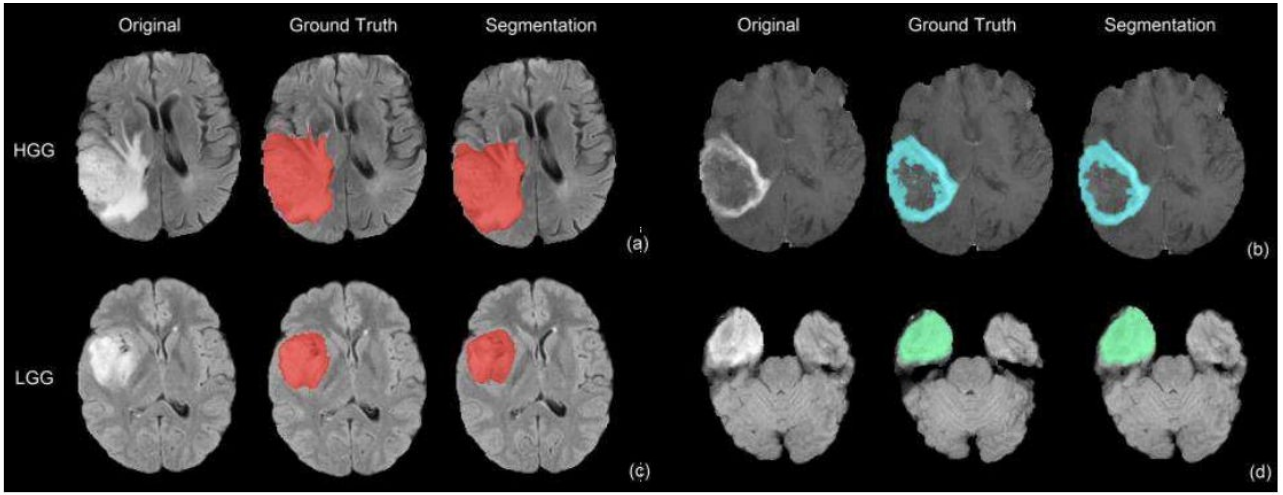
Body Part	Lungs		Clavicles		Heart	
Evaluation Metric	D	J	D	J	D	J
InvertedNet	0.972	0.946	**0.902**	**0.821**	0.935	0.879
All-Dropout	**0.973**	**0.948**	0.896	0.812	**0.941**	**0.888**
All-Convolutional	0.971	0.944	0.876	0.780	0.938	0.883
Original U-Net	0.971	0.944	0.880	0.785	0.938	0.883

Dice and Jaccard evaluation

Novikov et al. Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs. IEEE Trans. on Medical Imaging, 2018.

Dong et al. 2017

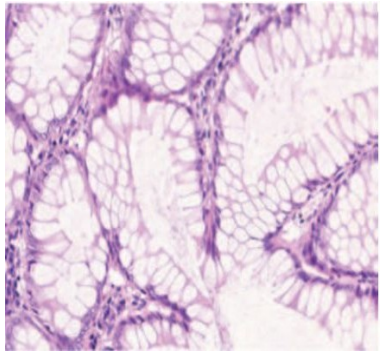
- Segmentation of tumors in brain MR image slices
- BRATS 2015 dataset: 220 high-grade brain tumor and 54 low-grade brain tumor MRIs
- U-Net architecture, Dice loss function



Dong et al. Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks. MIUA, 2017.

Richer visual recognition tasks: segmentation and detection

Classification

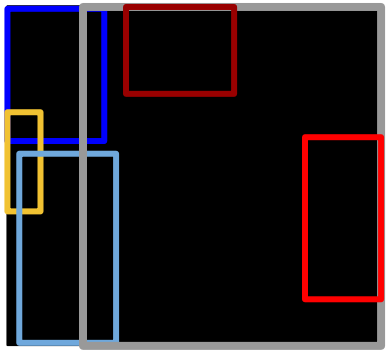


Output:
one category label for image (e.g., colorectal glands)

Semantic Segmentation

Output:
category label for each pixel in the image

Detection



Output:
Spatial bounding box for each **instance** of a category object in the image

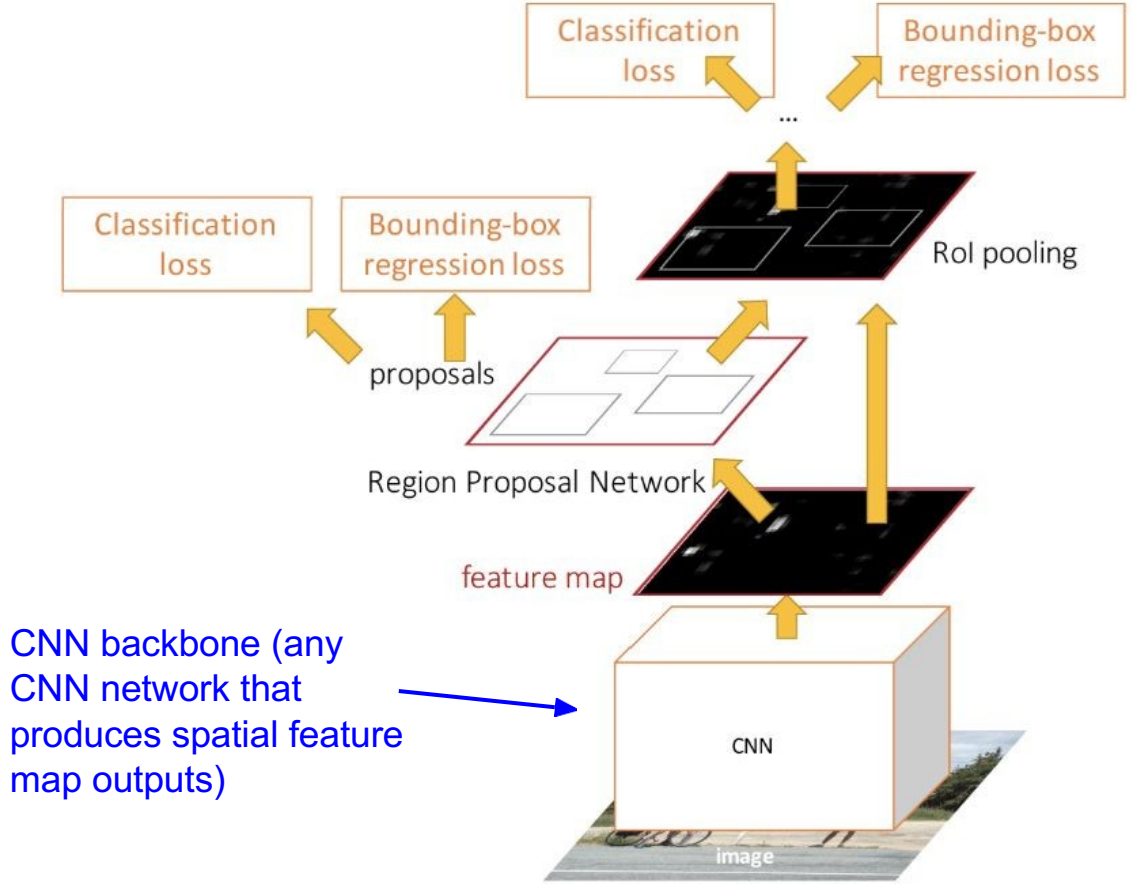
Instance Segmentation

Output:
Category label and instance label for each pixel in the image

Distinguishes between different instances of an object

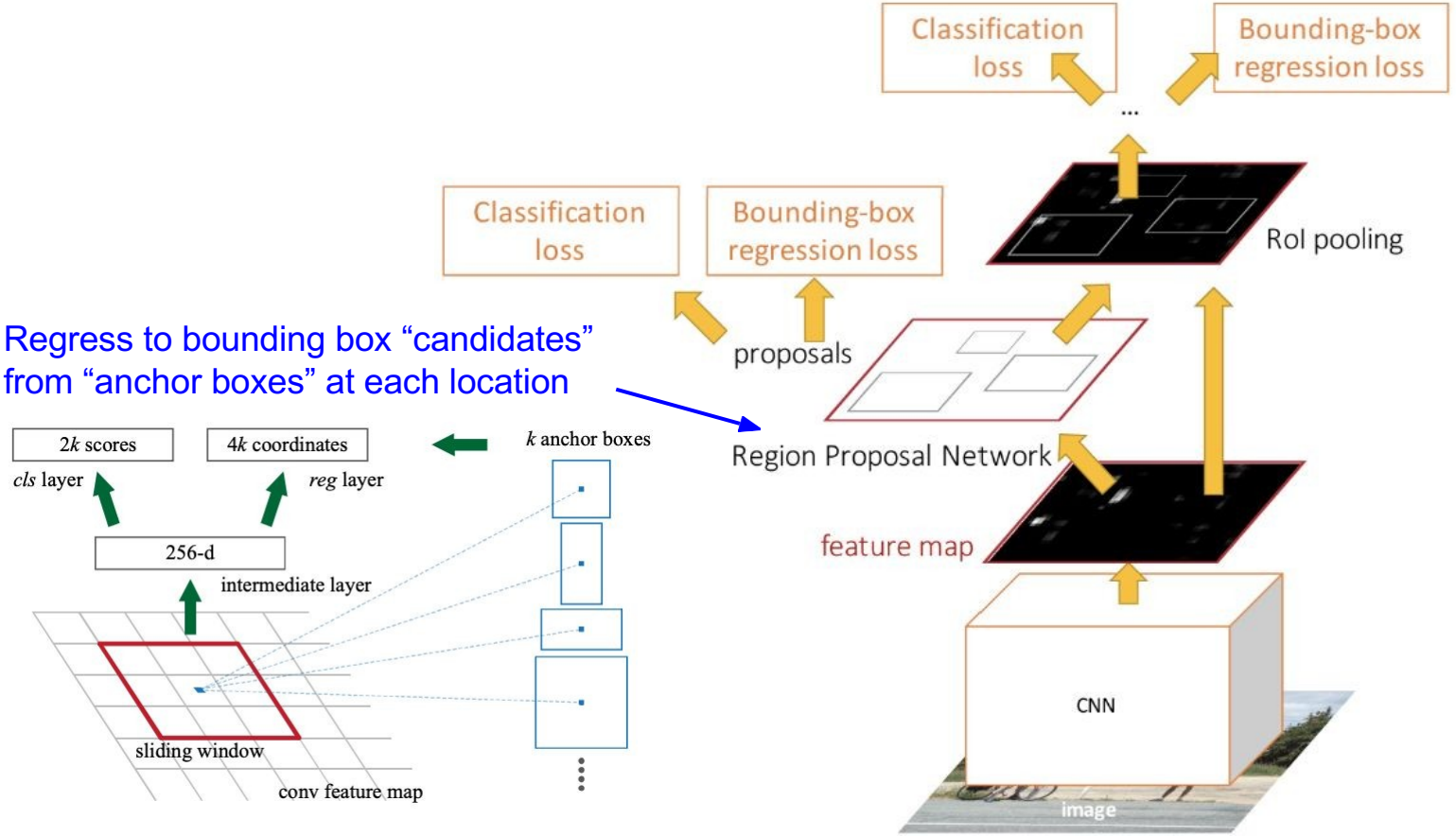
Figures: Chen et al. 2016. https://arxiv.org/pdf/1604.02677.pdf

