POLITECNICO DI TORINO

GRAINS
GRAphics and INtelligent Systems group

Deep learning applications in healthcare

Lia Morra

Politecnico di Torino - DAUIN

3rd Workshop SmartData@Polito

24/09/2018

Outline

- Brief introduction to computer aided diagnosis (or CAD)
- Deep learning in medical imaging: where do we stand?
- Medical images… and where to find them
- Challenges and lessons learnt in CAD research

The starting point

PREVENT DIAGNOSTIC ERRORS

Perceptual Misses

Lesions missed due to fatigue, distraction
Low disease prevalence
Small lesions
Satisfaction of search

Interpretation Errors

Inherent difficult problem
Unusual lesion characteristics
Reader inexperience
Cognitive biases

A shifting focus

PREVENT DIAGNOSTIC ERRORS

Perceptual Misses

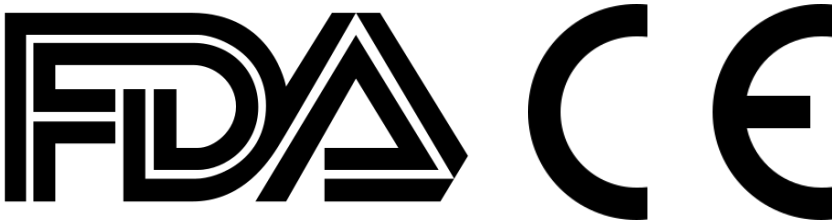
Interpretation Errors

Workflow optimization

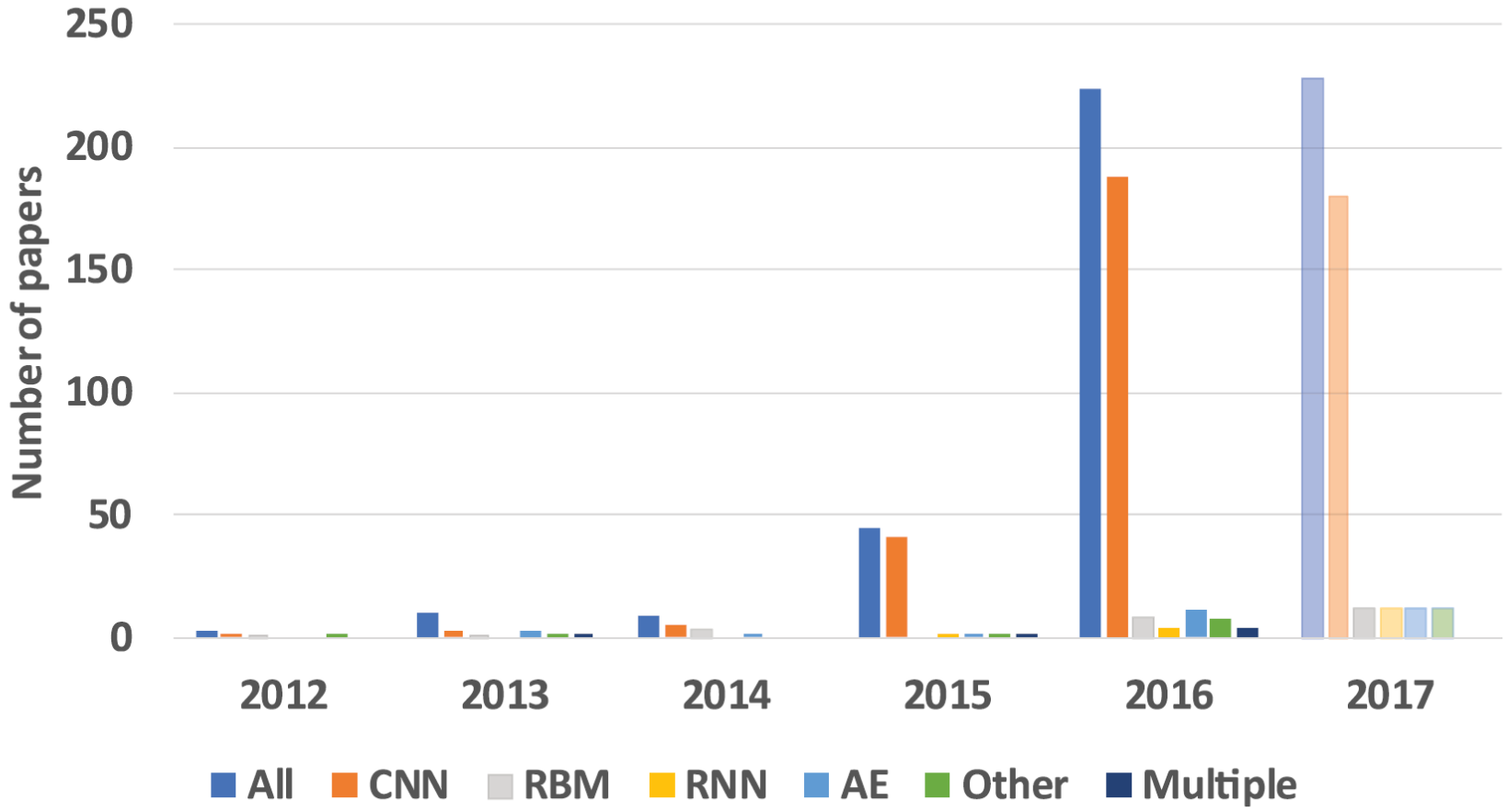
Increasing workload
Shift from 2D to 3D/4D/multi-modal imaging
Decreasing reimbursement rates
Advances in artificial intelligence

