THERAPIXEL

TECHNOLOGY. FOR LIFE. FOR ALL.

Special for MVA course

Deep Learning in practice: MammoScreen

MATHÉMATIQUES
VISION
APPRENTISSAGE

www.therapixel.com

February, 24 2020

by Yaroslav Nikulin,
Senior Research Scientist

Part I: Introduction

1. Therapixel
2. DL -> radiology
3. Breast cancer
4. DM DREAM Challenge

Therapixel: Medical Image Understanding

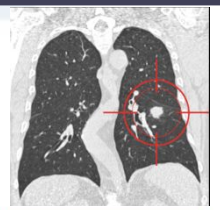
Mammscreen

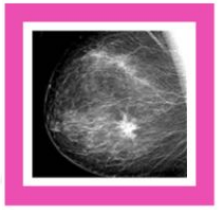
Clinical Study
Therapixel Cloud

2013 — Founded

2015 — Visualization SW

2016 — AI research

2017 — 
5th place Kaggle Data Bowl

2018 — 
1st place DREAM DM

2019

2020 — MammoScreen deployed

Olivier Clatz, PhD

Pierre Fillard, PhD

Breast Cancer Screening: some key stats

- 33M exams/year = 132M images in US alone
- $7.8 billion - cost of mammography screening in US (2010)
- 120 sec: average interpretation time.

1 out of 8

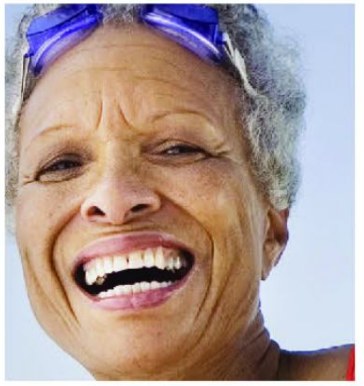
Woman affected during her lifetime

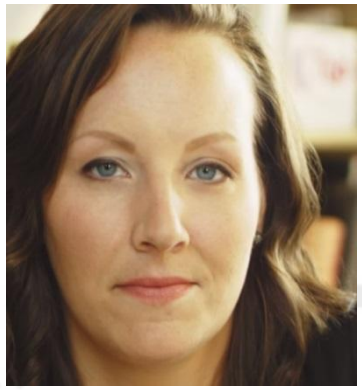
10 recall for

100 screened

5 cancers for

1000 screening





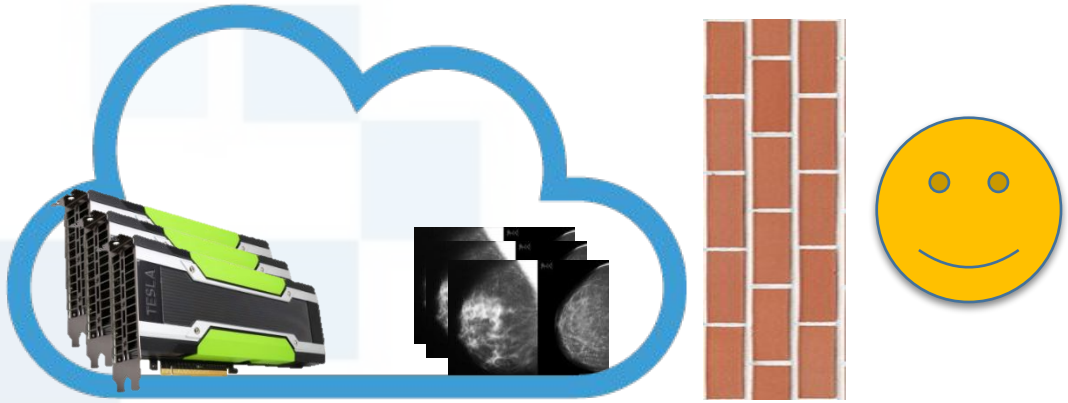
"If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future."

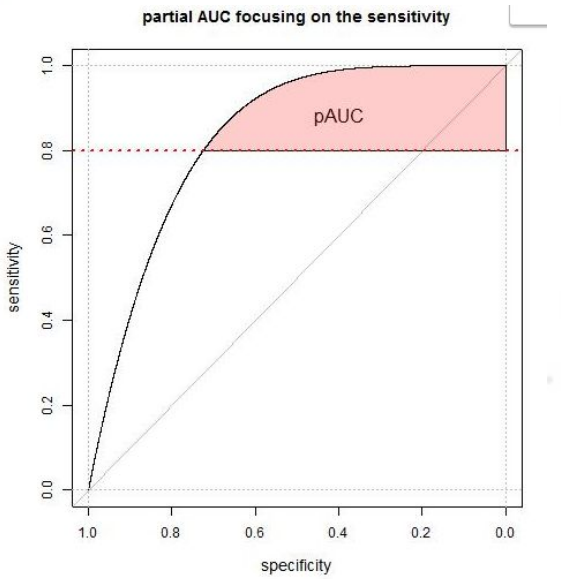
Andrew Ng, 2016

The Digital Mammography DREAM Challenge

Challenge setting:

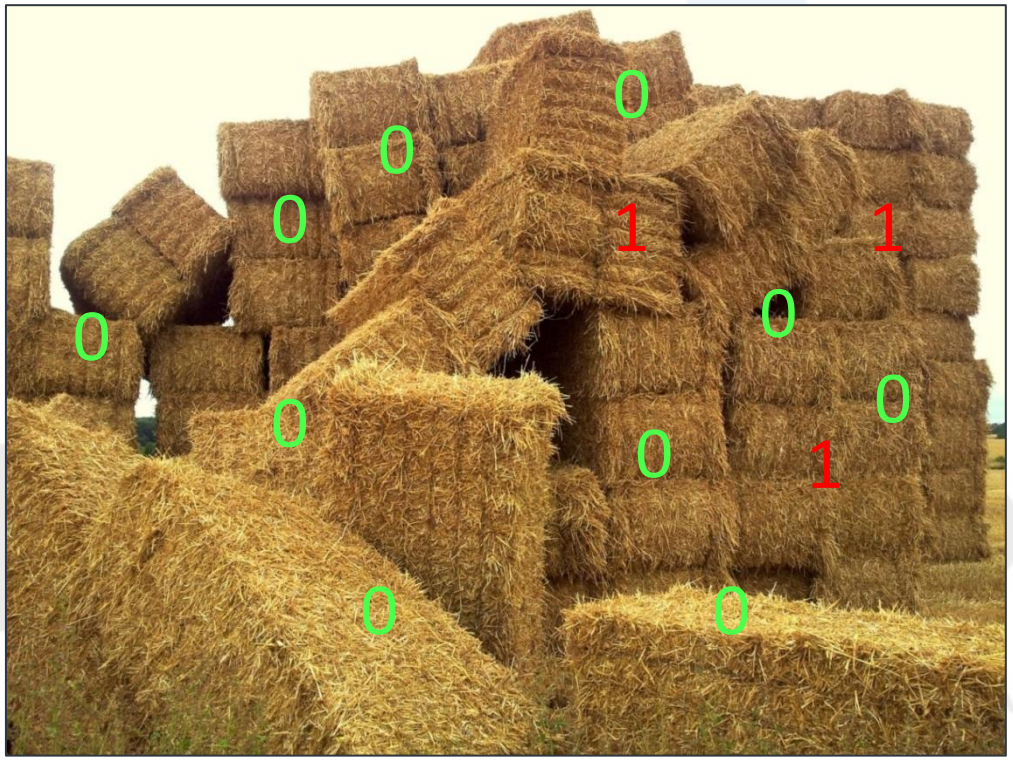
- Completely in the cloud
- 22 CPU cores **+** 2 GPUs
- 14 days / per team / round
- Performance measure:
  AUROC and **partial AUROC**





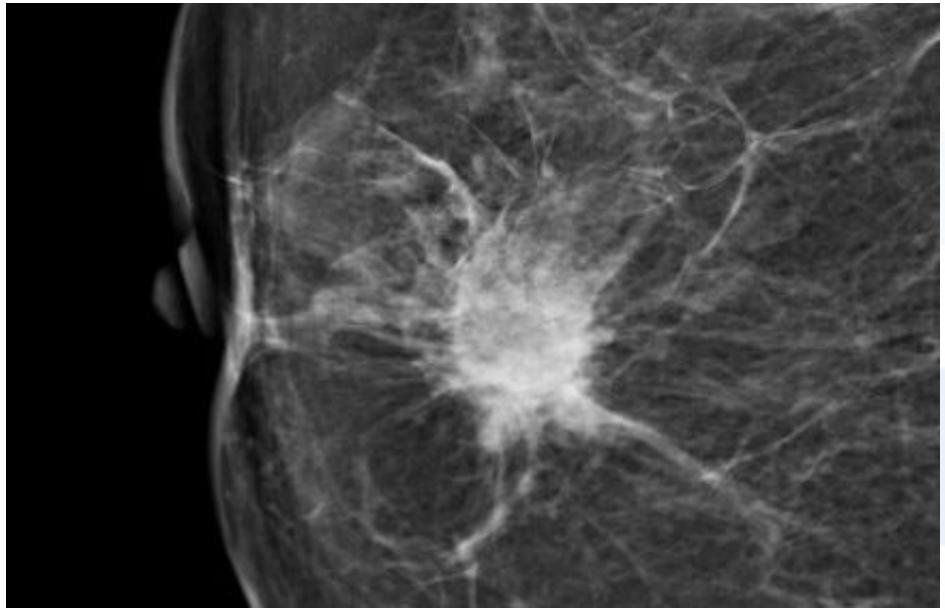
Why it is difficult - challenges of the Challenge

- 300k images
- Only 1114 (**0.35%!**) positive examples
- High resolution: from **3328x2560 to 5928x4728**.
- One single label per image: 0 or 1

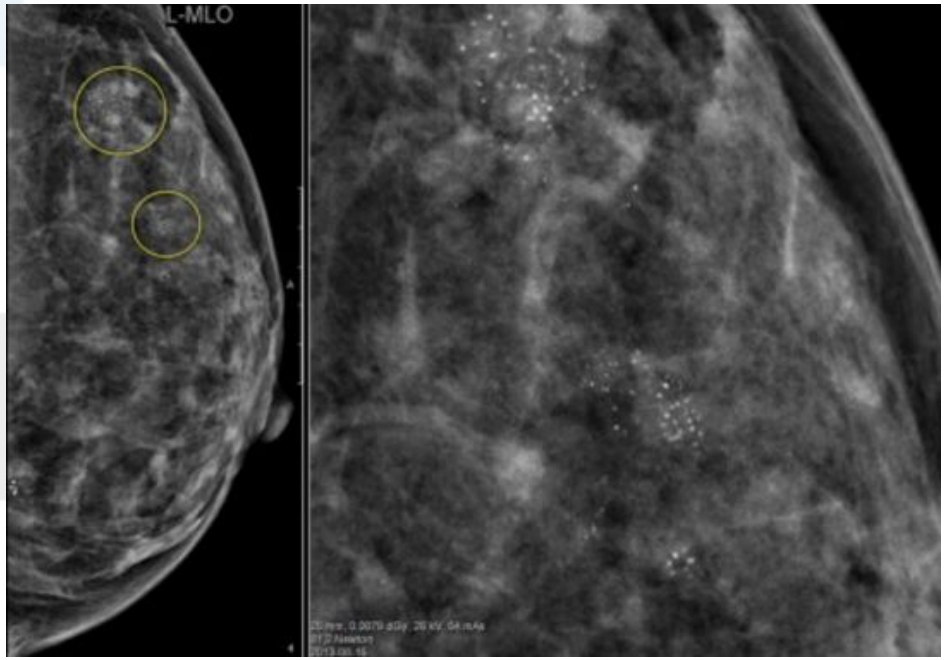


Now look for a needle in them…

Why it is difficult - challenges of the Challenge



- Different kinds of anomalies:
  - calcifications
  - masses
  - distorsions
  - asymmetries



- Different scales of anomalies: from micro-calcifications to big cancerous masses.