Object detection: Faster R-CNN

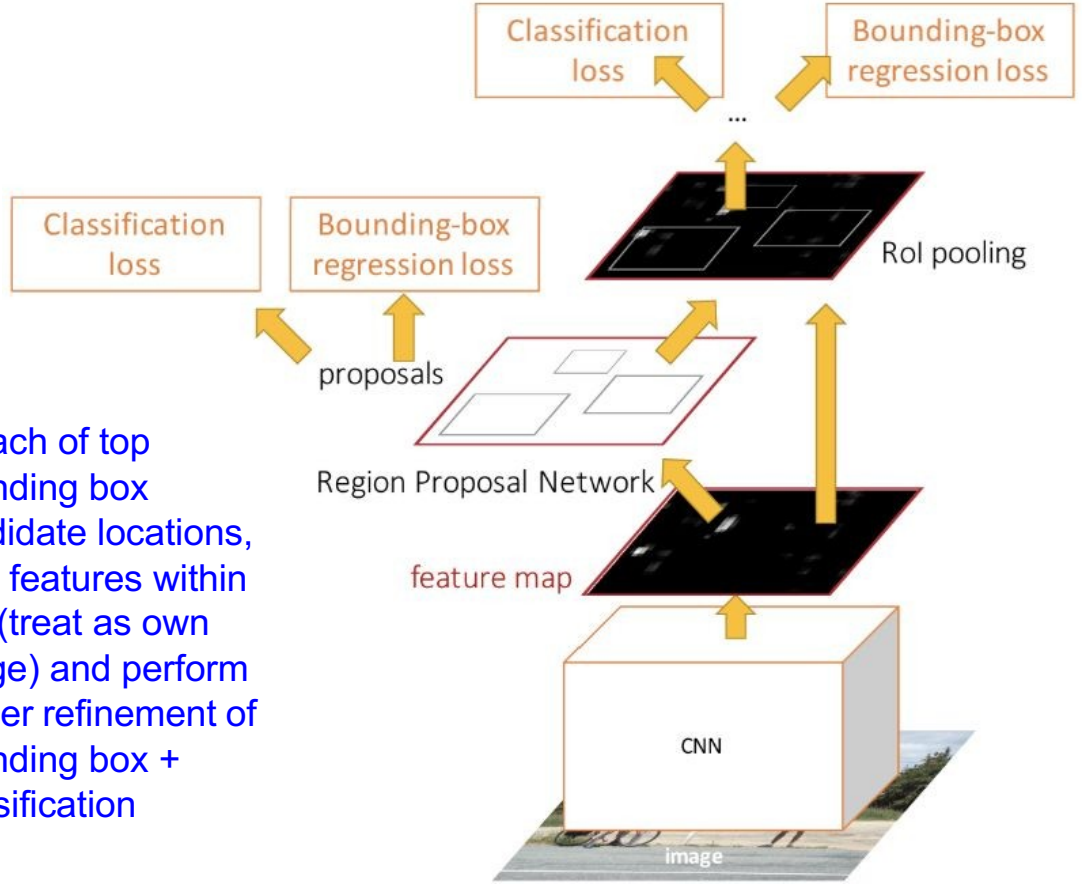


CNN backbone (any CNN network that produces spatial feature map outputs)

Object detection: Faster R-CNN



Object detection: Faster R-CNN



In each of top bounding box candidate locations, crop features within box (treat as own image) and perform further refinement of bounding box + classification

Evaluation of object detection

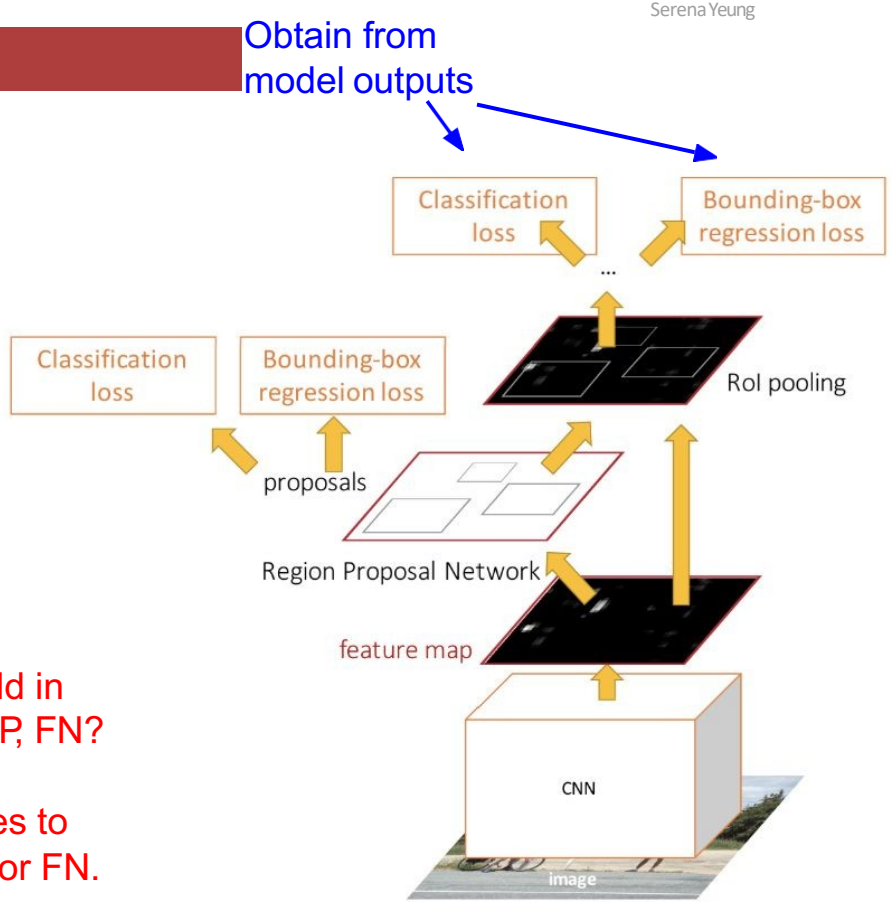
Standard output of object detection

For each class, a set of bounding box predictions with associated confidences:

- E.g., (x, y, h, w, c)

Bounding box

Class confidence

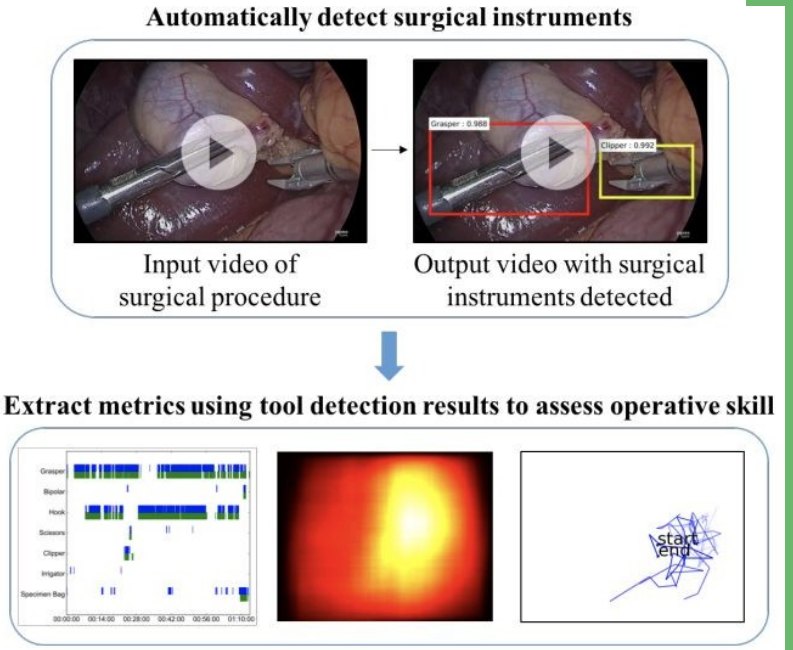


We have the class confidences to vary the threshold in plotting the PR curve. But how do we get TP, TN, FP, FN?