Software regulation

- From an industry perspective, diagnostic software are regulated as medical device
  - Emphasis on validation, pre- and post-market
  - Need to account for human factors
  - Strict requirements on software design
  - High impact on costs and time to market



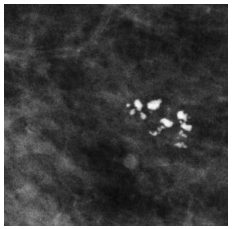
Deep learning in medical imaging



G. Litjens et al. / Medical Image Analysis 42 (2017) 60–88

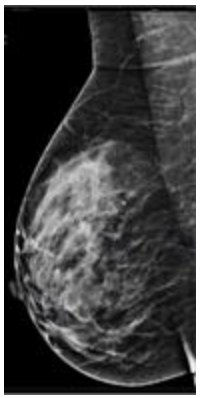
Brief recap on CNN

Patches

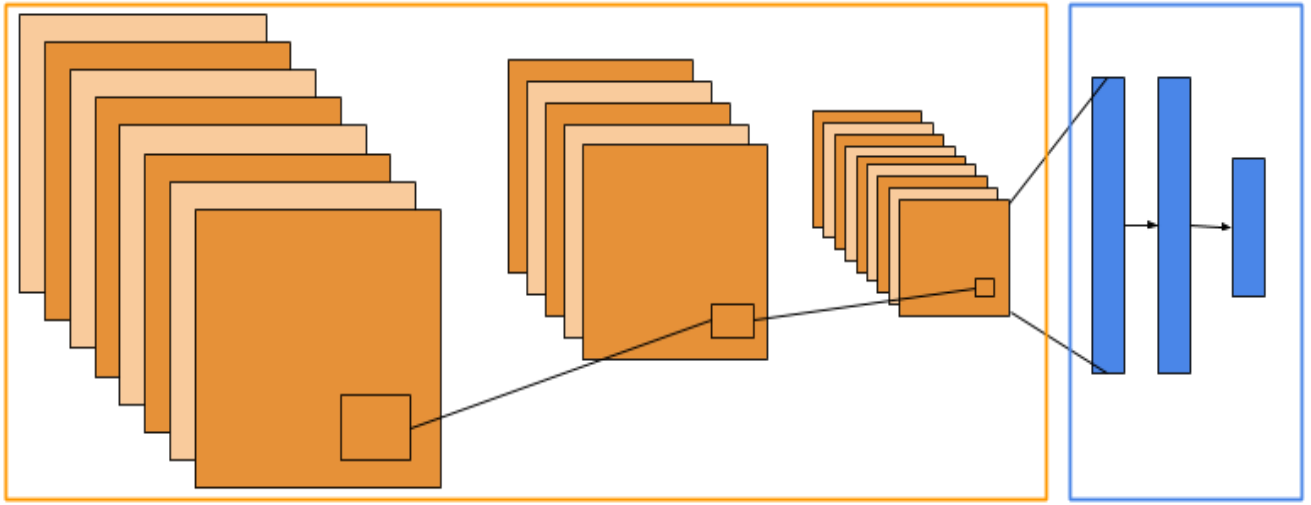


Many medical images are too large to fit into GPU memory

Images



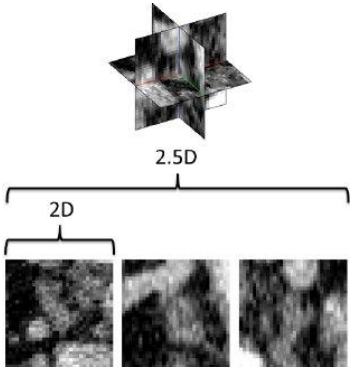
CLASSIFICATION



Diagnosis

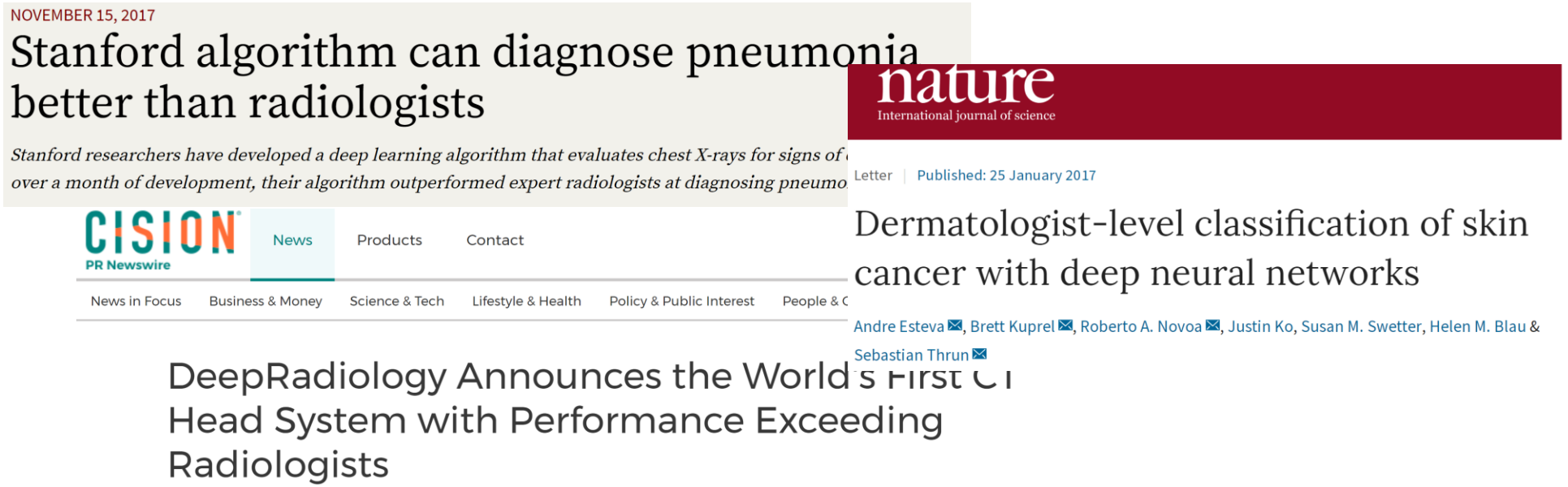
REPRESENTATION LEARNING

2.5 Conv



Dealing efficiently with 3D volumes is still an open issue!

Human-level performance?



NOVEMBER 15, 2017

Stanford algorithm can diagnose pneumonia better than radiologists

Stanford researchers have developed a deep learning algorithm that evaluates chest X-rays for signs of
over a month of development, their algorithm outperformed expert radiologists at diagnosing pneumo

CISION PR Newswire

News | Products | Contact

News in Focus | Business & Money | Science & Tech | Lifestyle & Health | Policy & Public Interest | People &

DeepRadiology Announces the World's First CT Head System with Performance Exceeding Radiologists

nature
International journal of science

Letter | Published: 25 January 2017

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

- Ability to achieve performance close or equal to radiologist level largely depends on the availability of large scale database (100,000x cases)
- Best results have been obtained with “simple” pathologies and images, mostly 2D

Data is the main limiting factor

#1 Lack of (annotated) training data

#2 Barriers to access (privacy, fragmentation)