Can be malignant OR benign!

Part II: Winning solution: dream_net

1. Data specificity
2. Hack: dense annotations
3. Patch model
4. Image model

Why is it very different from ImageNet?

In our approach, limited by several factors. Actually 3-5 times higher

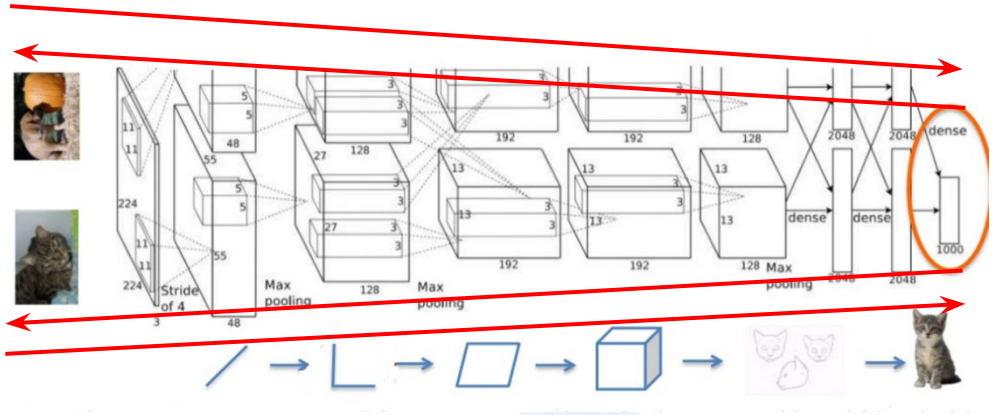
- Resolution: 1200x800 **vs** 224x224
- Zone of Interest : < 1% **vs** > 50%
- Number of classes : 2 **vs** 1000
- Highly imbalanced **vs** roughly balanced

Zone of Interest →

Why don’t DL results generalize always well to a new domain?

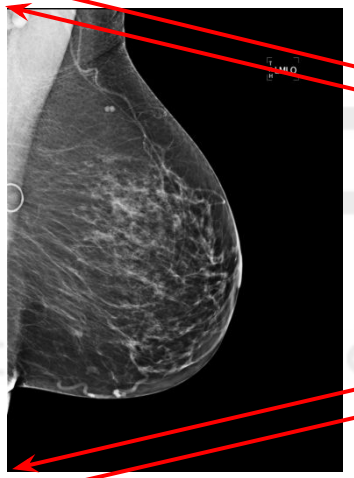
AlexNet (Krizhevsky et al. 2012)

Input size:
224x224



Output : 1 out of 1000
~ 10 bits of information

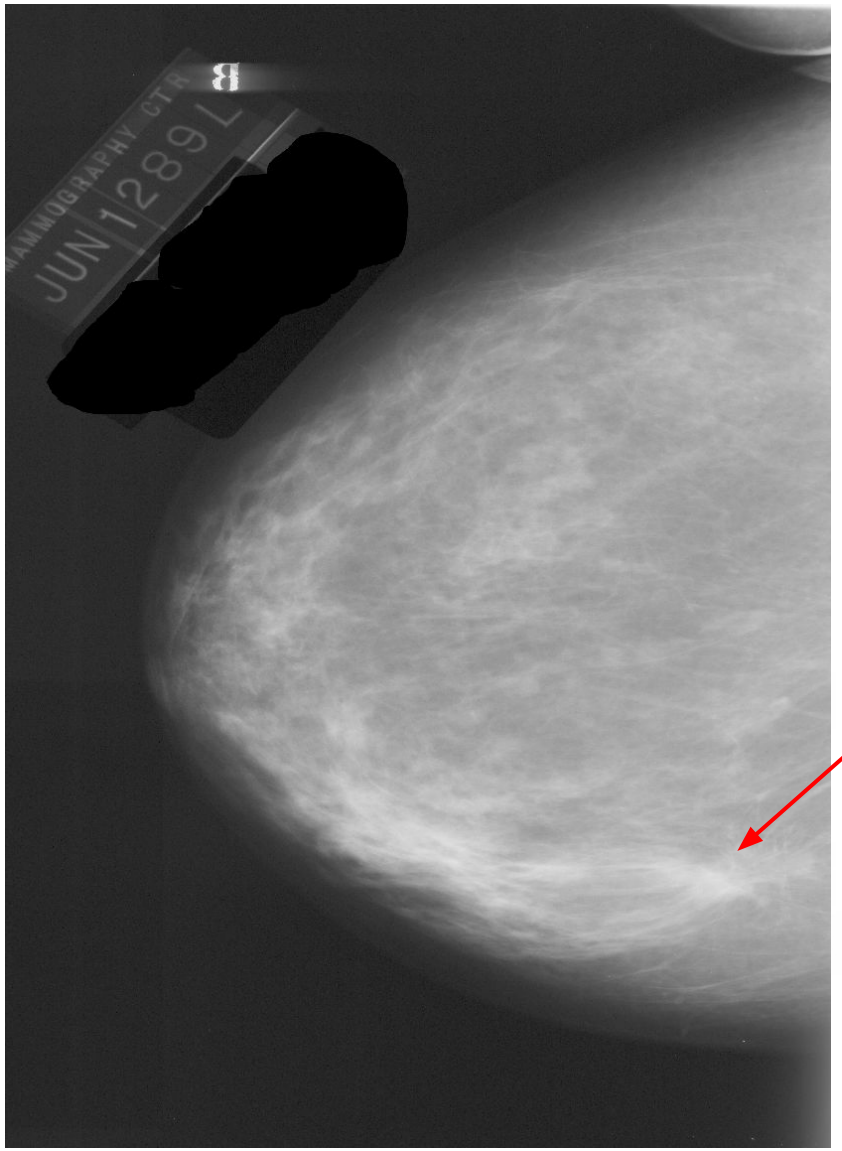
Input size:
~3500x2500



0 or 1

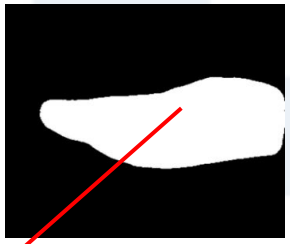
Output : 1 out of 2
= 1 bit of information