A: Choose an IOU threshold with ground truth boxes to determine if bounding box prediction is TP, TN, FP, or FN. Then can plot PR curve and obtain AP metric.

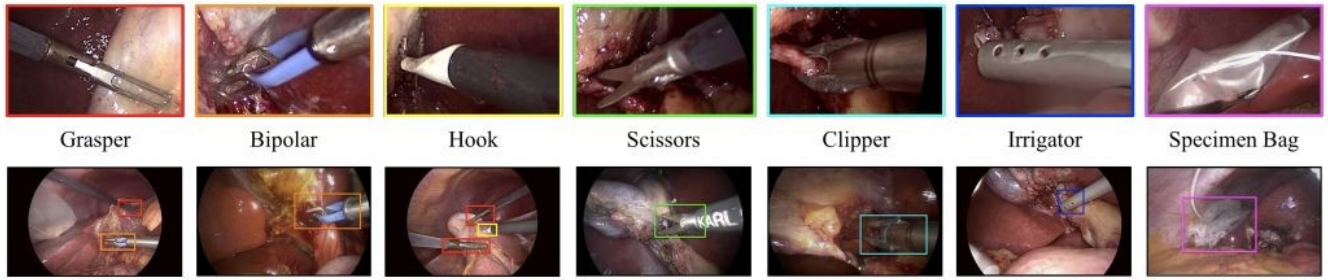
Jin et al. 2018

- Detection of surgical instruments in surgery videos (in each video frame)
- Surgical instrument movement over the course of a video can be used to extract metrics such as tool switching, and spatial trajectories, that can be used to assess and provide feedback on operative skill.
- Used M2cai16-tool dataset of 15 surgical videos. Annotated 2532 frames with bounding boxes of 7 tools.



Jin et al. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. WACV, 2018.

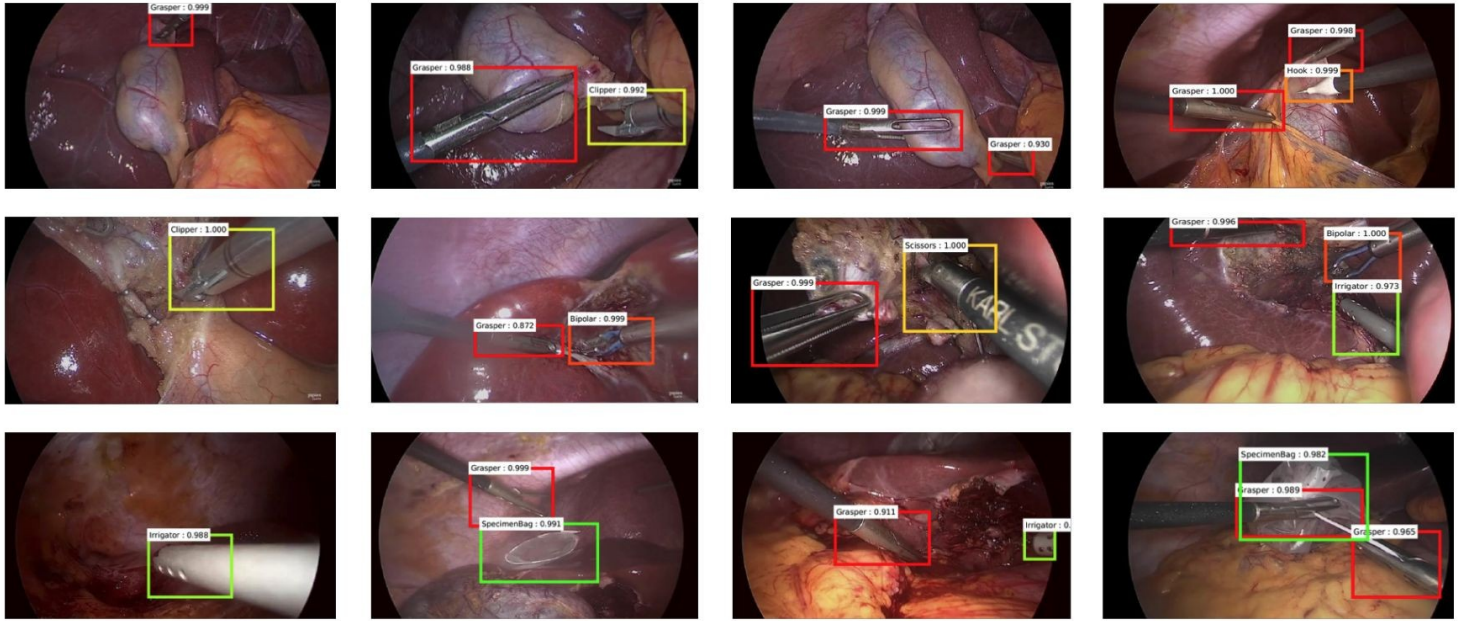
Jin et al. 2018



Tool	AP
Grasper	48.3
Bipolar	67.0
Hook	78.4
Scissors	67.7
Clipper	86.3
Irrigator	17.5
Specimen Bag	76.3
mAP	63.1

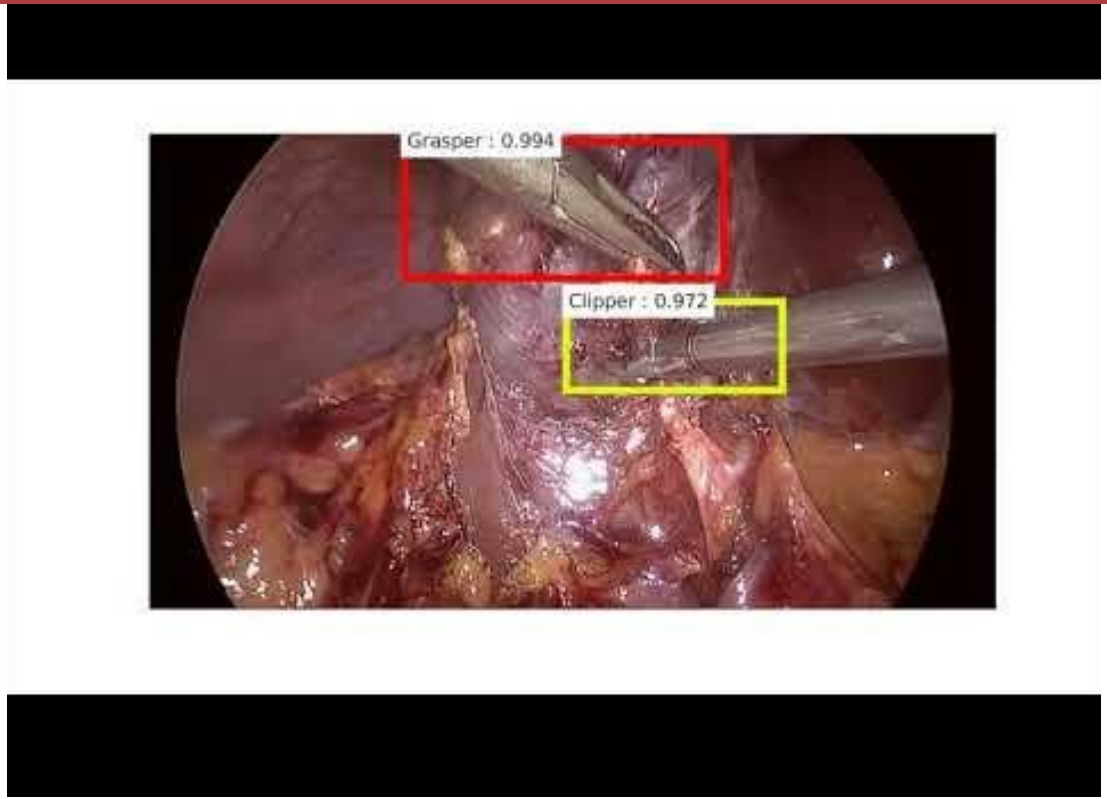
Jin et al. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. WACV, 2018.

Jin et al. 2018



Jin et al. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. WACV, 2018.

Jin et al. 2018



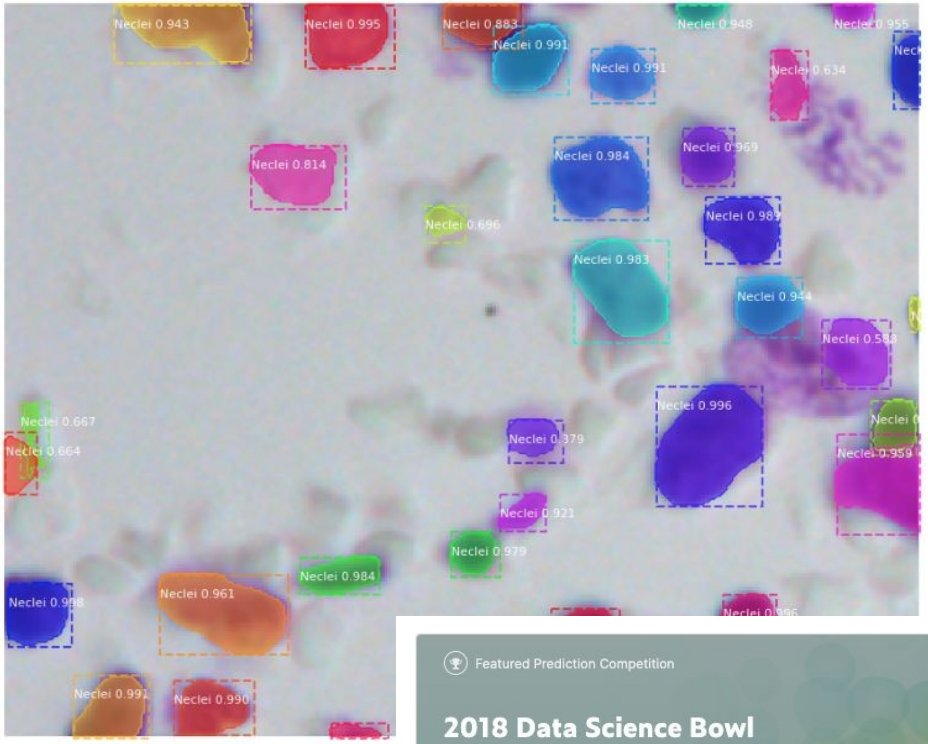
Jin et al. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. WACV, 2018.