#3 Data imbalance

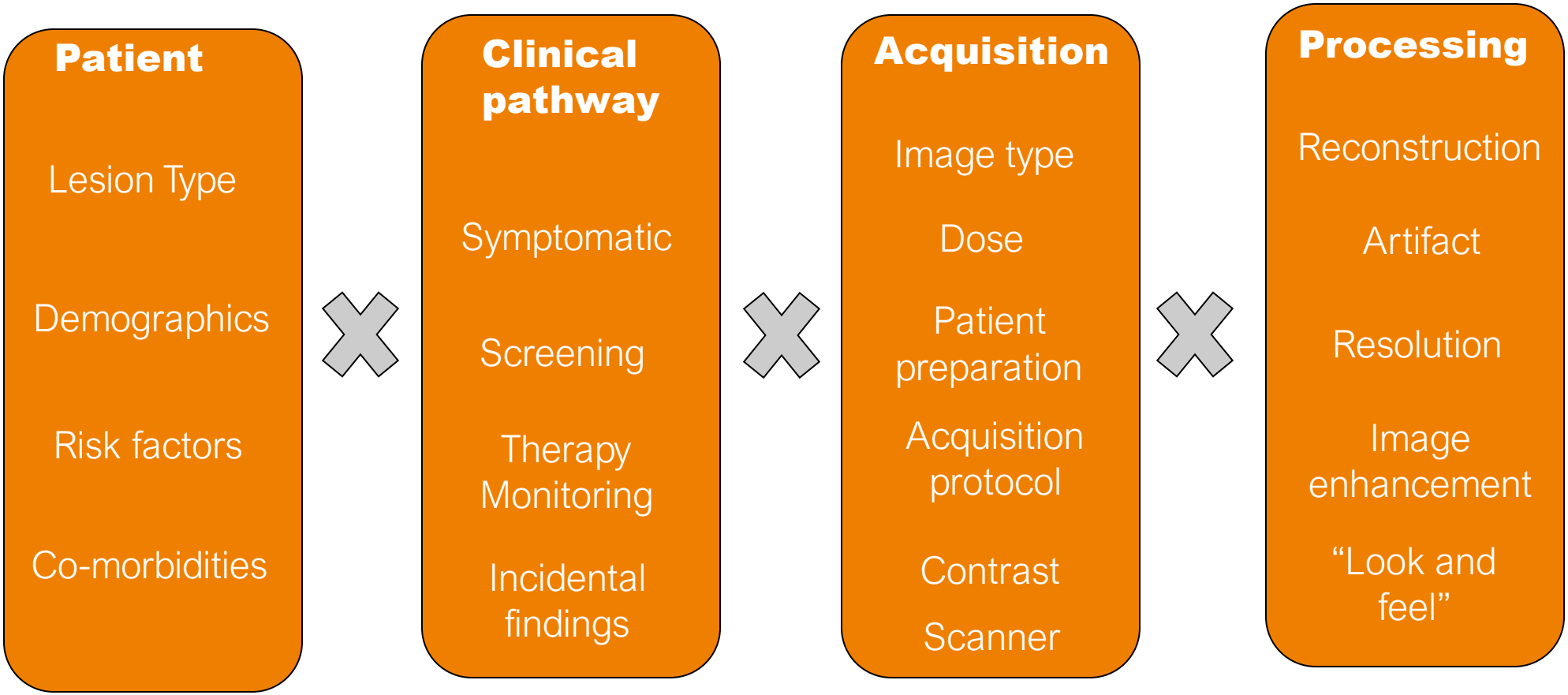
#4 Population biases

#5 Confounding factors

#6 Technological evolution

#7 Noisy ground truths

A complex universe

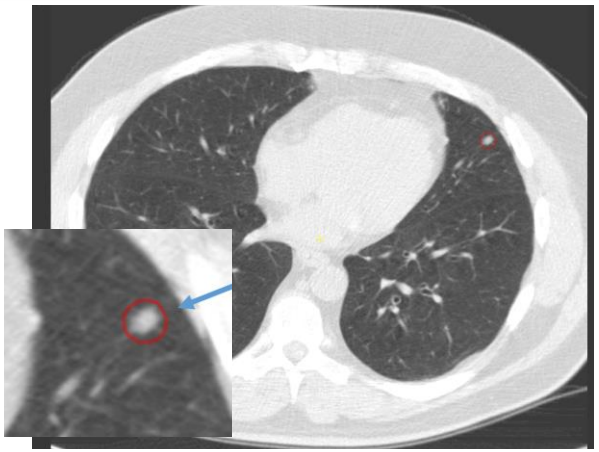


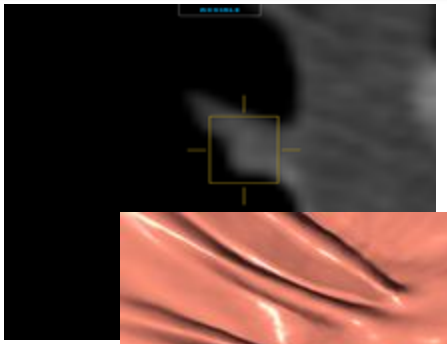
Population bias

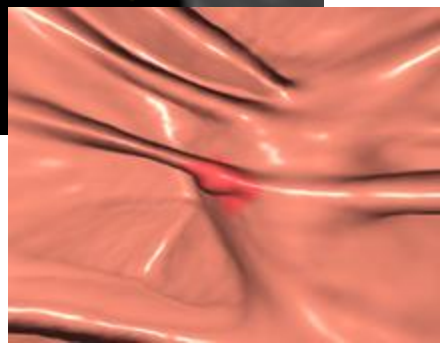
Symptomatic - diagnostic

Bias towards

- Large lesions
- Higher dose protocols
- Use of contrast agents
- Higher resolution
- High disease prevalence (selection bias)







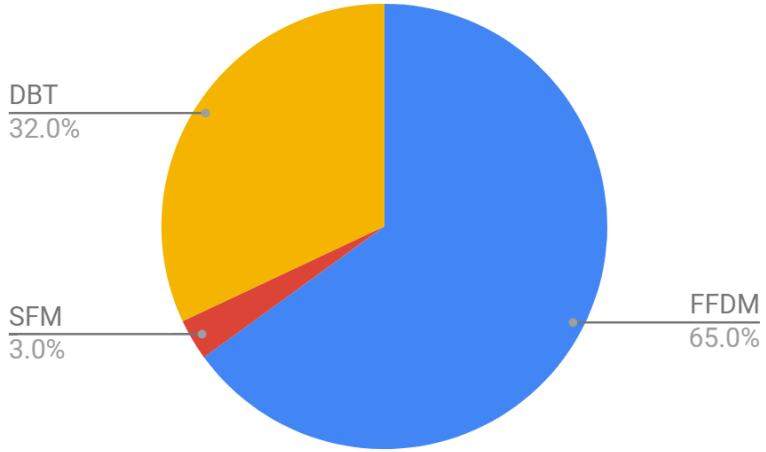
Bias towards

- Small (preneoplastic) lesions
- Patient-friendly protocols (low dose, lower resolution, no contrast agents)
- Low disease prevalence