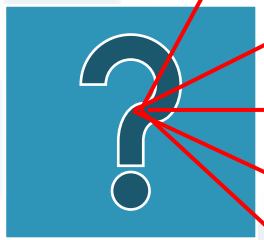
DDSM – bridge towards solution



	DDSM	DREAM
Total im	10k	640k
Positives	1807	1548
Info	mask&type	0 or 1

segmentation mask patch

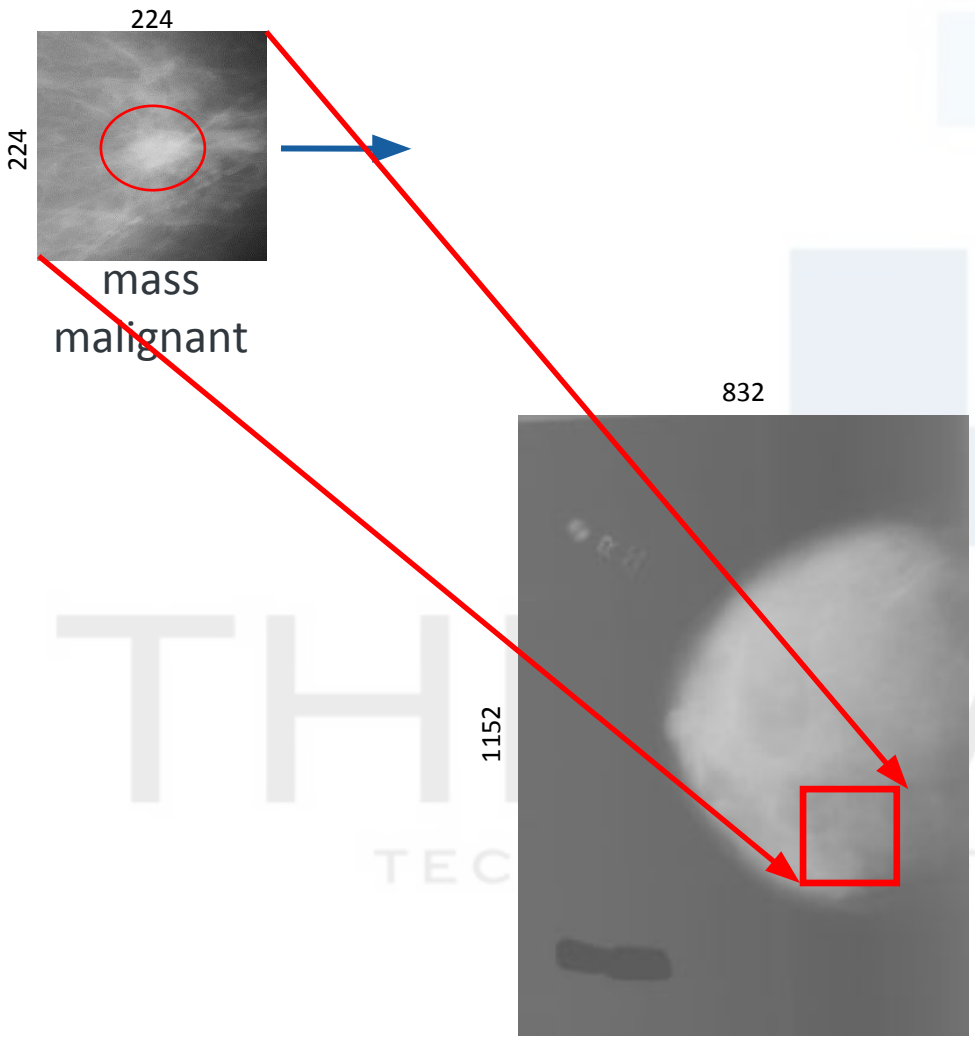




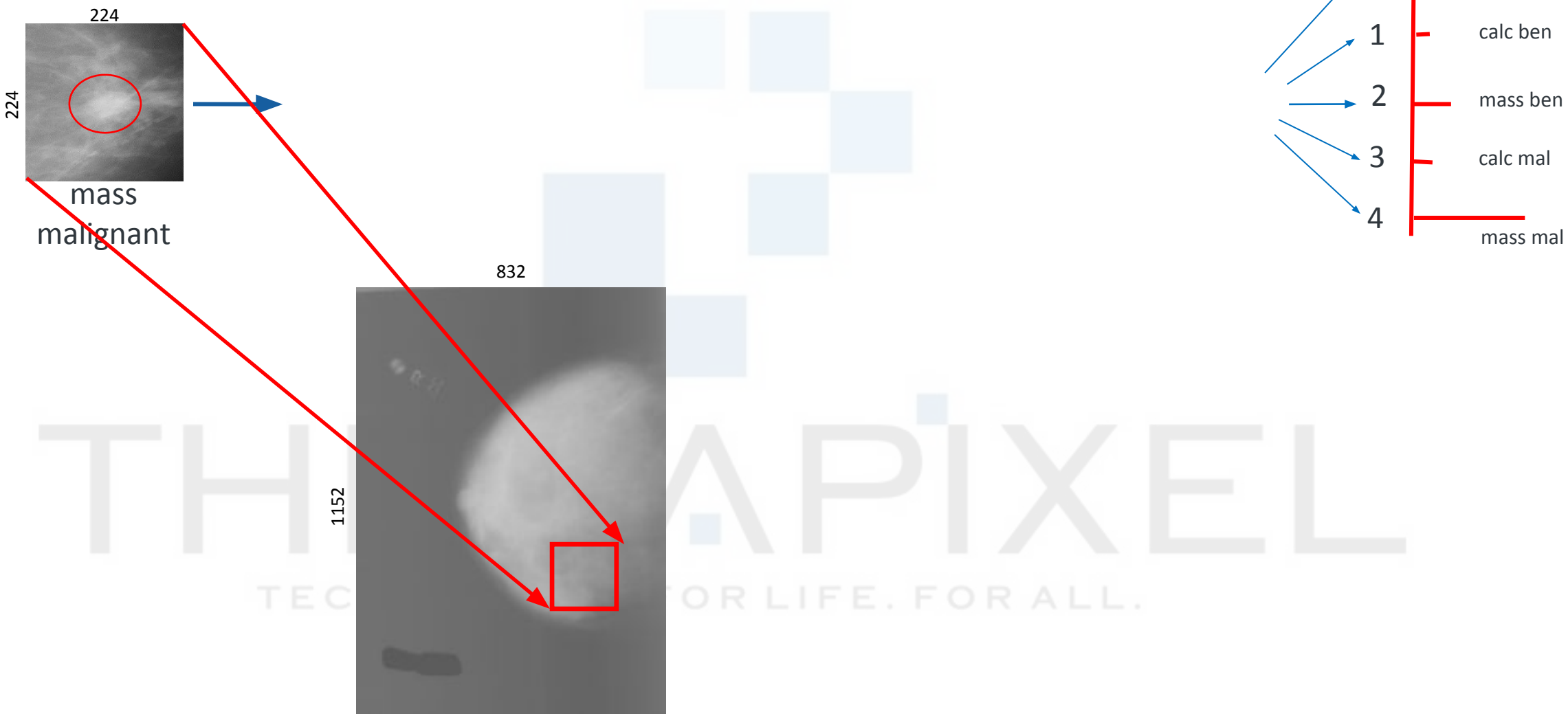
It would be great to:

- Make use of local info
- Make use of lesion type
- Still be able to train on DREAM

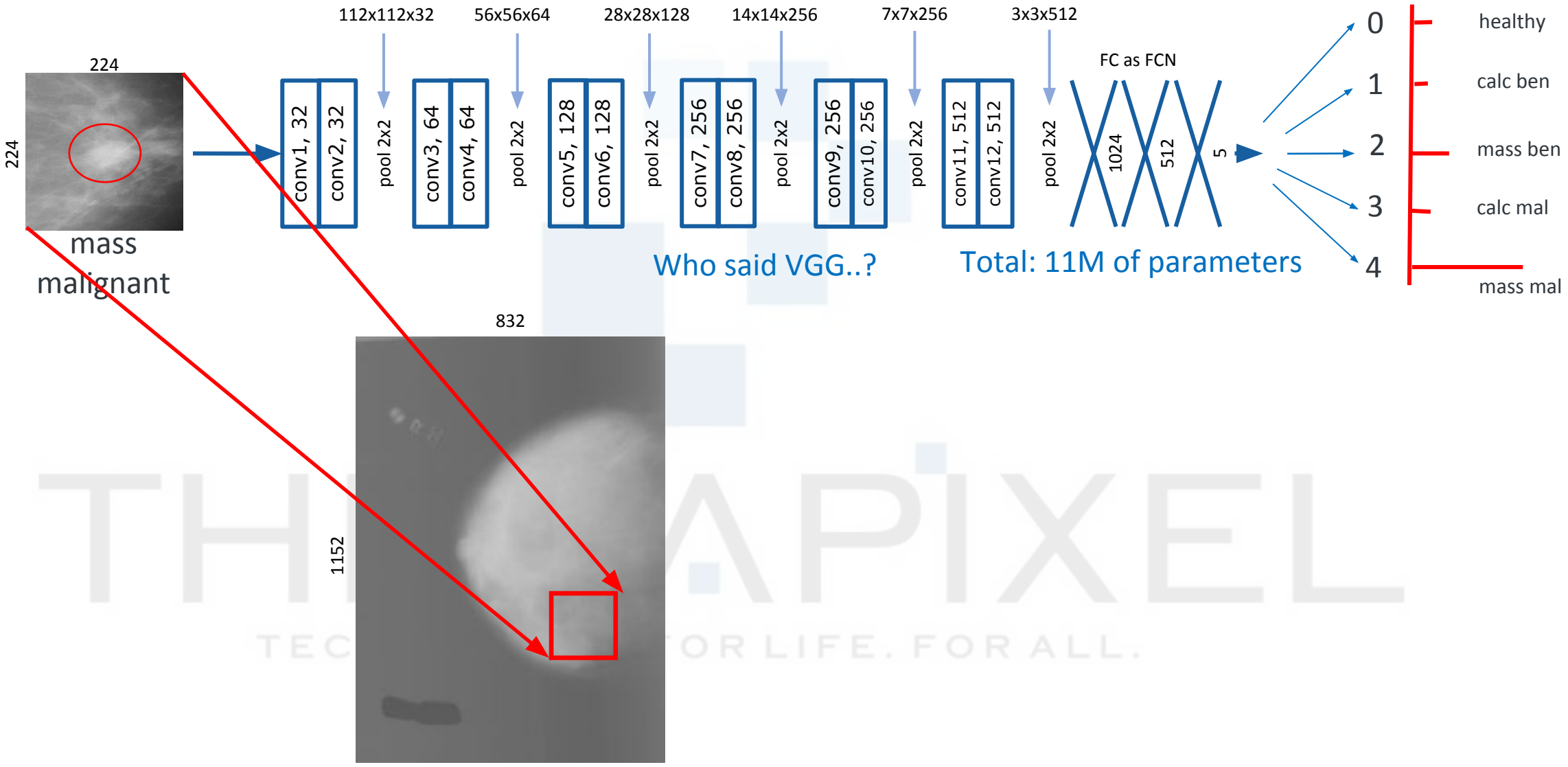
Patch model: Fully Convolutional Network



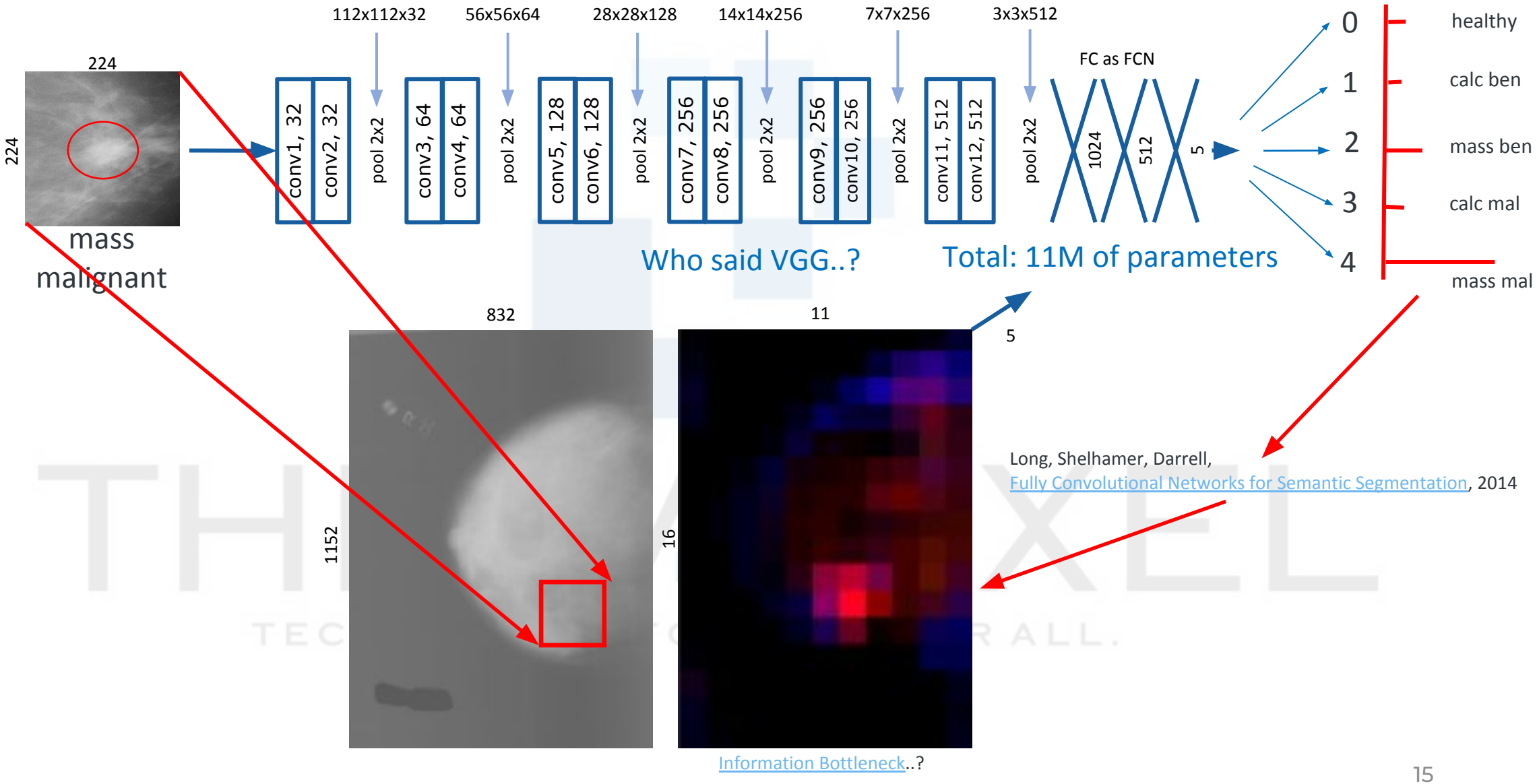
Patch model: Fully Convolutional Network