Other object detection architectures

- RCNN, Fast RCNN: older and slower predecessors to Faster-RCNN
- YOLO, SSD: single-stage detectors that change region proposal generation -> region classification two-stage pipeline into a single stage.
  - Faster, but lower performance. Struggles more with class imbalance relative to two-stage networks that filter only top object candidate boxes for the second stage.
- RetinaNet: single-stage detector that uses a “focal loss” to adaptively weight harder examples over easy background examples. Able to outperform Faster R-CNN on some benchmark tasks, while being more efficient

RetinaNet also worth trying for object detection projects!

Example: instance segmentation of cell nuclei

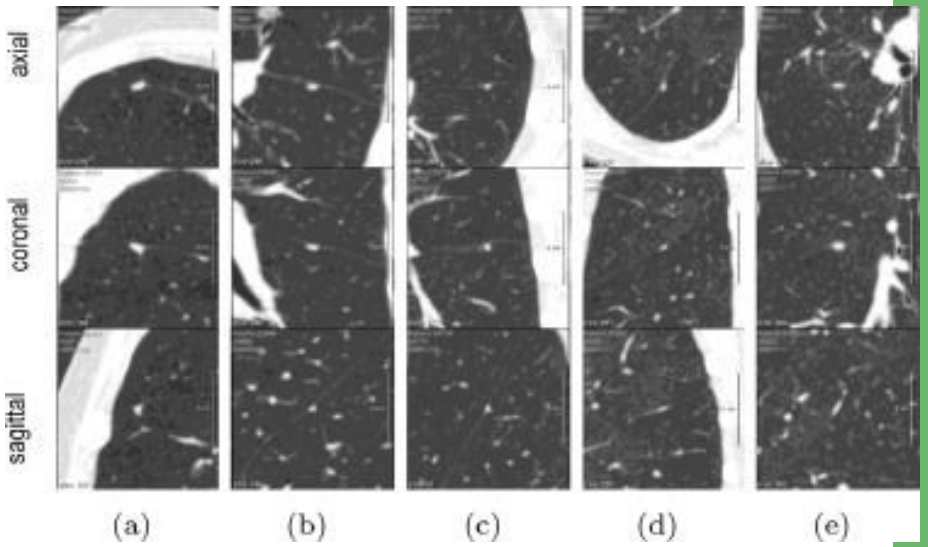


ADVANCED VISION MODELS FOR HIGHER-DIMENSIONAL (3D AND VIDEO) DATA

How do we handle 3D data?

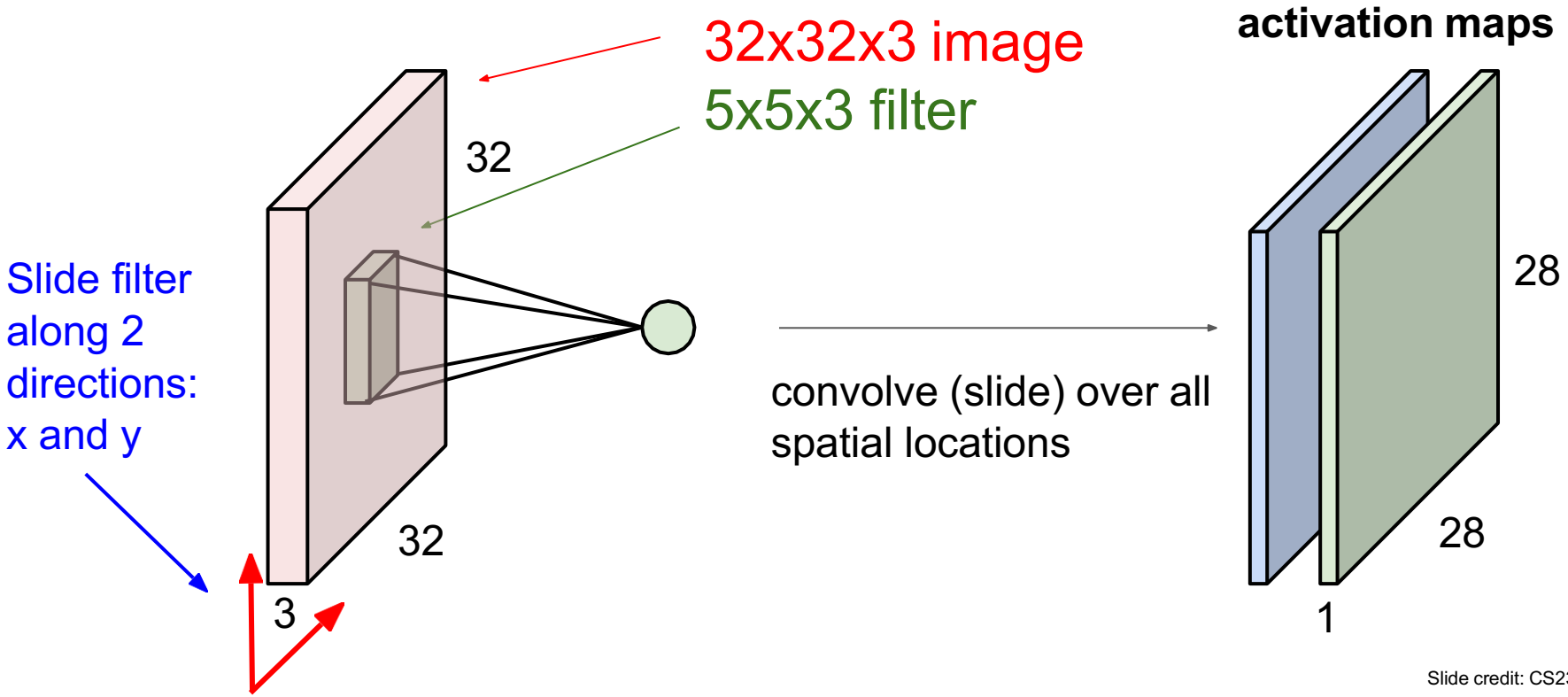
Recall: Ciompi et al. 2015

- Task: classification of lung nodules in 3D CT scans as peri-fissural nodules (PFN, likely to be benign) or not
- Dataset: 568 nodules from 1729 scans at a single institution. (65 typical PFNs, 19 atypical PFNs, 484 non-PFNs).
- Data pre-processing: prescaling from CT hounsfield units (HU) into [0,255]. Replicate 3x across R,G,B channels to match input dimensions of ImageNet-trained CNNs.



Ciompi et al. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Medical Image Analysis, 2015.

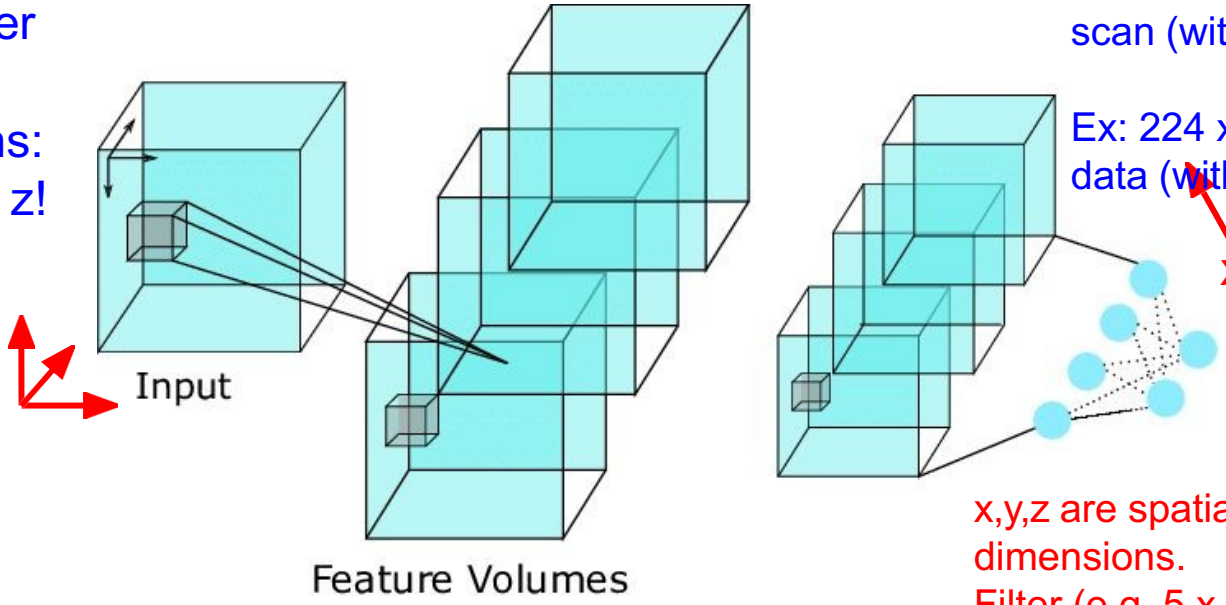
Remember 2D convolutions



Slide credit: CS231n

3D convolutions

Slide filter along **3** directions: x, y, and z!



When might you use 3D convolutions?