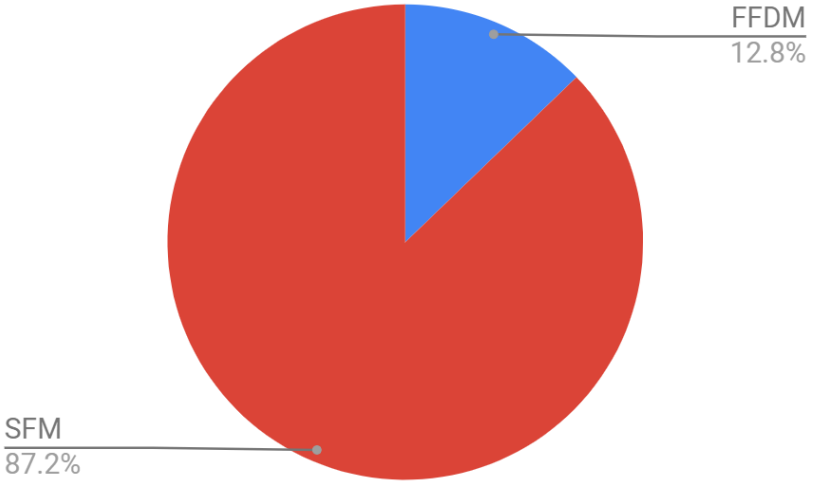
Asymptomatic - screening

Technological evolution

Datasets become stale with technical advances in acquisition modalities



Mammography market share (2016 - US)



Distribution of publicly available datasets

Breast screening transition from Screen Mammography (SFM) to Digital Mammography (FFDM) and then to Tomosynthesis or 3D mammography (DBT)

Establishing ground truth

Radiologist reports

- Most practical choice
- Radiologists' sensitivity is not perfect
  - usually in the 70% - 95% range, depending on the application
- High inter- and intra-rater variability

Independent ground truth

- The best practice is to rely on an independent device or procedure, ideally with higher accuracy (such as biopsy)
- Alternatively, use clinical follow-up (6 months to 2 years)
- May require clinical trials if additional procedures are not medically justified

Example I: LIDC-IDRI lung dataset

Only 1940 (26.3%) of the 7371 nodules demonstrate complete agreement

Table 1

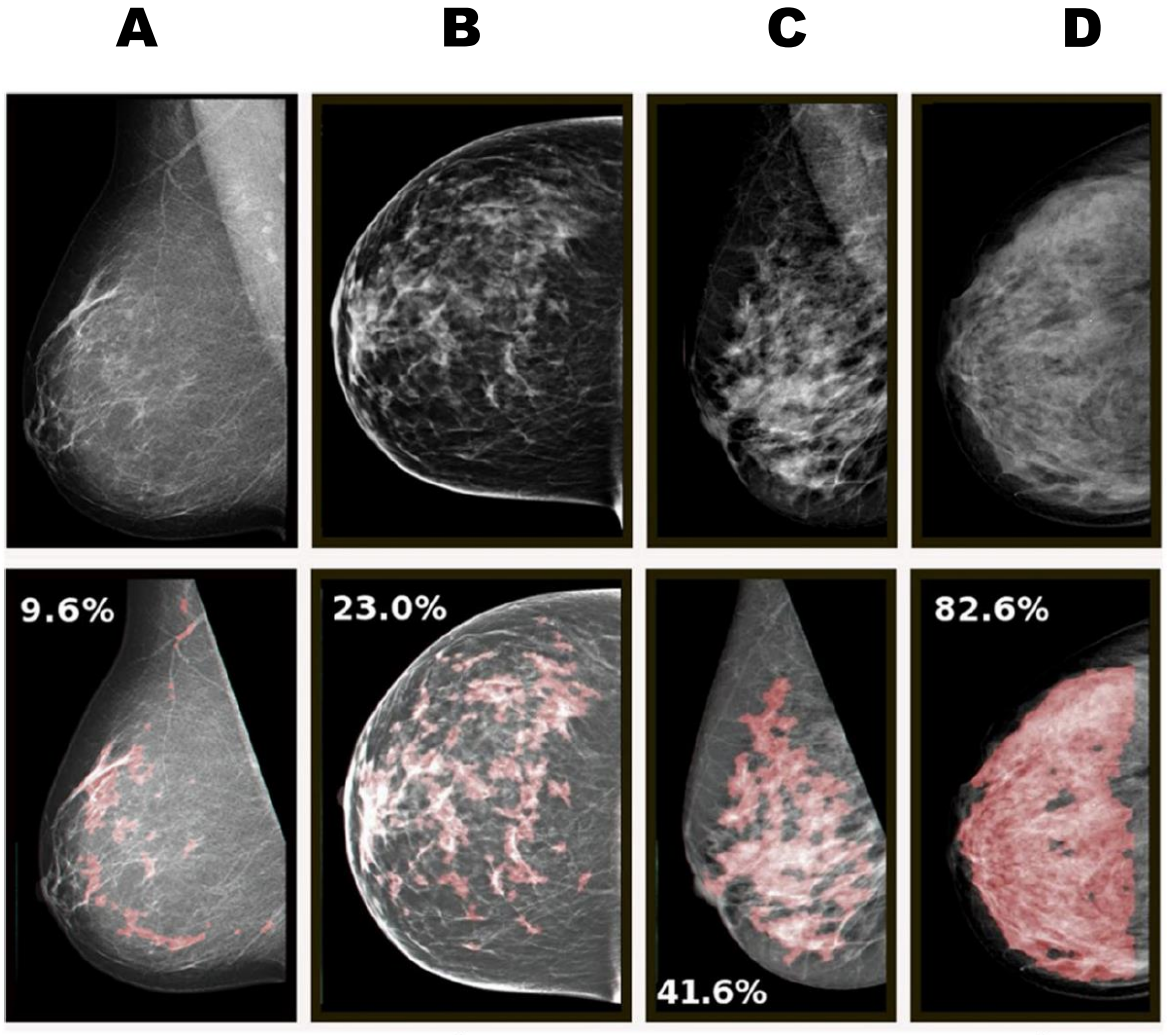
Summary of lesions identified by LIDC/IDRI radiologists across all 1018 CT scans.