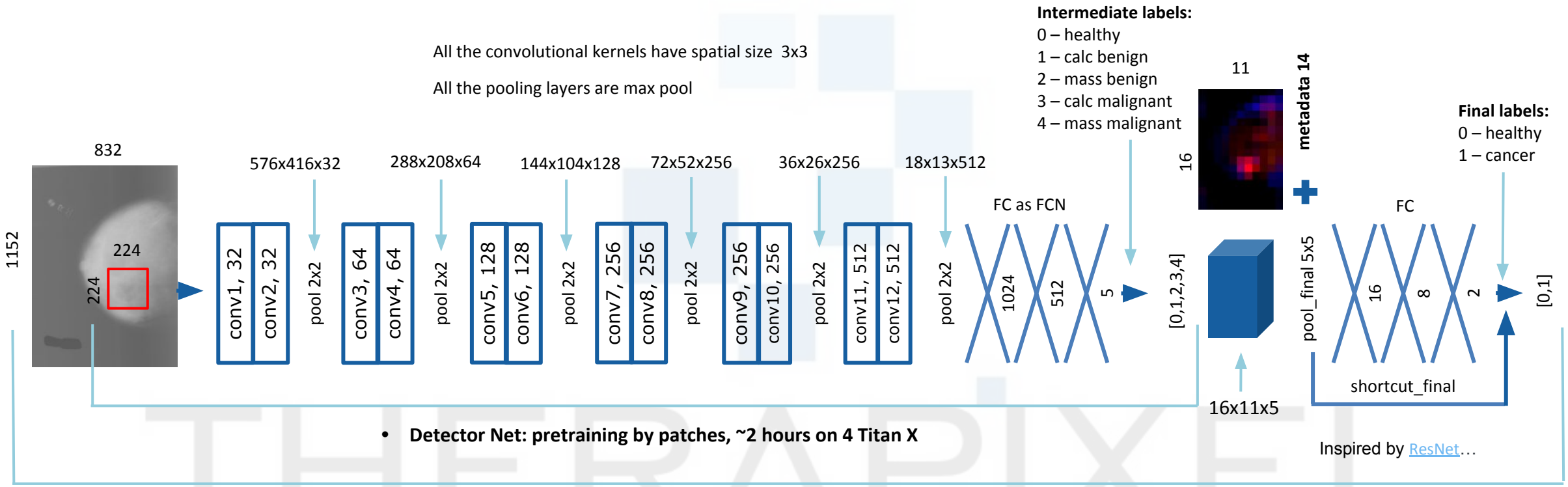
Patch model: Fully Convolutional Network



Patch model: Fully Convolutional Network



From patch to image model: final pooling and some more layers



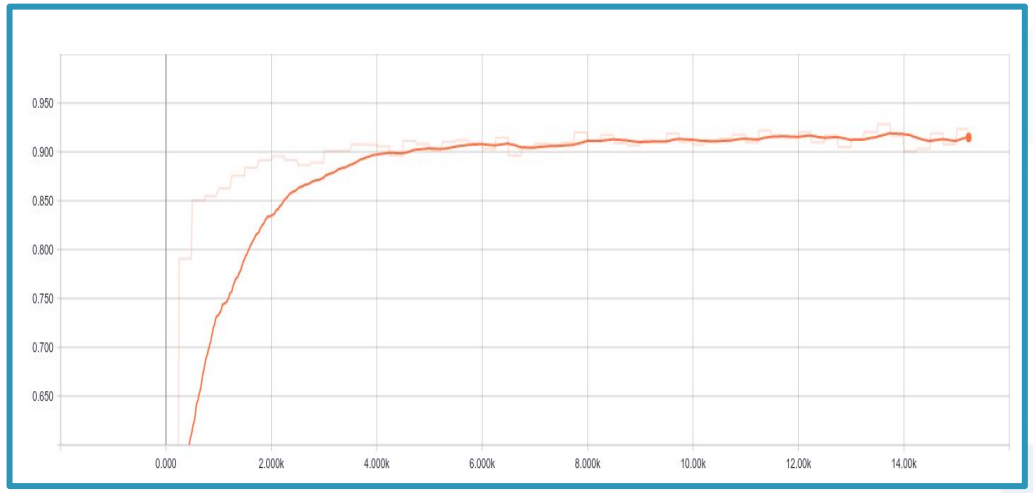
- **Detector Net: pretraining by patches, ~2 hours on 4 Titan X**
- **End-to-end finetuning by images, ~20 hours on 4 Titan X**

Important to train on images:

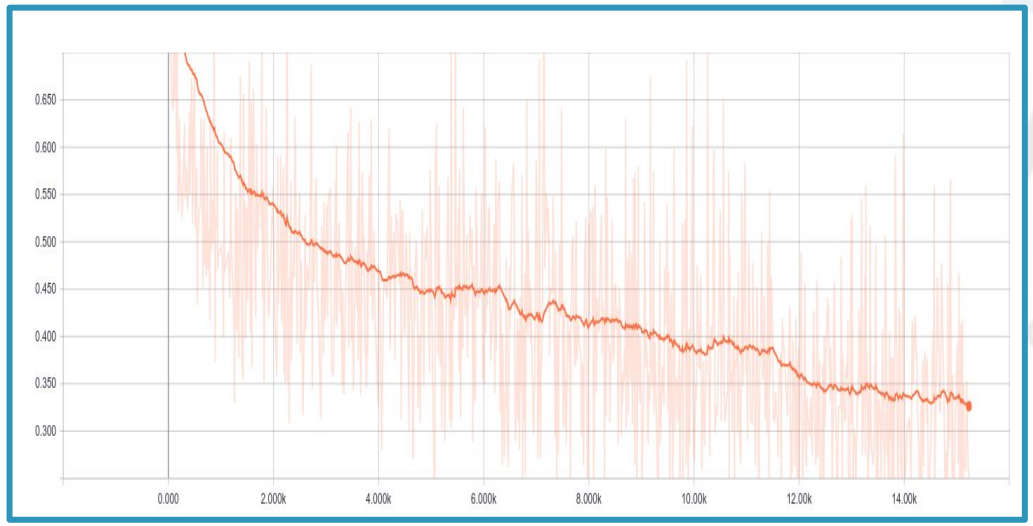
- Final pool 5x5
- Adjust learning rate
- Linear shortcut

Some technical details: training procedure and EMA

AUC per breast (DDSM)



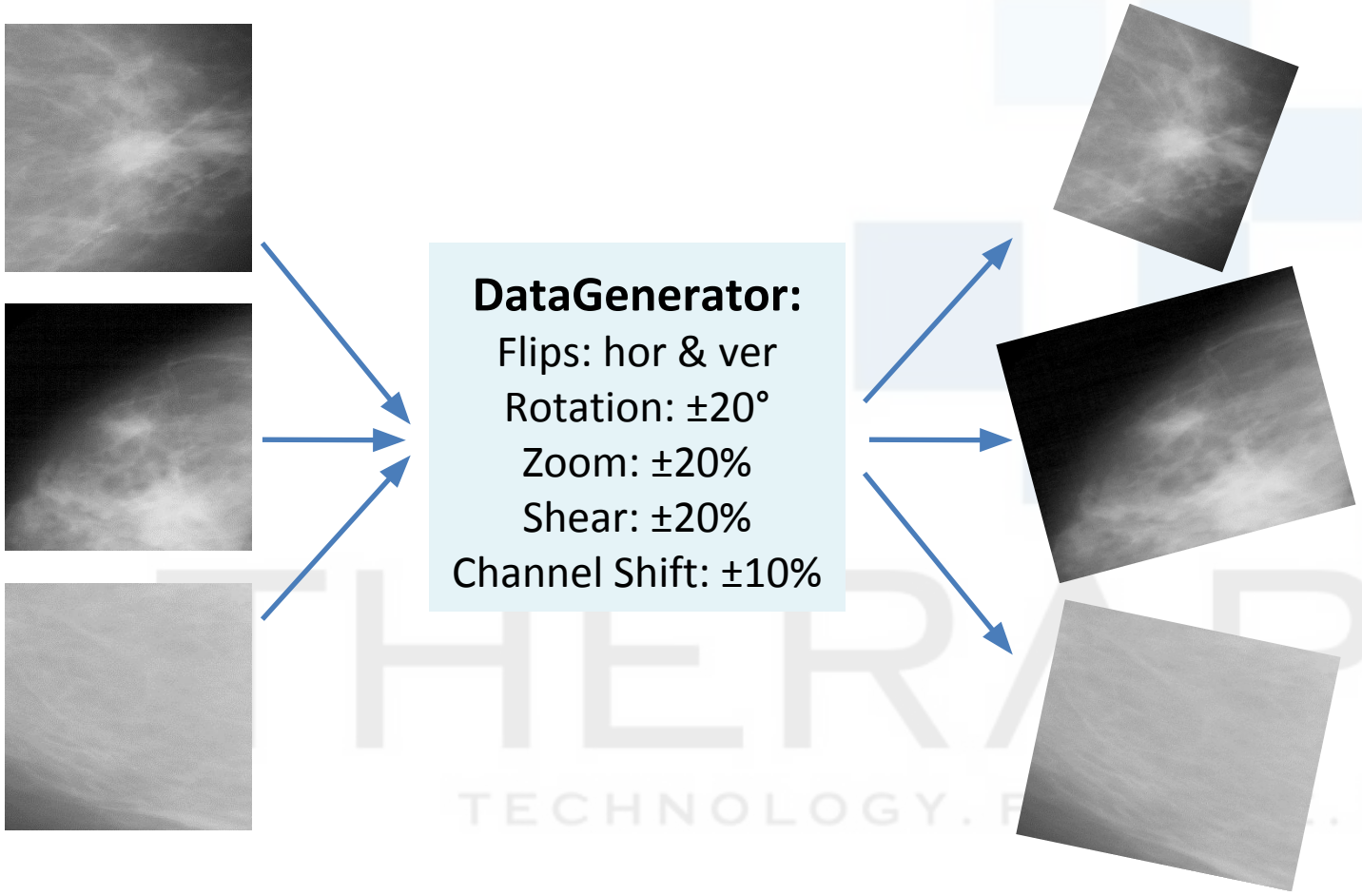
Loss



default

- DetectorNet on patches from scratch: Adam, lr 0.001
- Restore DetectorNet weights and Adam variables
- On images (partially restored): Adam, lr 0.0001
- Send it to the cloud and use as a starting point
- Finetuning on DREAM data: Adam, lr 0.0001 and Exponential Moving Averages (0.9)
- Restore EMA (0.9), finetune with SGD, lr 0.0001