Ex: 224 x 224 x 1 x 256 3D CT scan (with 256 slices)

Ex: 224 x 224 x 3 x 500 video data (with 500 temporal frames)

x y z

channels (e.g. R,G,B)

x,y,z are spatial and/or temporal dimensions.
Filter (e.g. 5 x 5 x 3 x 10 filter) goes all the way through the "channels" dimension as before.

For video data, 3rd dimension is time

Figure credit:
https://www.researchgate.net/profile/Deepak_Mishra19/publication/330912338/figure/fig1/AS:723363244810254@1549474645742/Basic-3D-CNN-architecture-the-3D-filter-is-convolved-with-the-video-in-three-dimensions.png

Now: 3D CNNs for lung nodule classification

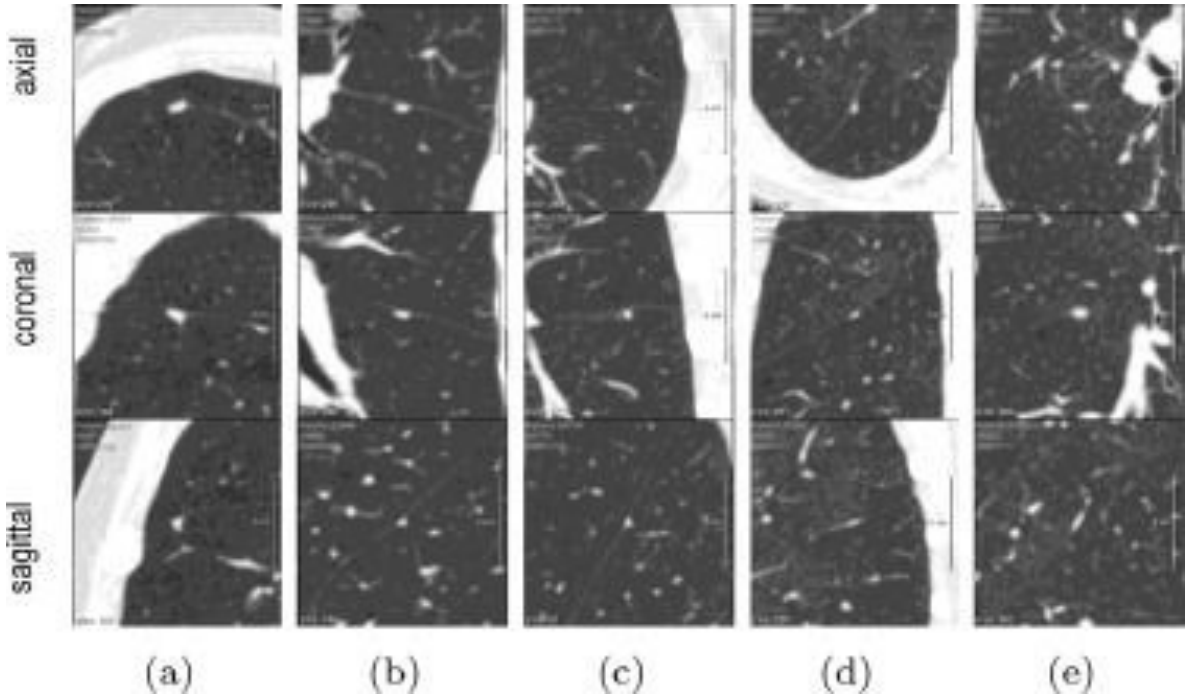
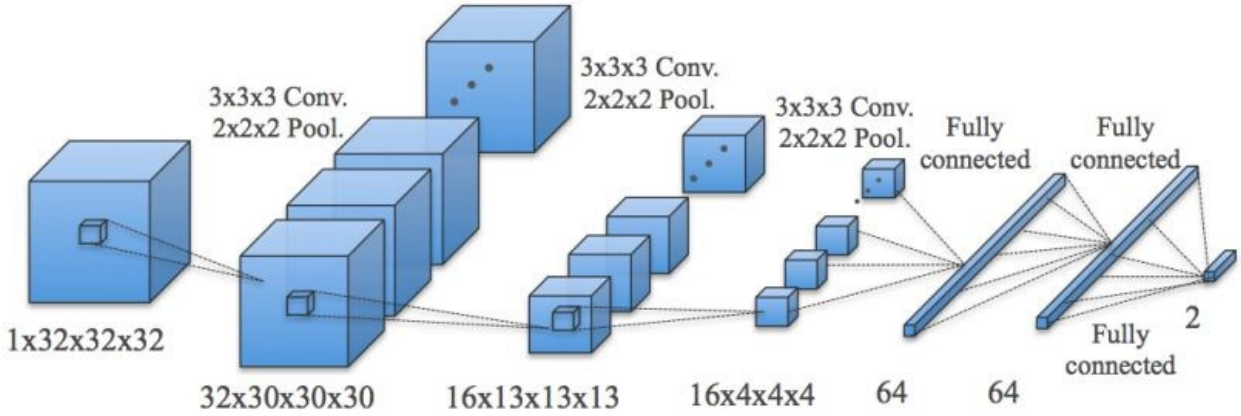


Figure credit: Ciompi et al. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Medical Image Analysis, 2015.

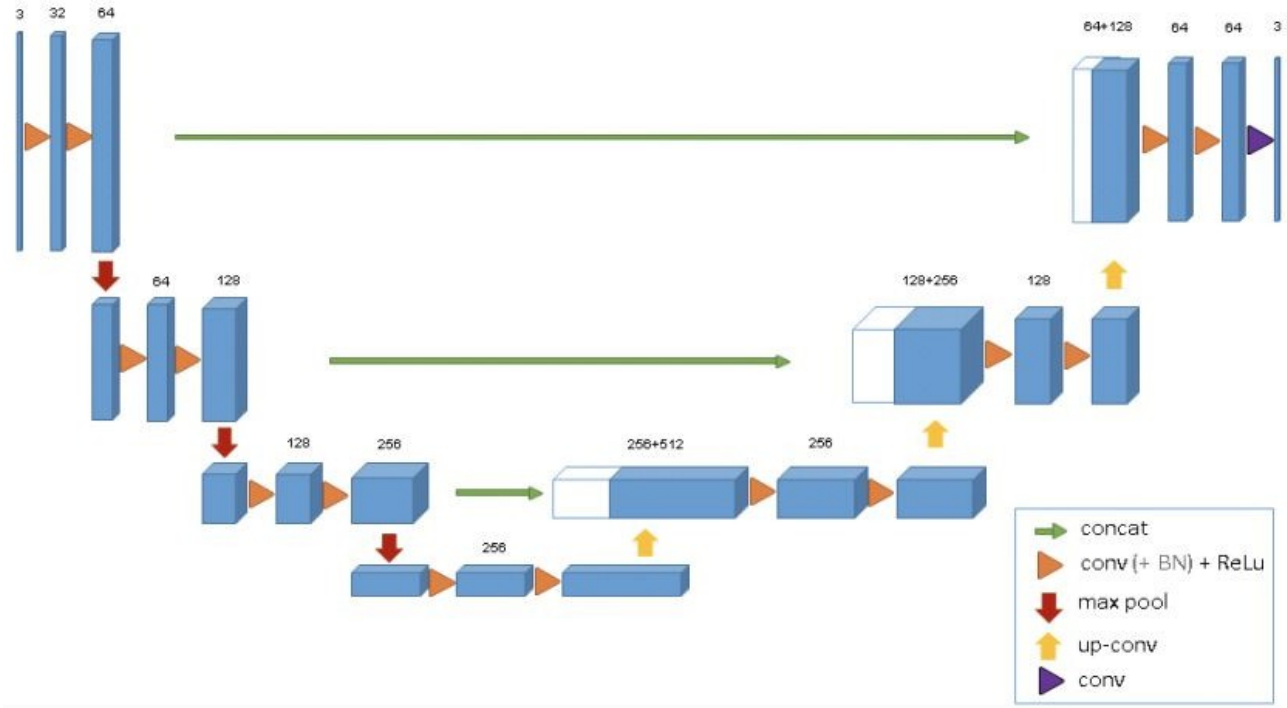
Huang et al. 2017

- Simple 3D CNN for lung nodule classification
- Used image processing approaches to extract candidate nodules, then 3D CNN to classify the surrounding volume
- Used the Lung Image Database Consortium (LIDC) Dataset, with 99 3D CT scans



Huang et al. Lung Nodule Detection in CT Using 3D Convolutional Neural Networks. ISBI 2017.