Description	Number of lesions
At least one radiologist assigned either a nodule≥3 mm mark or a nodule<3 mm mark	7371
At least one radiologist assigned a nodule≥3 mm mark	2669
All four radiologists assigned a nodule≥3 mm mark	928
All four radiologists assigned a nodule≥3 mm mark or all four radiologists assigned a nodule<3 mm mark	1940
All four radiologists assigned either a nodule≥3 mm mark or a nodule<3 mm mark	2562

S.G. Armato et al / Med Phys 2011 38 (2): 915 - 931

Example II: Breast density

Radiologist BI-RADS class

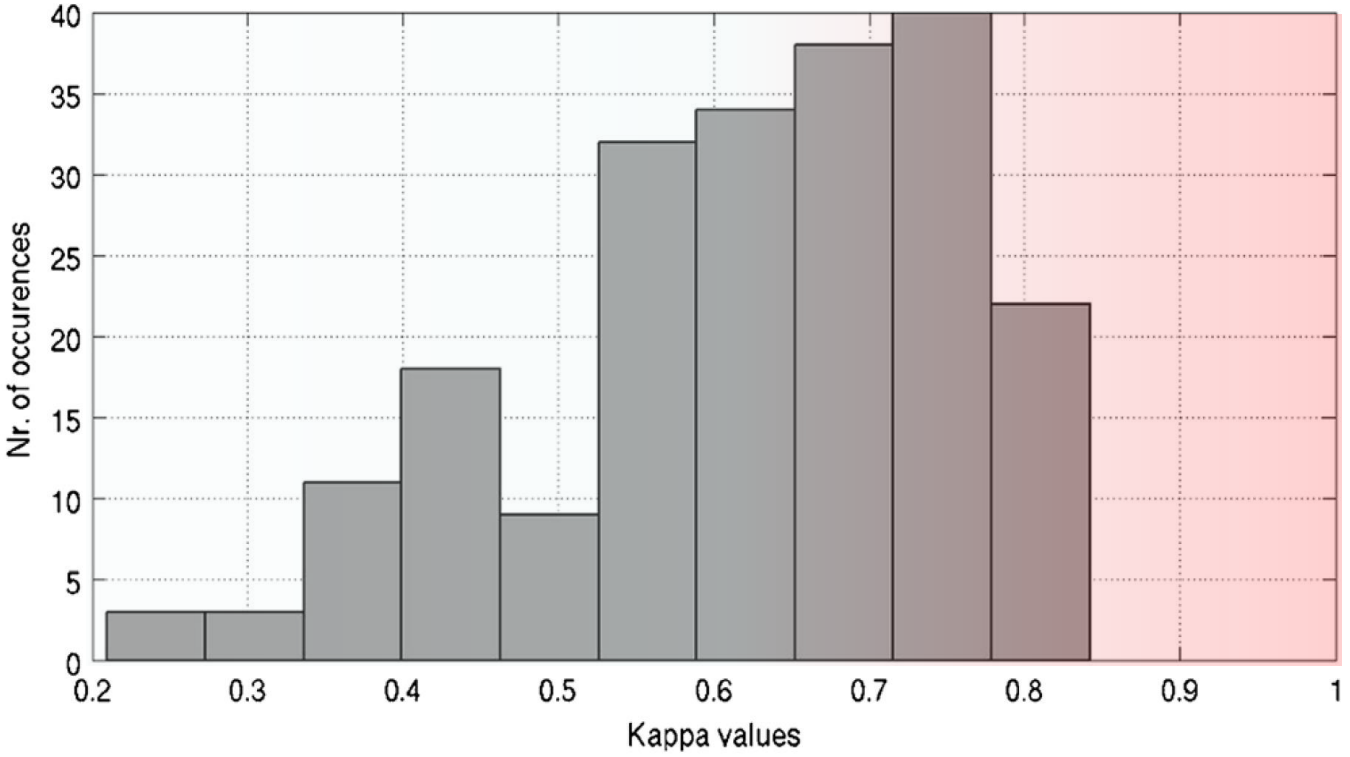


Automatic software

D. Sacchetto et al / Eur Radiol 2016 26(1):175-83

Example II: Breast density

Distribution of pairwise agreement in a group of 21 radiologists shows moderate agreement in discriminating dense vs non-dense breasts