Some technical details: data



- Batches are balanced
- Data Augmentation is crucial
- It also helps during the inference (4 flips → ~+1%AUC)
- Averaging everything works well

From DreamNet to MammoScreen

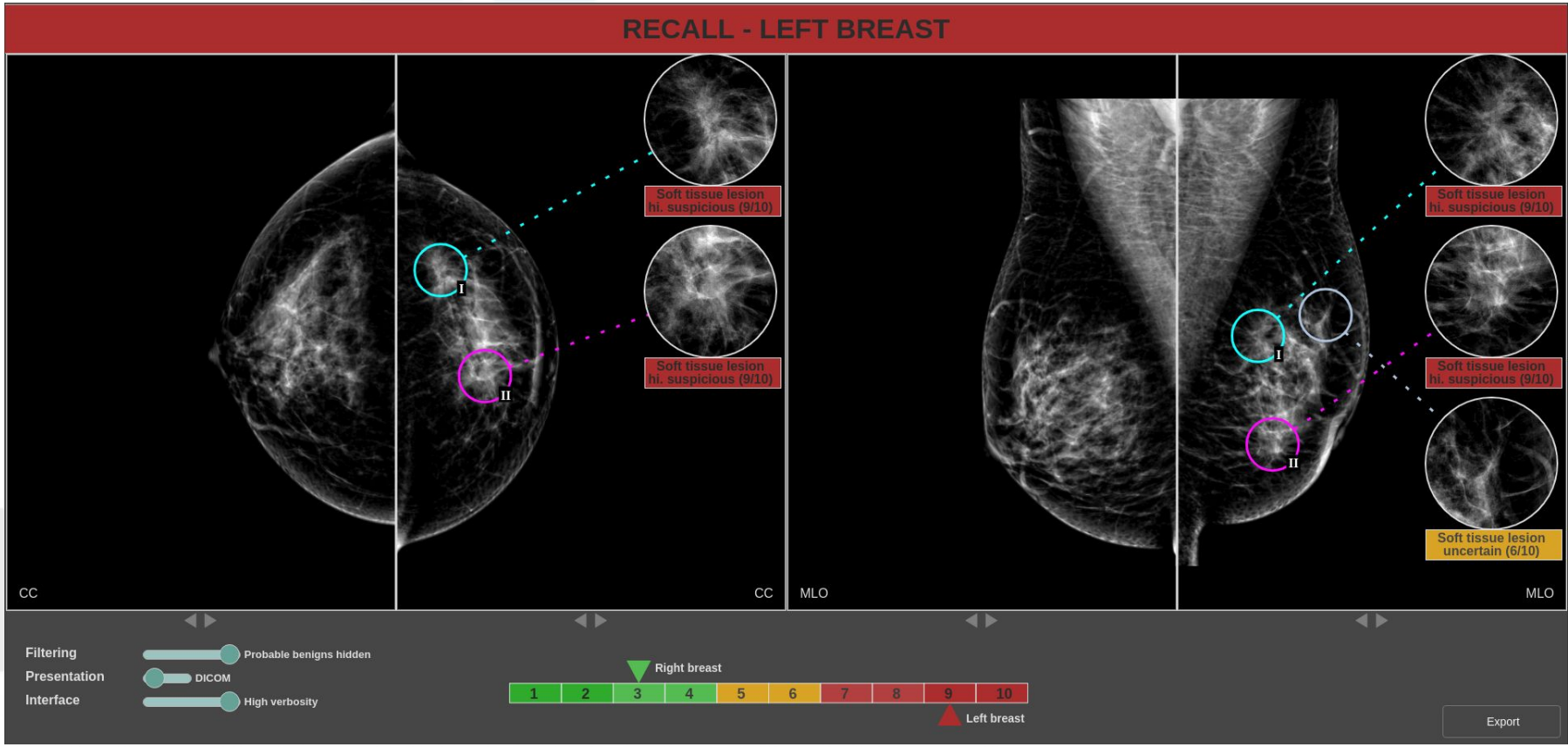
Models overview

1. 50 model instances of 5 model families (all for perf)
   1.1. Adjusted architectures
   1.2. Custom training procedures

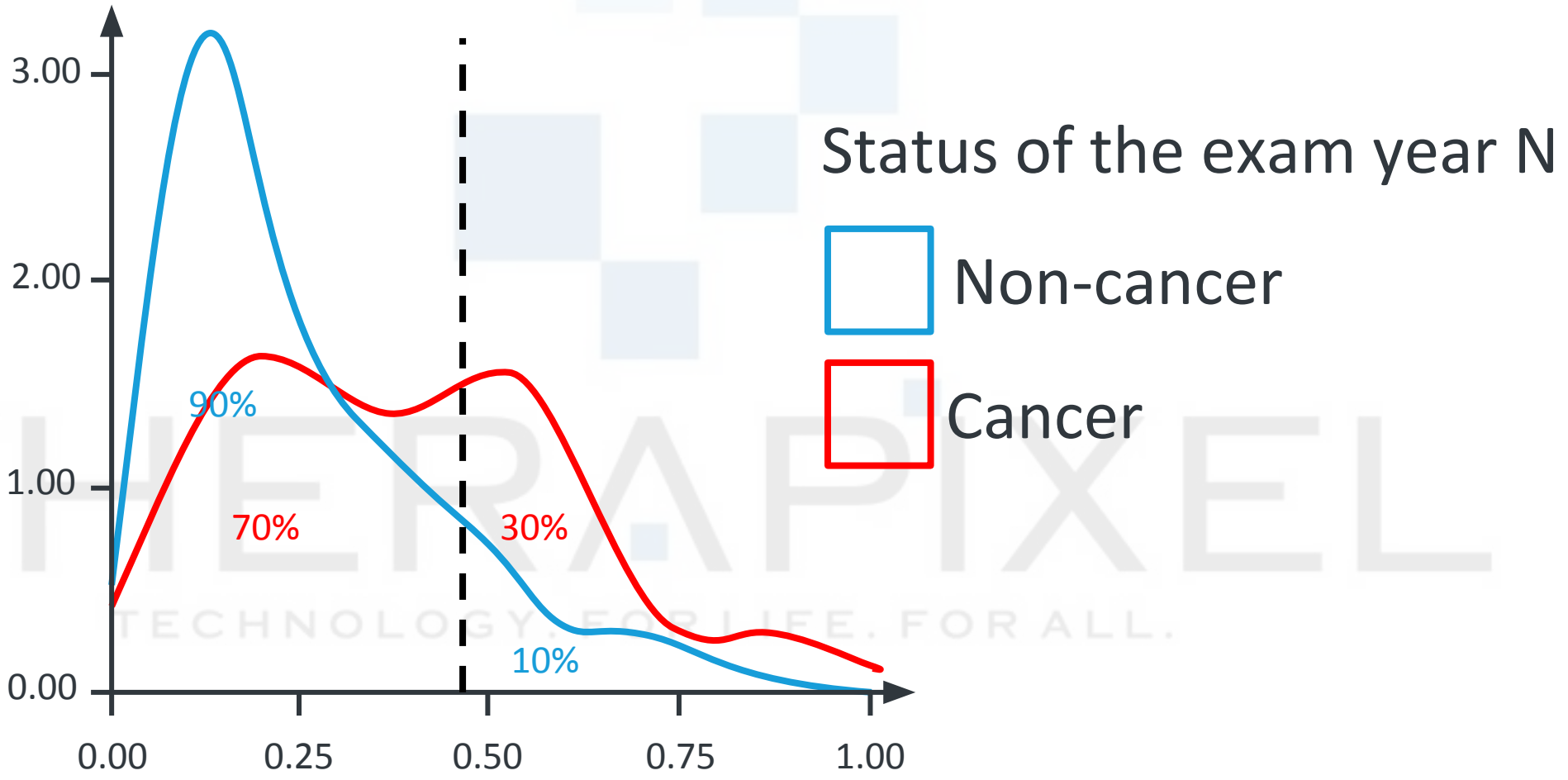
2. Lesion detection:
   2.1. RetinaNet
3. Patch malignancy score:
   3.1. CaracNet
   3.2. CaracNetSymmetry
4. Image malignancy score:
   4.1. DreamNet
   4.2. DreamNetSymmetry