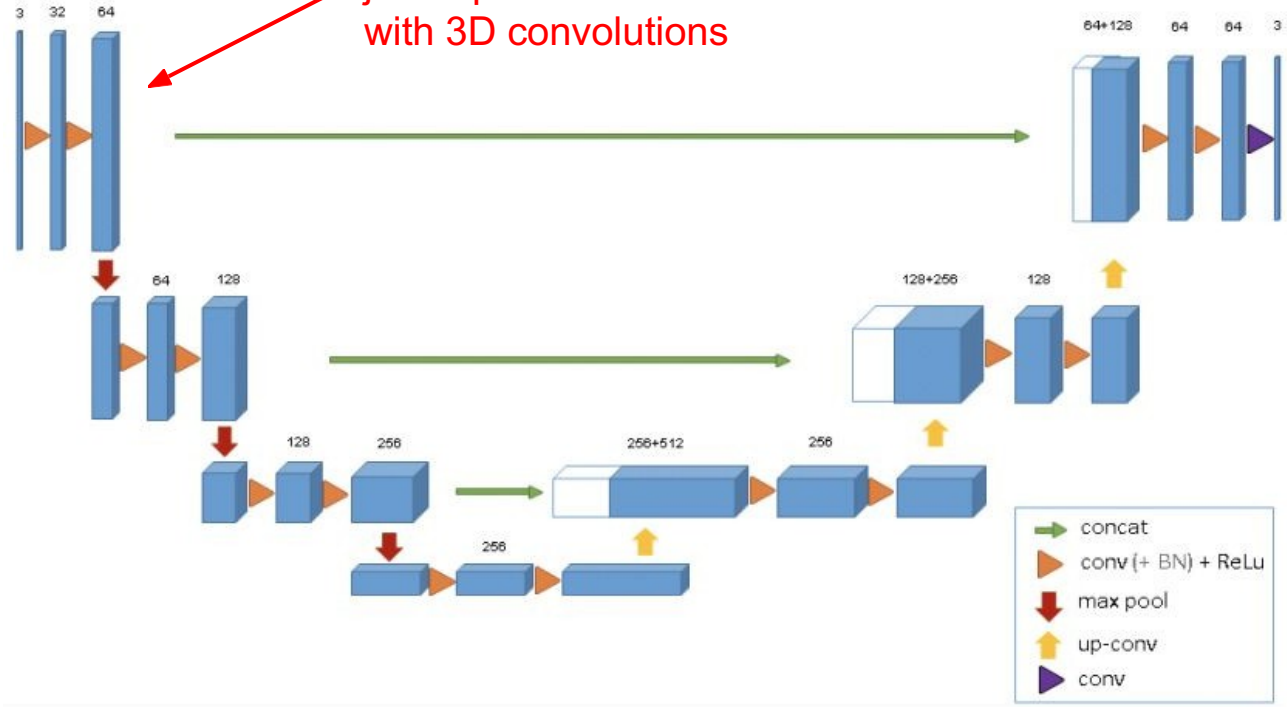
E.g. 3D U-Net



Cicek et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. MICCAI 2016.

E.g. 3D U-Net

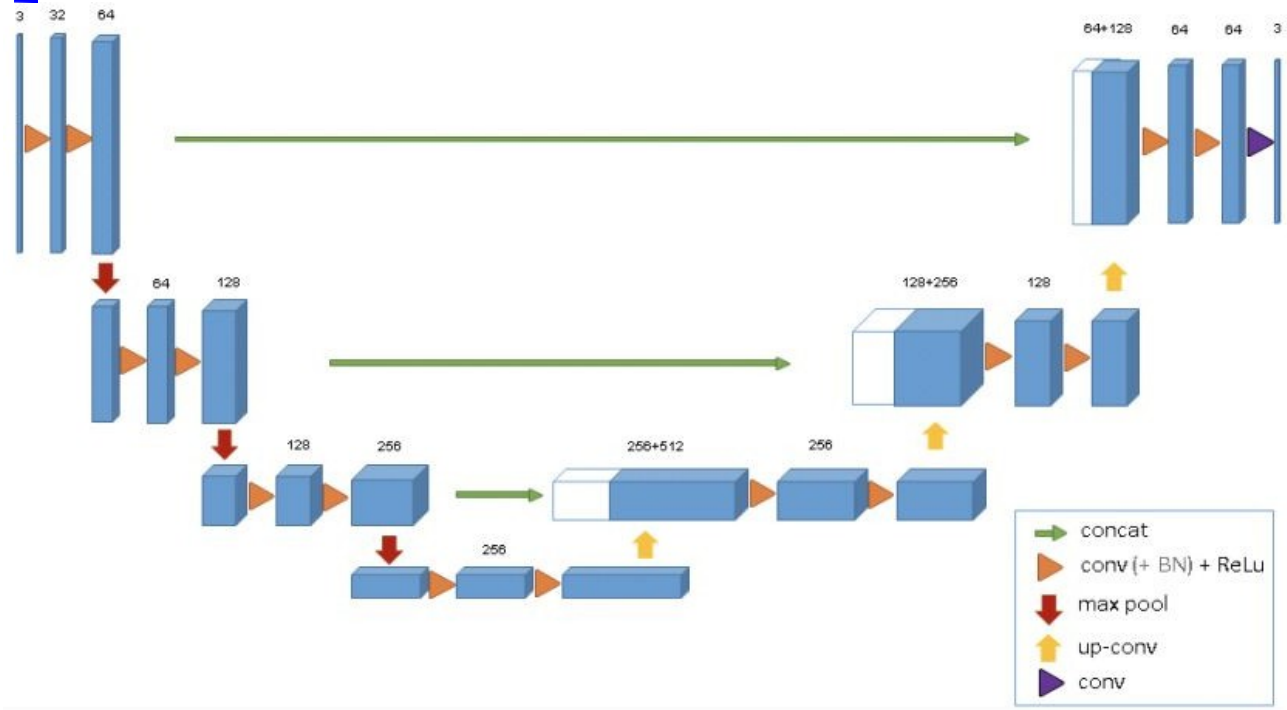
Same structure as 2D version, just replace all 2D convolutions with 3D convolutions



Cicek et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. MICCAI 2016.

E.g. 3D U-Net

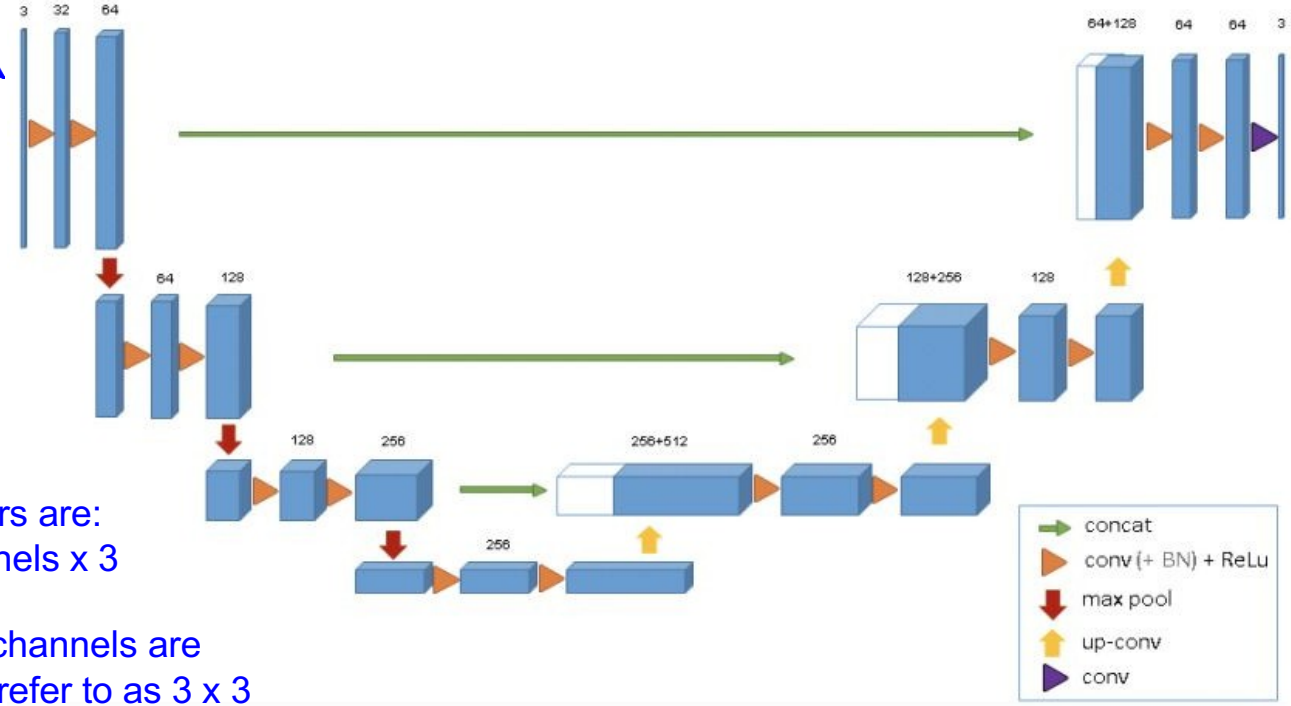
Channels



Cicek et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. MICCAI 2016.

E.g. 3D U-Net

Ex. input: 132 x 132 x 3 x 116



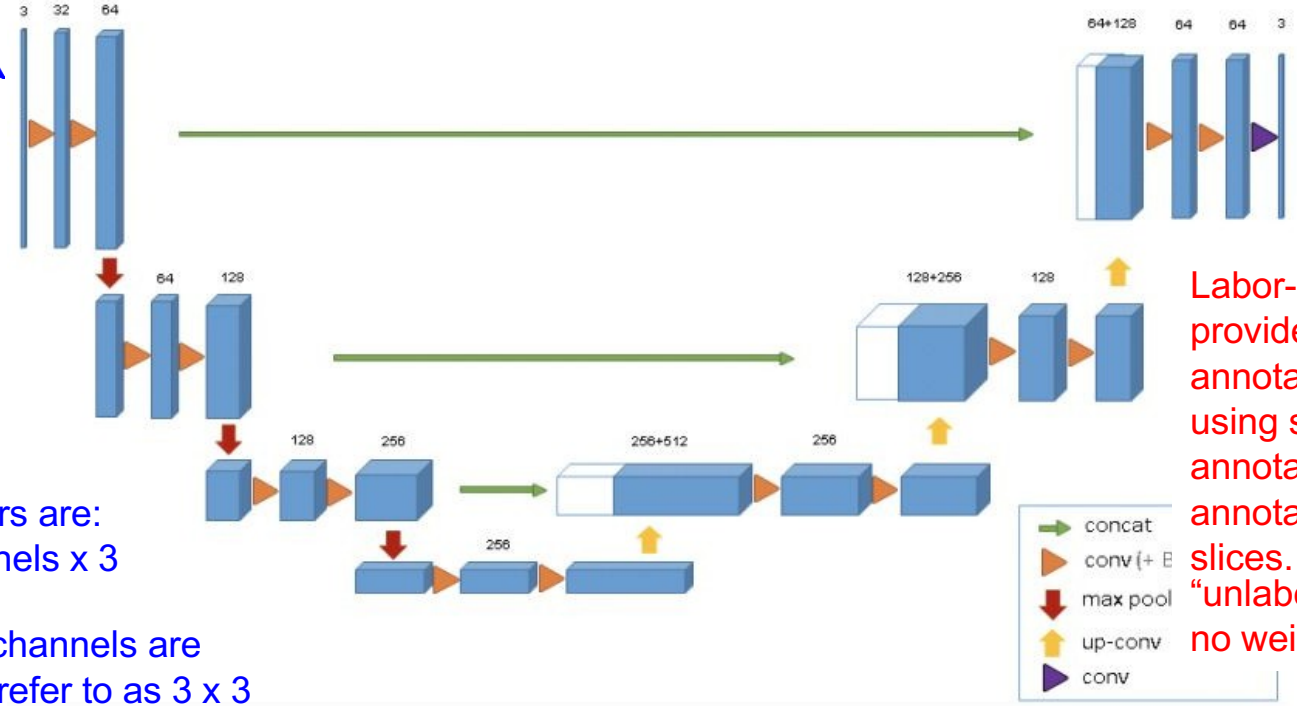
3D conv filters are:
3 x 3 x channels x 3