D. Sacchetto et al / Eur Radiol 2016 26(1):175-83

Key points

#1 Handle heterogeneous data with care

#2 Test properly

#3 Explore transfer & multi-task learning

#4 Beware of human factors

Key lessons #1 – Handle with care

- Beware of the range of disease, patient and imaging features
- Test for built-in biases
- If training on heterogeneous data, at least test on realistic data

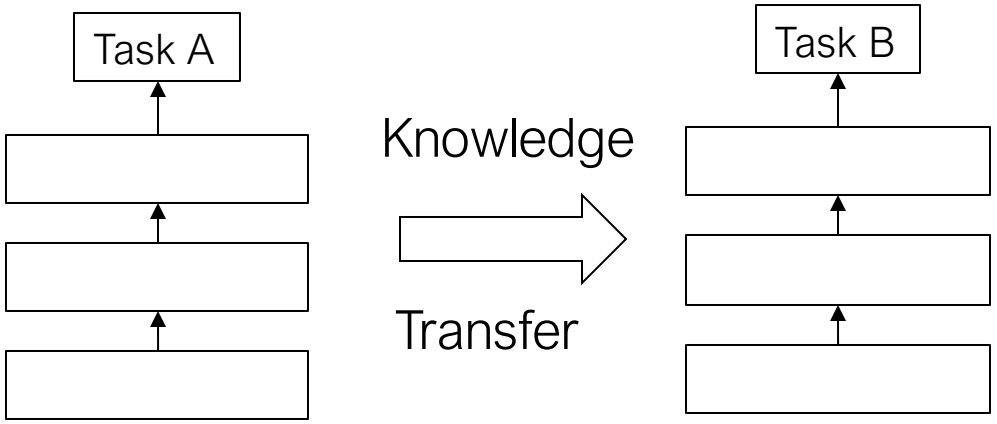
Key lessons #2 – Test properly

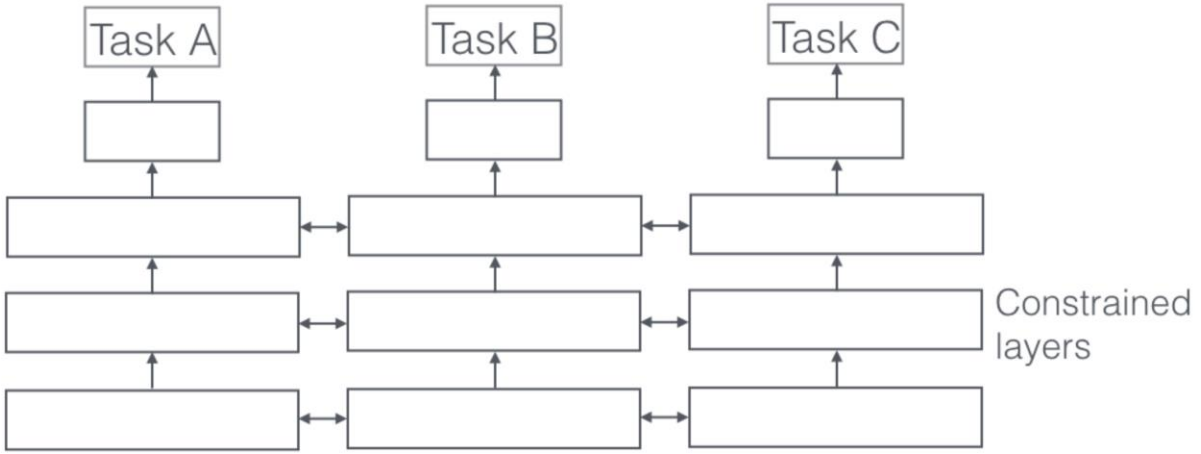
> Ideally, the composition of the test data set should match the target population to which the CAD system is intended to be applied. Also, image acquisition and patient preparation parameters should ideally be representative of those found in the target population. To allow proper interpretation of test results, inclusion and exclusion criteria must be clearly stated and must be justified as necessary. The distribution of the known covariates (e.g., lesion type and size, disease stage, organ characteristics, patient age, etc.) should be specified, and any significant departure from those of the target population should be identified and discussed in the publications that report the result of performance assessment.