5. Ensembled, calibrated

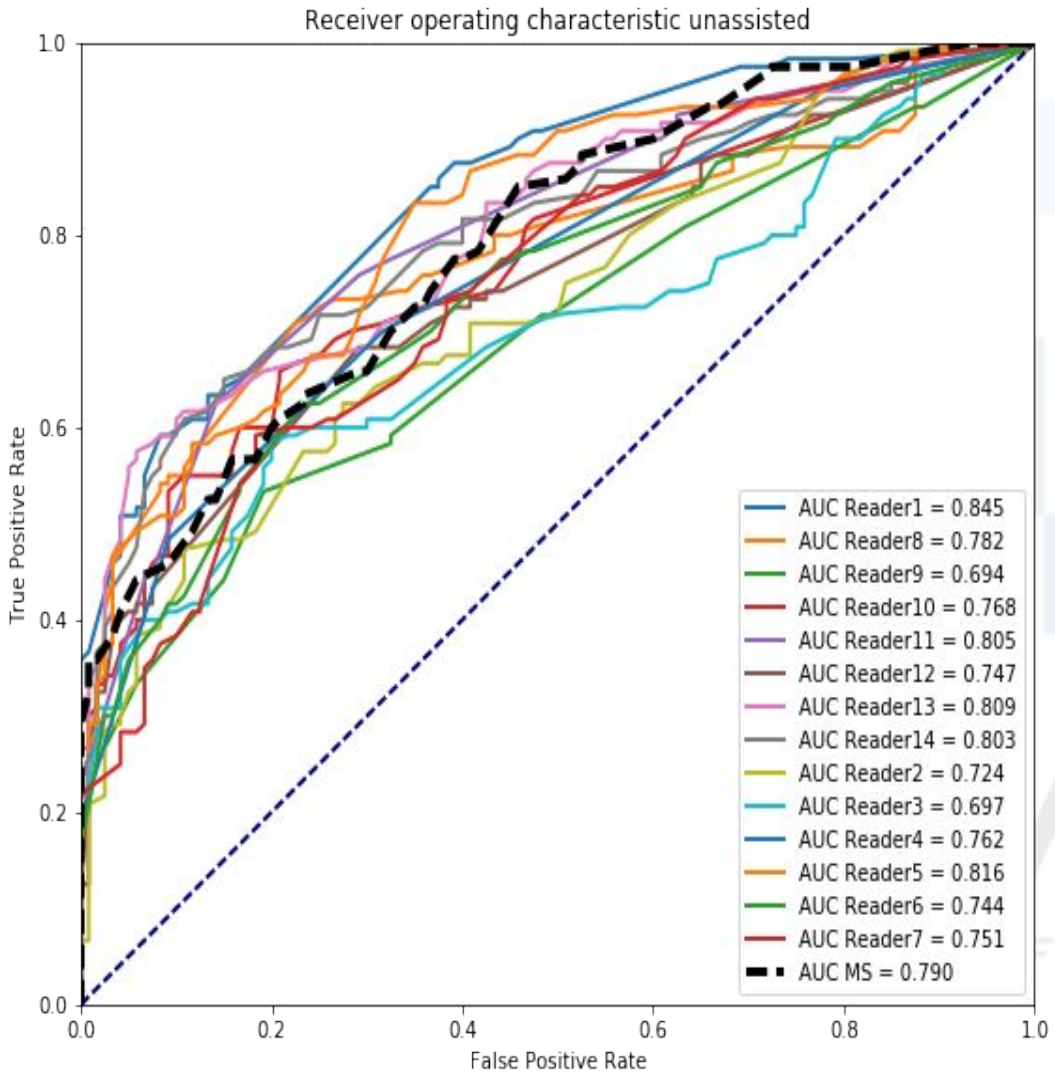
6. Lesions paired

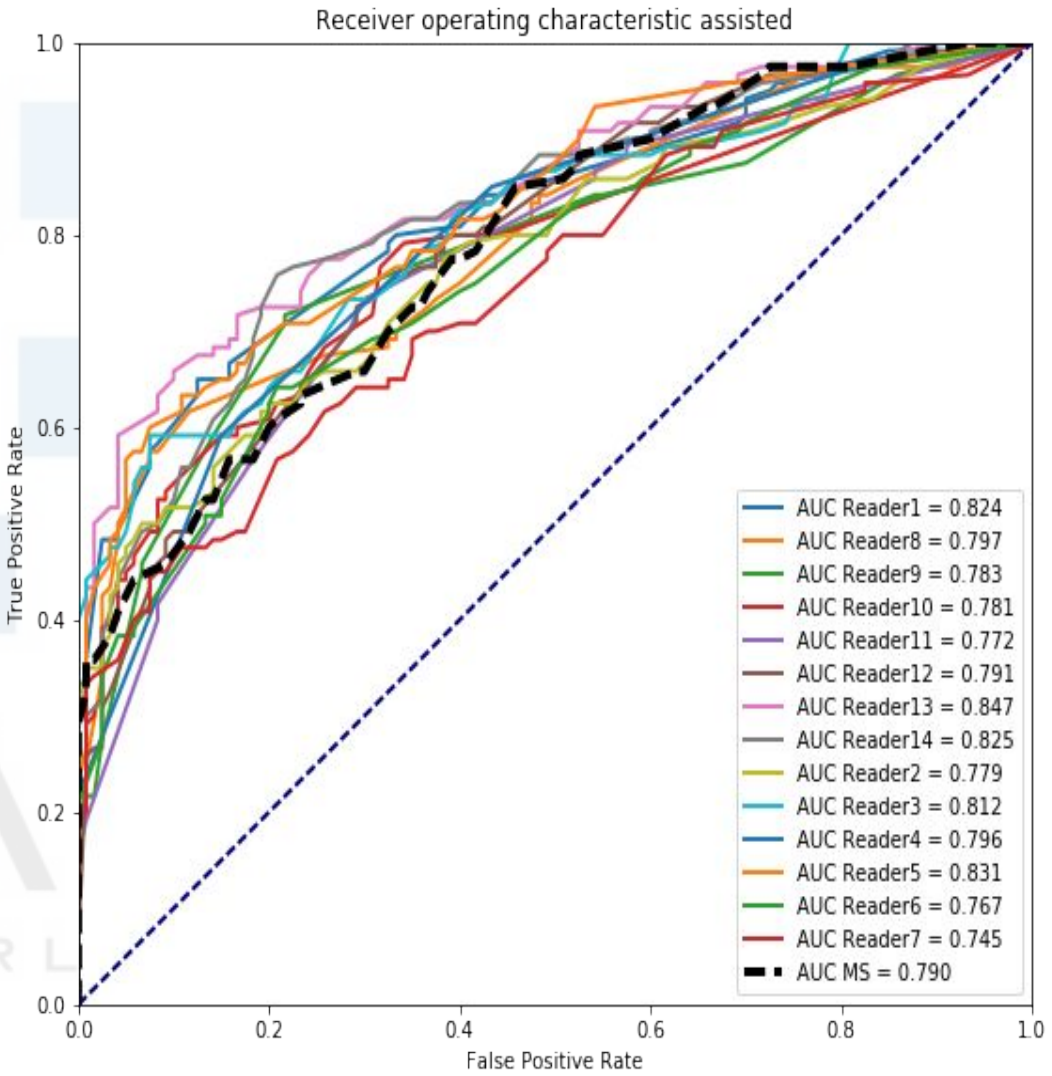


Model's output distribution on exams year N-1

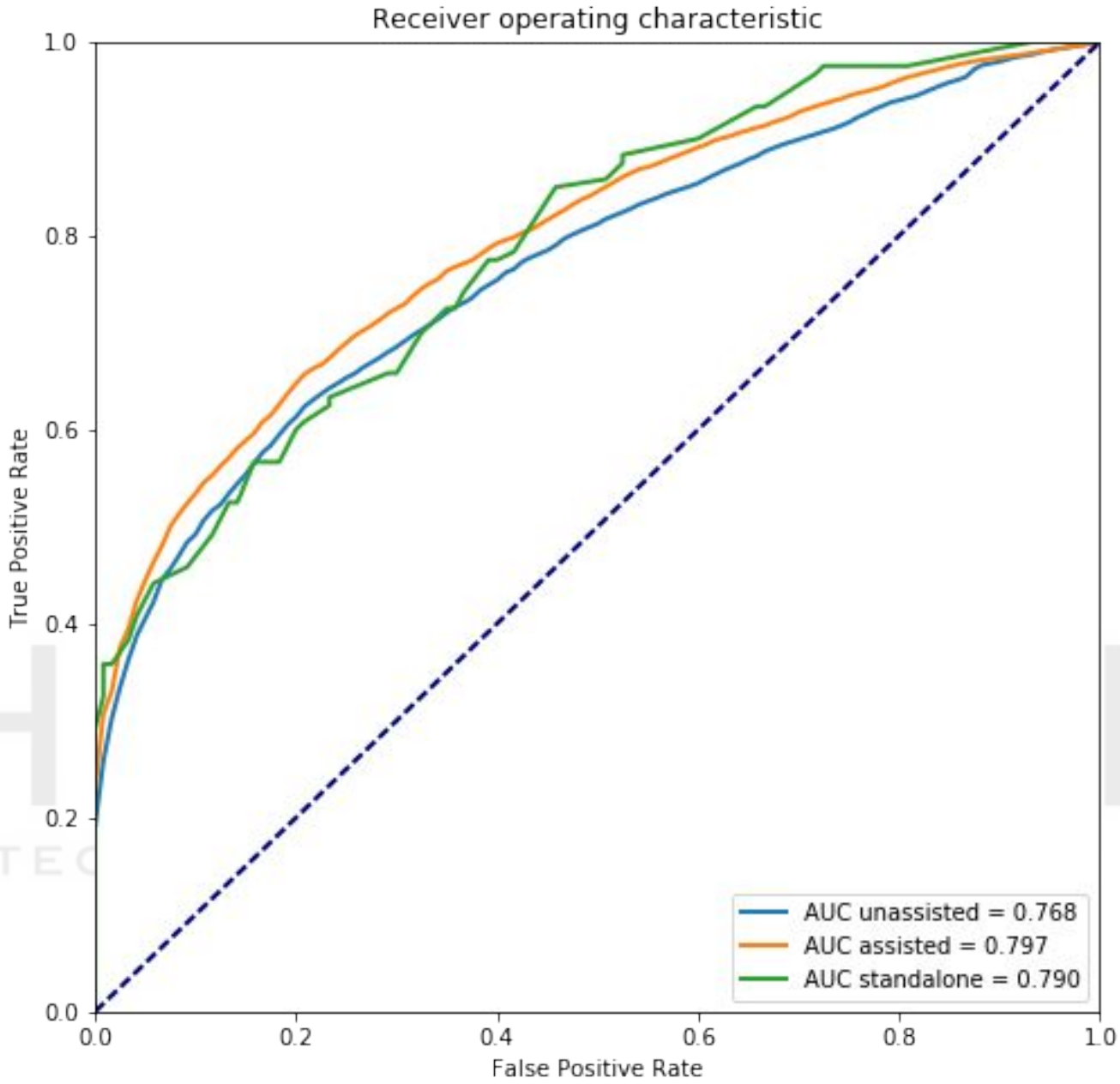


Clinical study summary

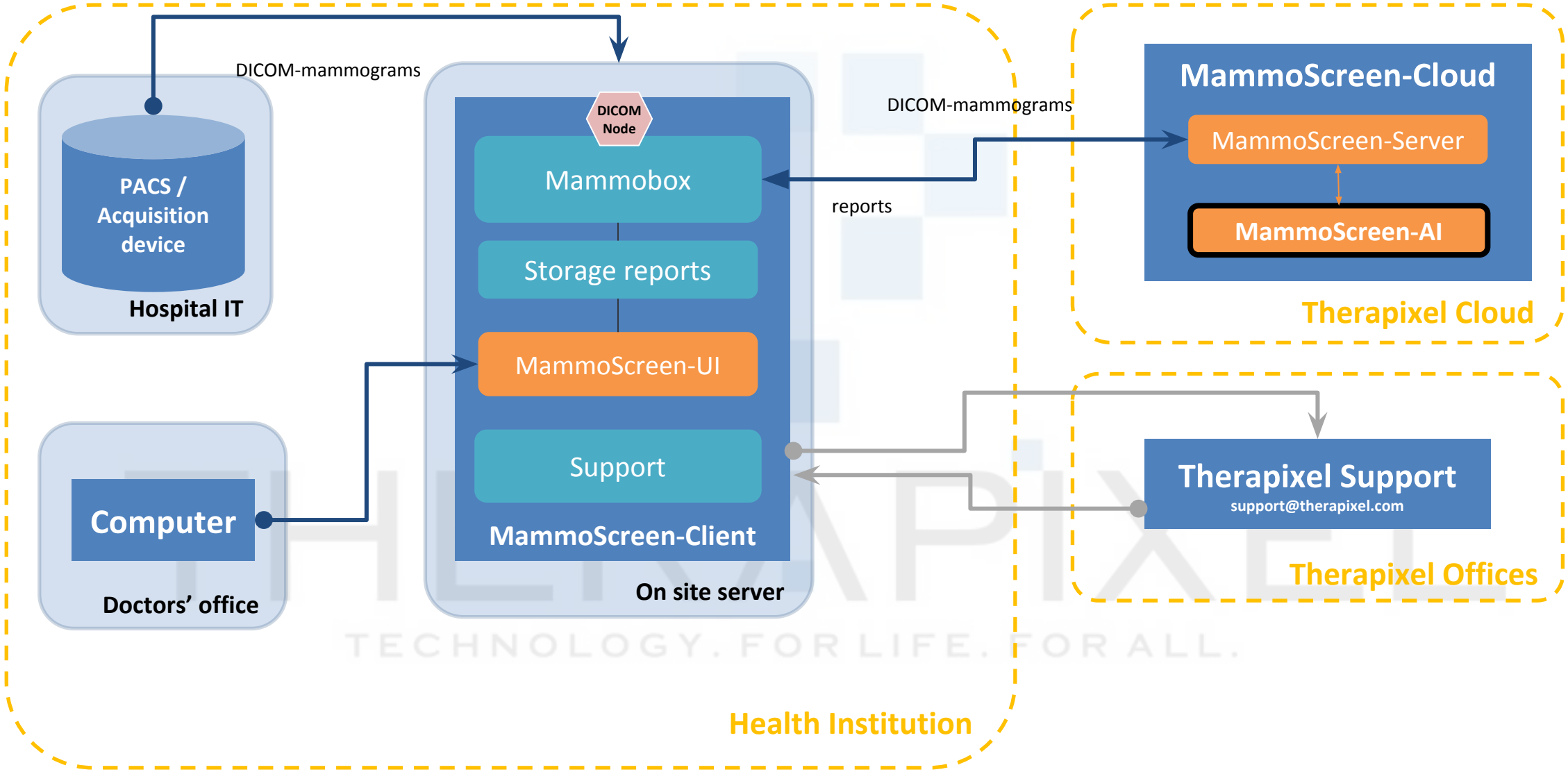




Clinical study summary



MammoScreen in Production

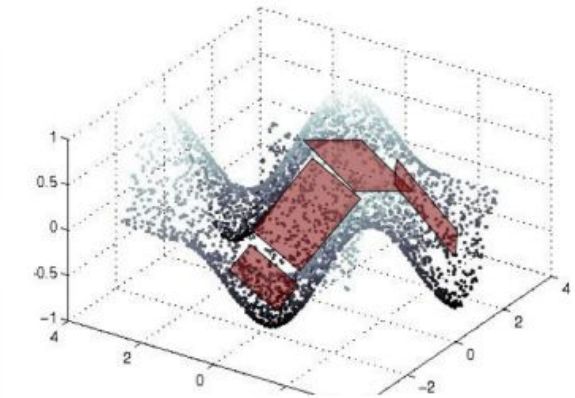


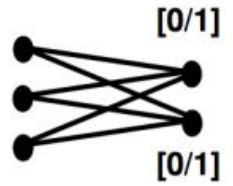
Part III: Some specific advices

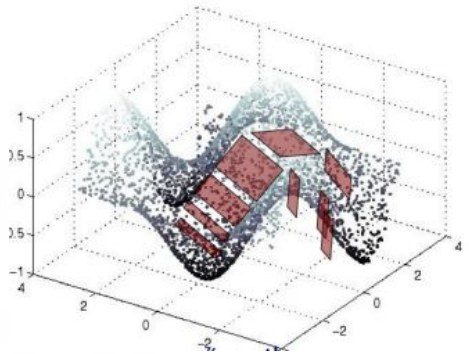
1. DL is a paradigm, not a method
2. Hyperparameters' space is too vast
3. Develop your intuition
4. Understand the internal dynamics of NN

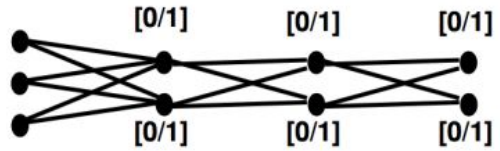
Some specific advices and practical moments

- Adapt model to your problem
- good data and gradient flow: “well-wired net”
- Adjust architecture !
- Deep = complex, but cheap









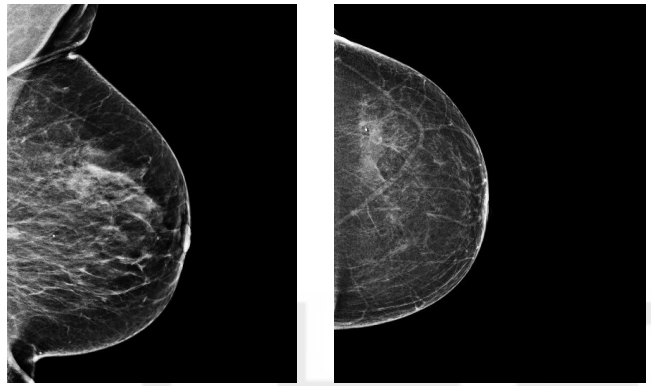