Sometimes channels are implicit, can refer to as 3 x 3 x 3 conv filter

Cicek et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. MICCAI 2016.

E.g. 3D U-Net

Ex. input: 132 x 132 x 3 x 116



3D conv filters are:
3 x 3 x channels x 3

Sometimes channels are implicit, can refer to as 3 x 3 x 3 conv filter

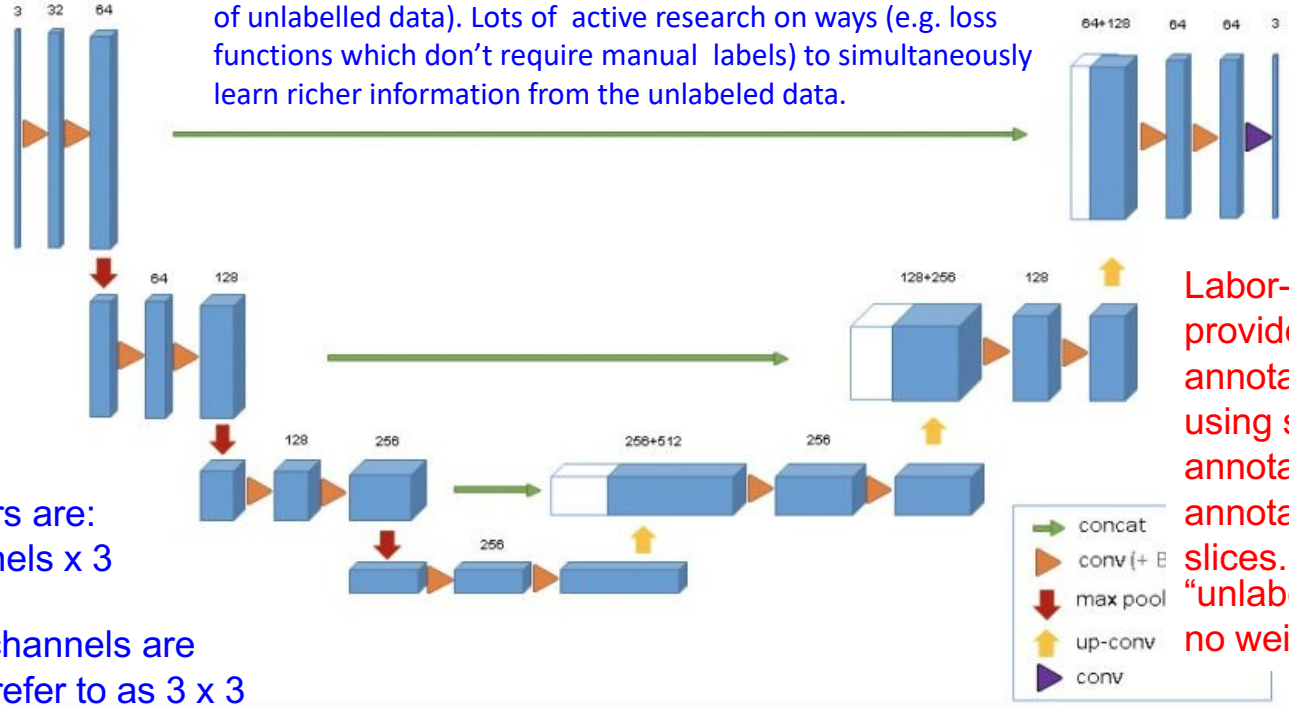
Labor-intensive to provide ground truth 3D annotation. Train instead using sparse annotations: a handful of annotated xy, xz, yz 2D slices. All others are "unlabeled" pixels with no weight in the loss.

Cicek et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. MICCAI 2016.

E.g. 3D U-Net

Ex. input: 132 x 132 x 3 x 116

Semi-supervised learning: learning from datasets that are partially labeled (small amount of labeled data + larger amount of unlabelled data). Lots of active research on ways (e.g. loss functions which don't require manual labels) to simultaneously learn richer information from the unlabeled data.



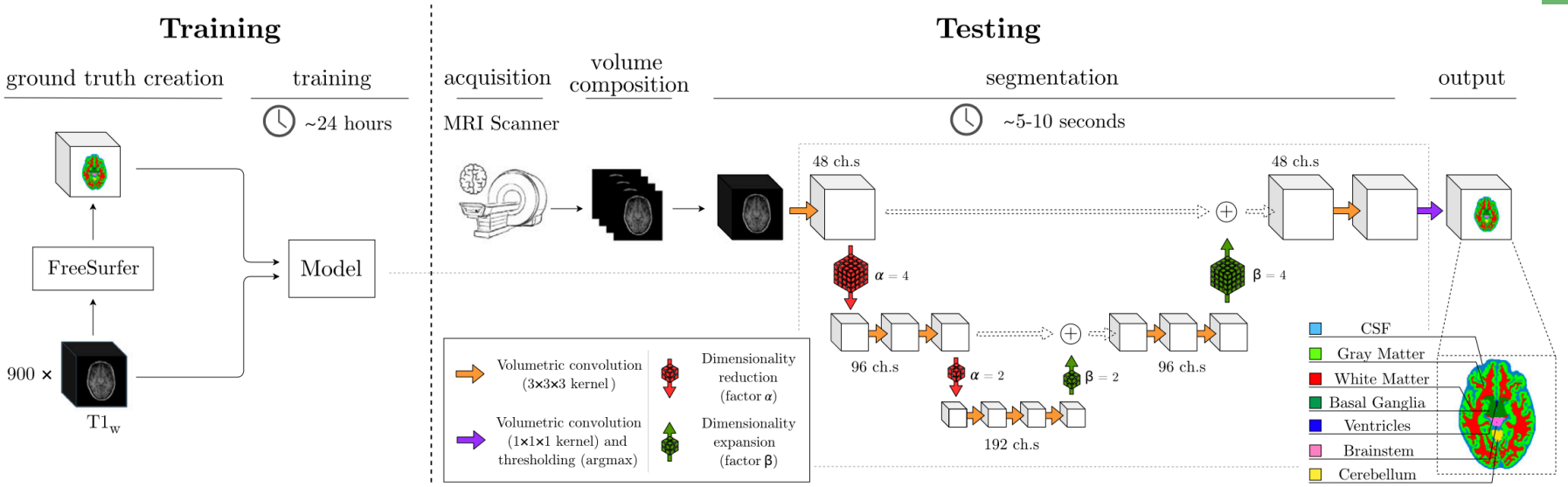
3D conv filters are:
3 x 3 x channels x 3

Sometimes channels are implicit, can refer to as 3 x 3 x 3 conv filter

Labor-intensive to provide ground truth 3D annotation. Train instead using sparse annotations: a handful of annotated xy, xz, yz 2D slices. All others are "unlabeled" pixels with no weight in the loss.

Cicek et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. MICCAI 2016.