N. Petrick et al / Med Phys 2013 40(8)

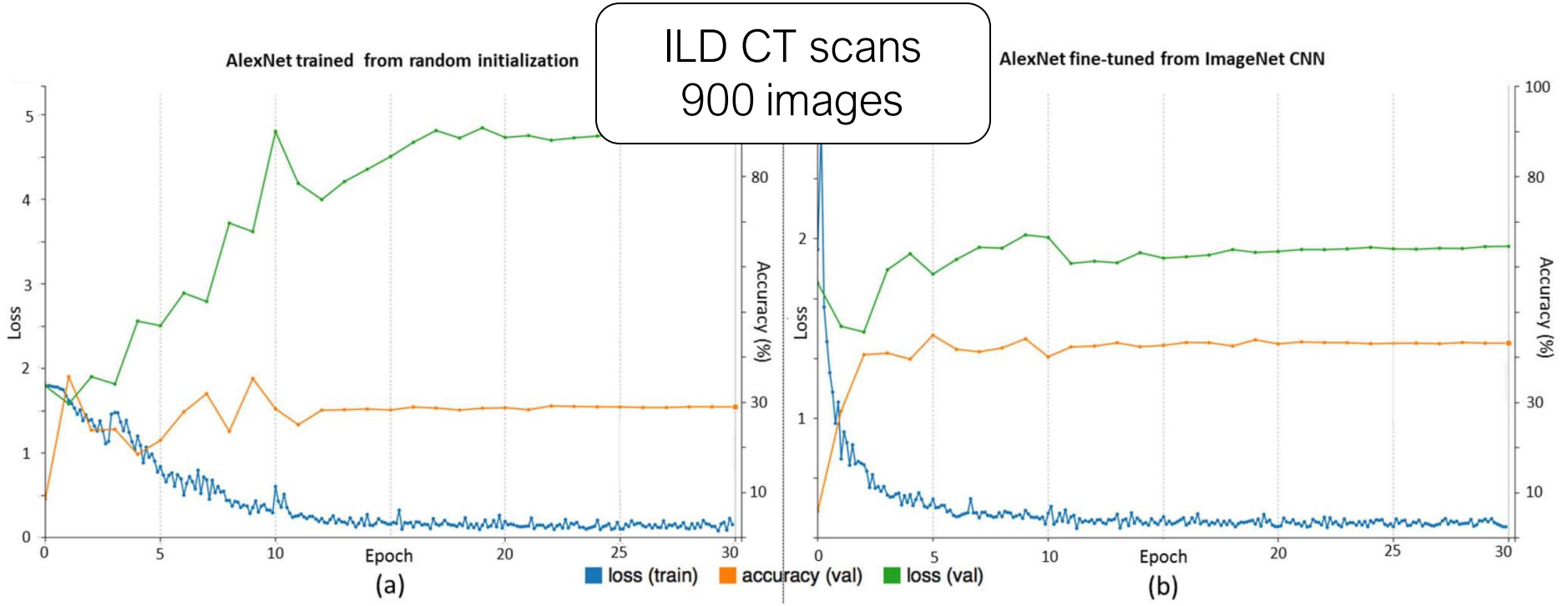
- "Ideally", an independent representative testing set should be used to assess performance
- Beware of selection biases
- Beware of verification biases
- Beware of test-set re-use
- Estimate confidence interval
- Use performance metrics that are insensitive to prevalence

Transfer and multi-task learning





Transfer learning (I)



Fine-tuning from ImageNet pre-trained networks works well in literature for medium datasets but...

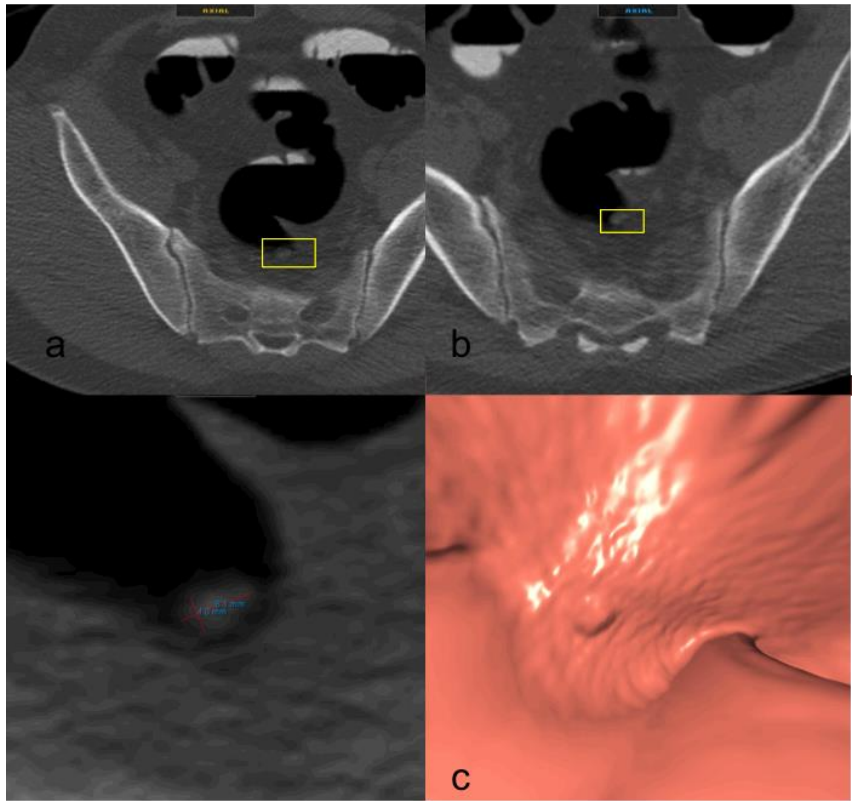
…. highly constrained in architectural choices!

H-C. Shin et al / IEEE T Med Imag 2016 35(5)

Key lesson #4 – Human factors

- Radiologist is ultimately responsible for the diagnosis….
- Practically, effect on radiologist performance has often proven less than expected
- Deep learning? We'll see

Example of colonic polyp detected by a commercial CAD system, but discarded by the radiologist



Questions?