Slide credit: 1)G. Montúfar et al, On the Number of Linear Regions of Deep Neural Networks 2) Marc'Aurelio Ranzato slides 3) Introduction to Deep Learning by Iasonas Kokkinos

Neural nets are good at NOT learning the right problem

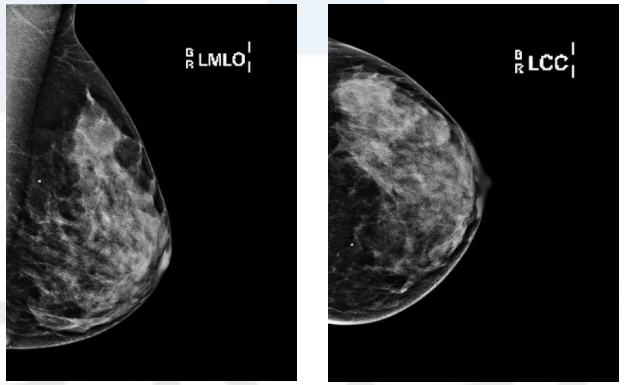
Observation:

If the networks find a latent variable in your data that correlates well with their optimization problem, they will use it. It might not be what you want!

Initial train set
(too few malignant images)



Enriching this train set
with malignant images

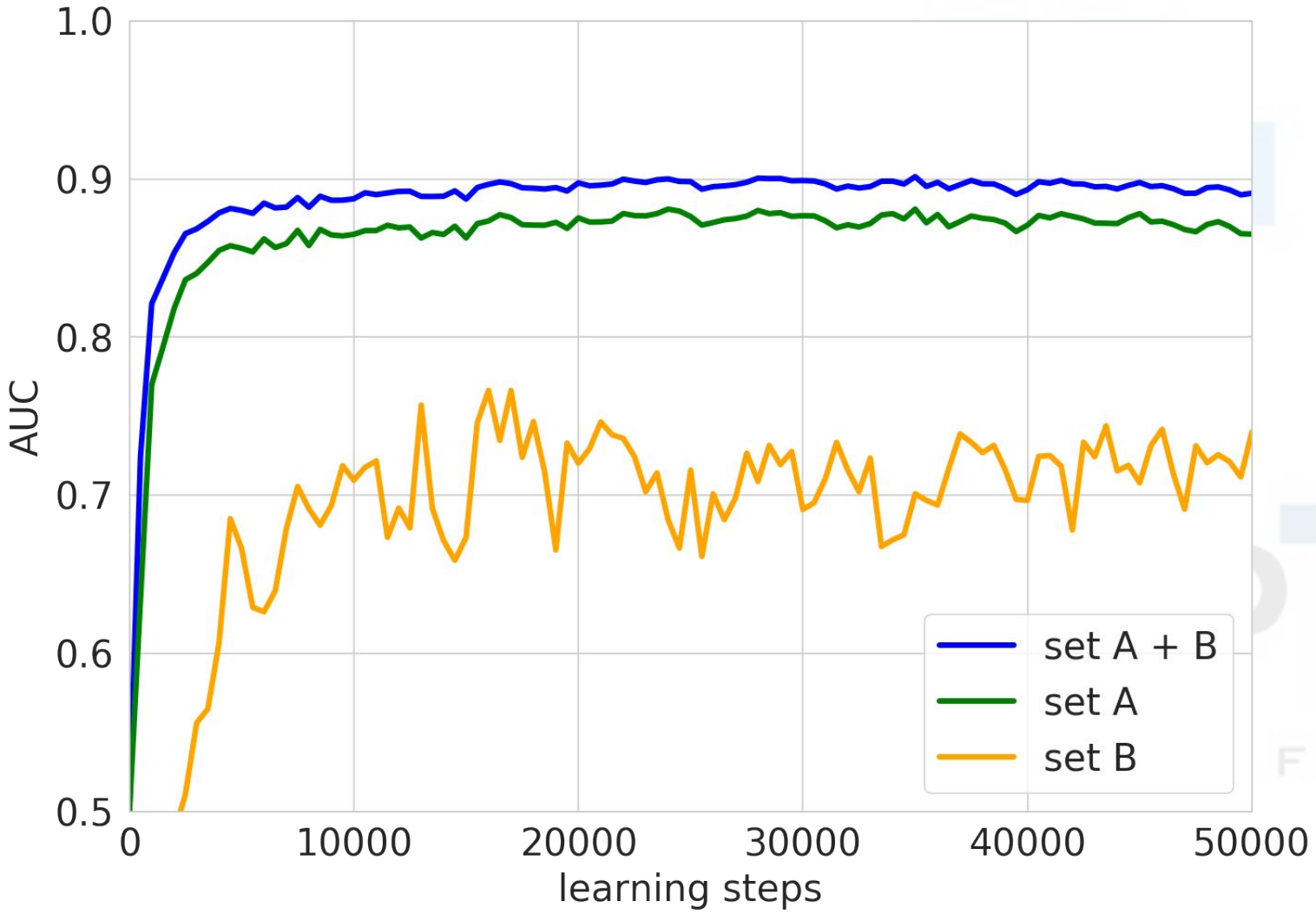


⇒ what do you think the network did?

Message:

Non-obvious biases in training data may be exploited by the network. You need to realize and control that effect. In particular, be suspicious about skyrocketing performances when injecting new source of data.

Look at the right metrics… on the right data



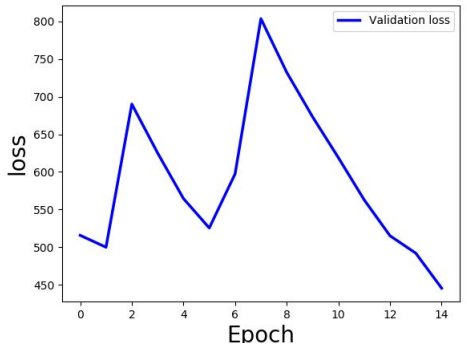
Problem ?

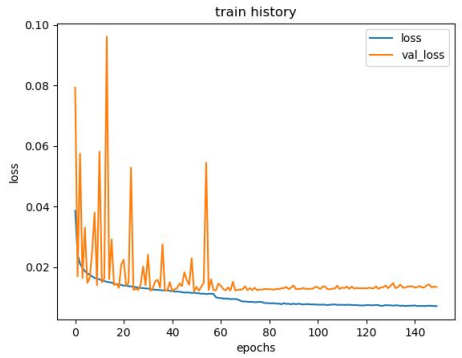
- Inhomogeneous set of validation data

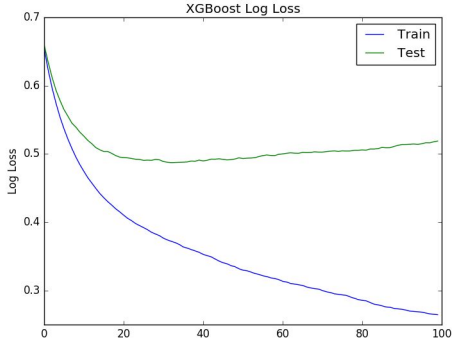
Solution ?