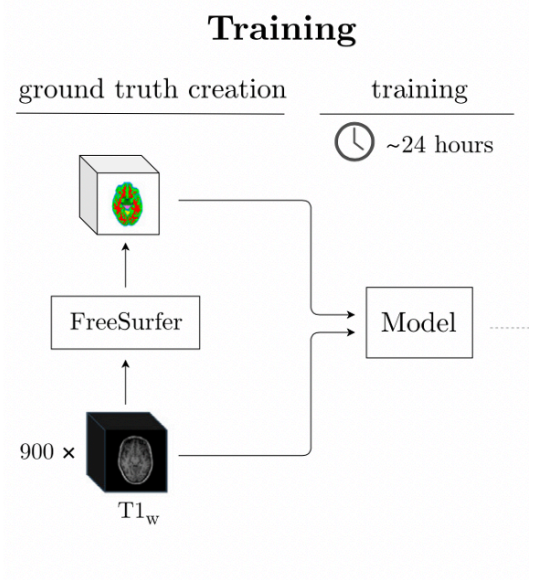
E.g. 3D U-Net in CEREBRUM Brain segmentation (Bontempi et.al)

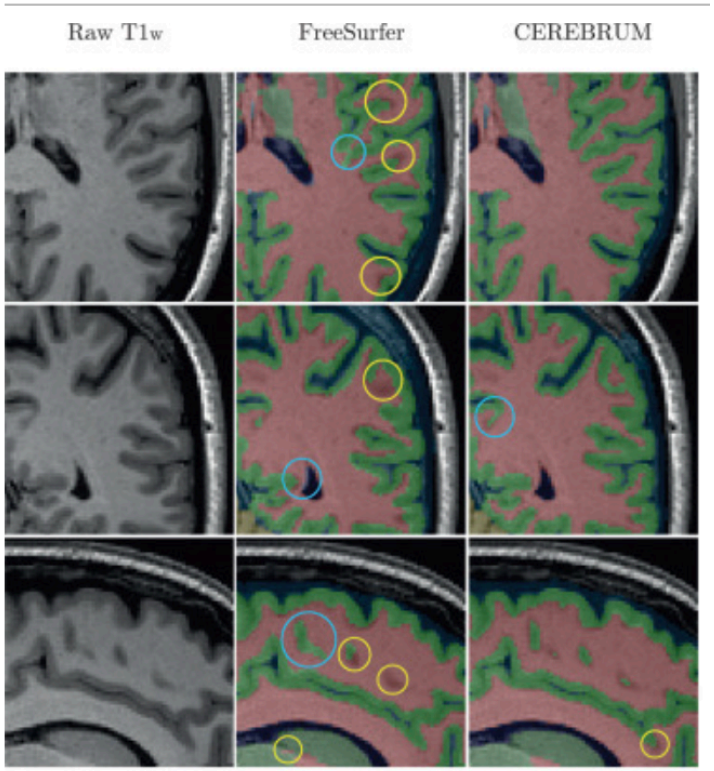


CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decodeR for weakly-supervised sEgmentation of BRain strUctures from out-of-the-scanner MRI
Dennis Bontempi , Sergio Benini , Alberto Signoroni , Michele Svanera , Lars Muckli
Medical Image Analysis, 2020
https://doi.org/10.1016/j.media.2020.101688

E.g. 3D U-Net in CEREBRUM Brain segmentation (Bontempi et.al)

- Why CEREBRUM is better than FreeSurfer?
  - Exploits weakly supervised learning
  - Out of the scanner method, does not require pre-normalization

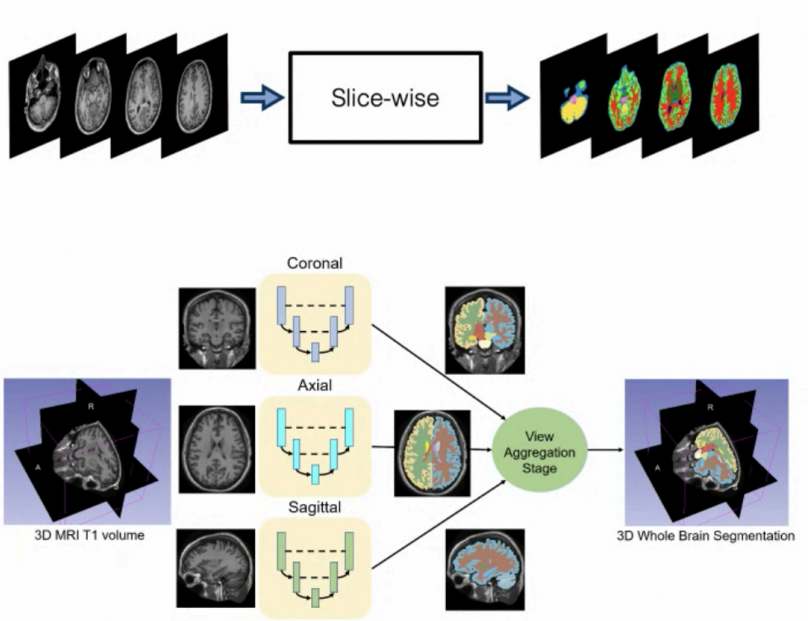


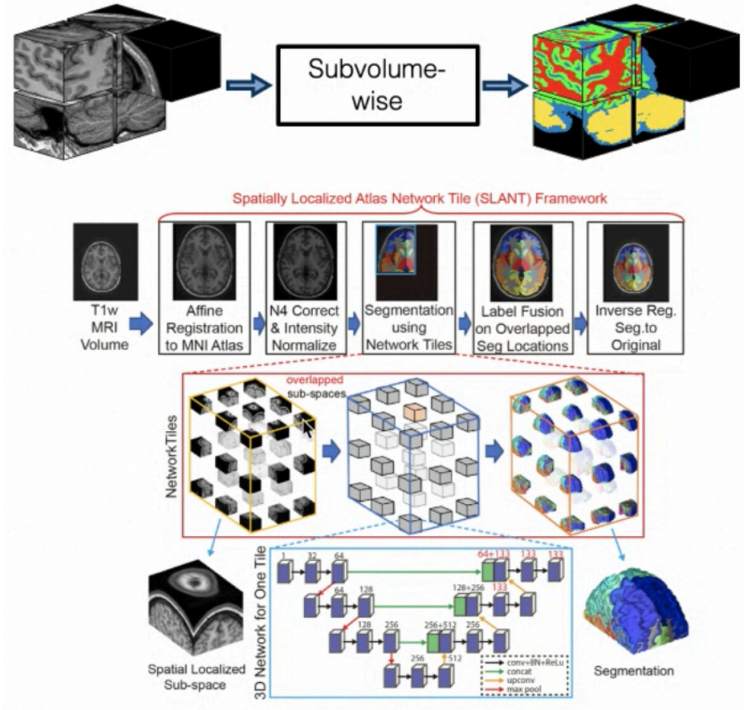


CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decodeR for weakly-supervised sEgmentation of BRain strUctures from out-of-the-scanner MRI
Dennis Bontempi , Sergio Benini , Alberto Signoroni , Michele Svanera , Lars Muckli - Medical Image Analysis, 2020 https://doi.org/10.1016/j.media.2020.101688

E.g. 3D U-Net in CEREBRUM Brain segmentation (Bontempi et.al)

- Why CEREBRUM is better than other DL- based systems?
  - 3D processing is more coherent and lead to better results (higher accuracy and confidence values) compared to methods based on slices or subvolumes

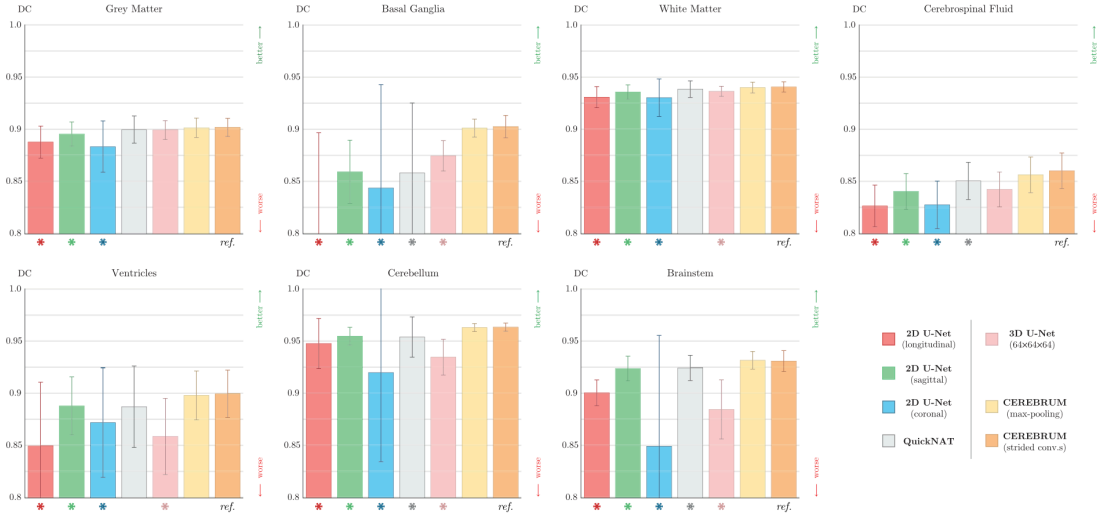


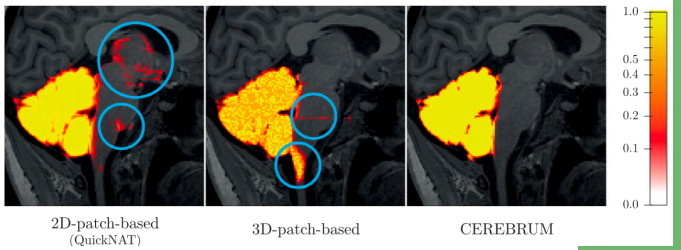


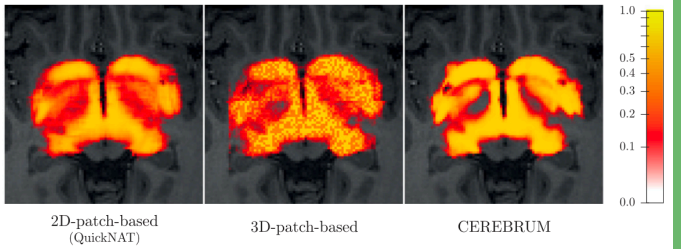
QuickNAT '19, FastSurfer '20 | SLANT '18, AssemblyNet '20

E.g. 3D U-Net in CEREBRUM Brain segmentation (Bontempi et.al)

- Why CEREBRUM is better than other DL- based systems?
  - 3D processing is more coherent and lead to better results (higher accuracy and confidence values) compared to methods based on slices or subvolumes







CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decodeR for weakly-supervised sEgmentation of BRain strUctures from out-of-the-scanner MRI
Dennis Bontempi , Sergio Benini , Alberto Signoroni , Michele Svanera , Lars Muckli - Medical Image Analysis, 2020 https://doi.org/10.1016/j.media.2020.101688