- Split the validation set in homogenous subsets

Fight overfitting

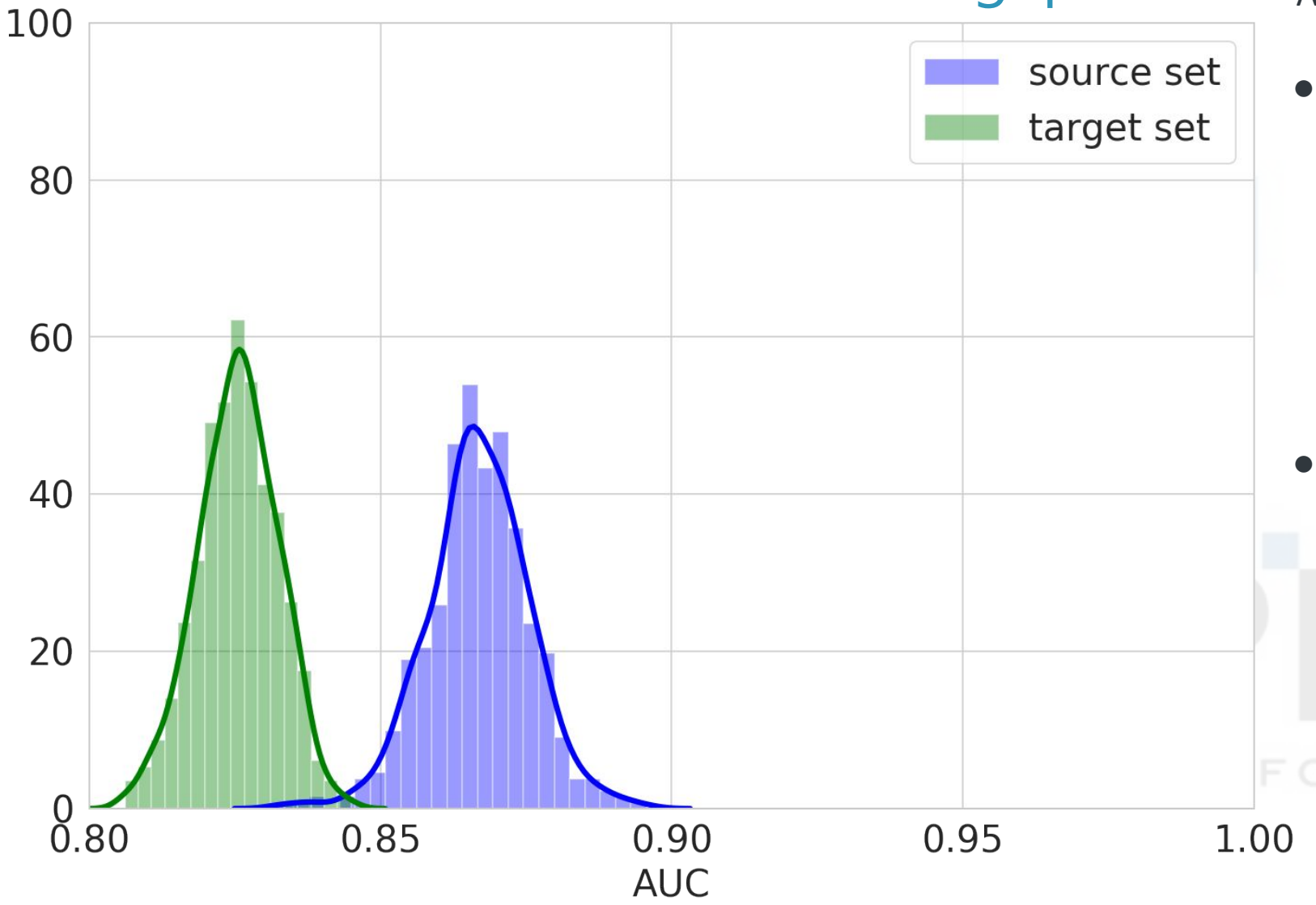






- Prefer smooth, stable learning inside a certain L2 sphere
- Good condition number
- Long plateau of low valid metric
- (Very) noisy, spiky, non-stable = bad
- How to improve:
  - more examples
  - data augmentation
  - adapt model, regularization, loss
  - validate more often
  - debug model, data
- And only then “early stopping” - checkpoint with best valid metric

Mind the generalization gap

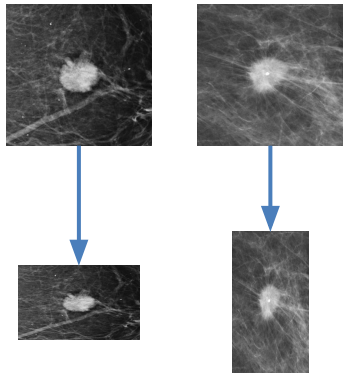


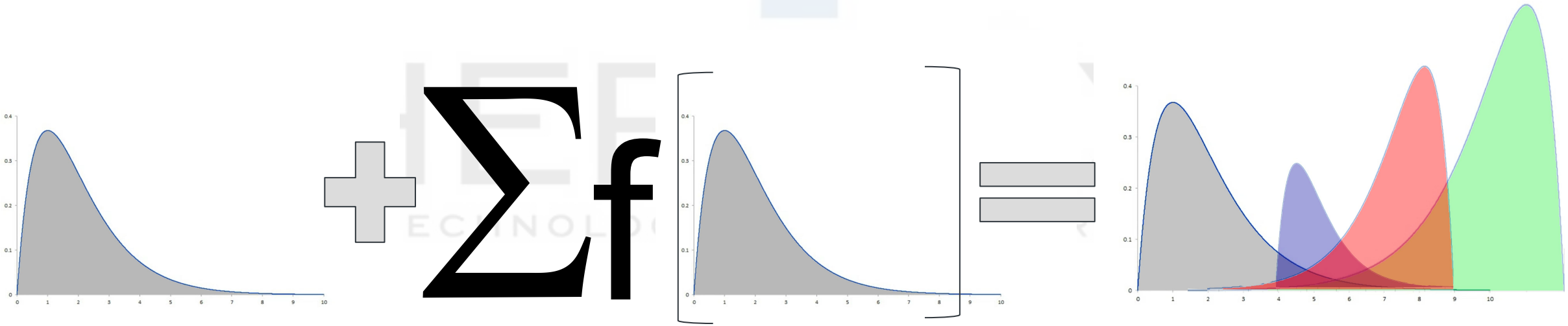
A note on overfitting and “advertising” stats:

- Overfitting happens on several levels:
  - training data
  - validation data
  - test data = overfit dataset
  - overfit a particular problem
  - overfit a particular domain (?)
  - overfit human style of thinking (??)
- In particular, performance of DL model on mammographies depends on:
  - Device used for mammography
  - Skills of technician
  - Screening period (1-1.5-2 years)
  - Positive/negative ratio, closely linked to
  - Fraction of truly difficult cases
  - Population (country)
  - …

Standardize the input distribution

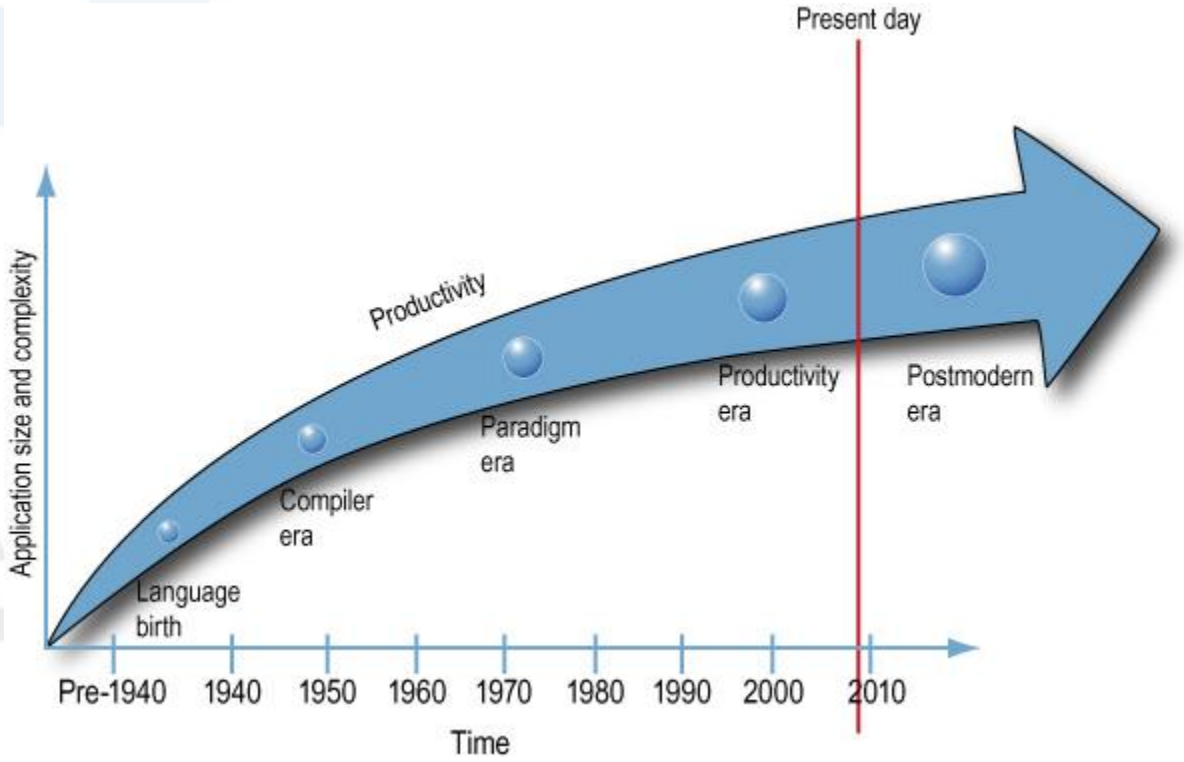
- All breast lesions in the world form a (very complex) distribution
- Eliminate everything not linked to natural variability
- Example: device1 -> 3600x2400, device2 -> 3600x3600
- When both image sets rescaled to 1200x800 -> 2 modes





Data Science 2020 = Software Engineering 2000

- Visual Studio 1st release: 1997
- Development process and paradigm evolving
- Data becomes 2nd part of your code
- Software 2.0 stack (©Andrej Karpathy)
- IDEs for ML models are yet to come?
  TensorFlow Extended (TFX)?



Thank you for your attention!

Q&A session