MATHÉMATIQUES
VISION
APPRENTISSAGE

Special for MVA course

Deep Learning in practice: MammoScreen

Therapixel

Yaroslav Nikulin,
Senior Research Scientist

January, 21 2019

Part I: Introduction

- Therapixel
- DL -> radiology
- Breast cancer
- DM DREAM Challenge

Therapixel: Medical Image Understanding

Therapixel

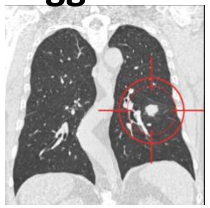
2019 Clinical Study comparing MammoScreen to radiologists

2018 Mammoscreen

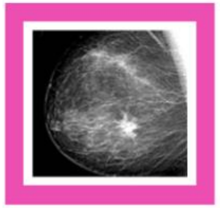
2017 1st place DREAM DM
5th place Kaggle Data Bowl

2016 AI research

Kaggle Bowl



DREAM DM



2015 Visualization SW

2013 Founded

Olivier Clatz, PhD

inria inside

Pierre Fillard, PhD

Breast Cancer Screening: some key stats

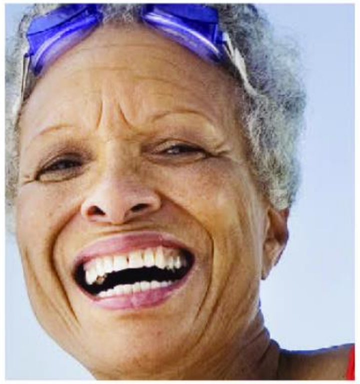
- 33M exams/year = 132M images in US alone
- $7.8 billion - cost of mammography screening in US (2010)
- 120 sec: average interpretation time.

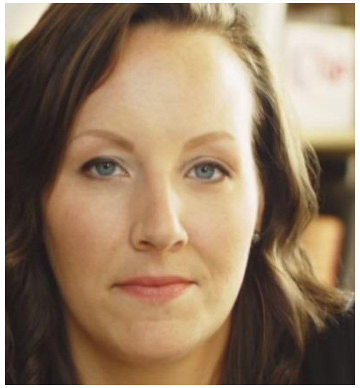
1 out of 8

Woman affected during her lifetime

10 recall for **100** screened

5 cancers for **1000** screening





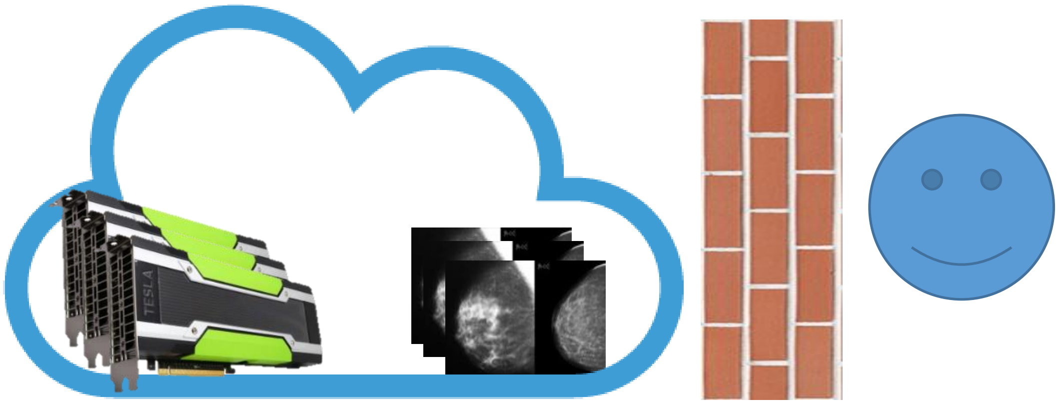
"If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future."

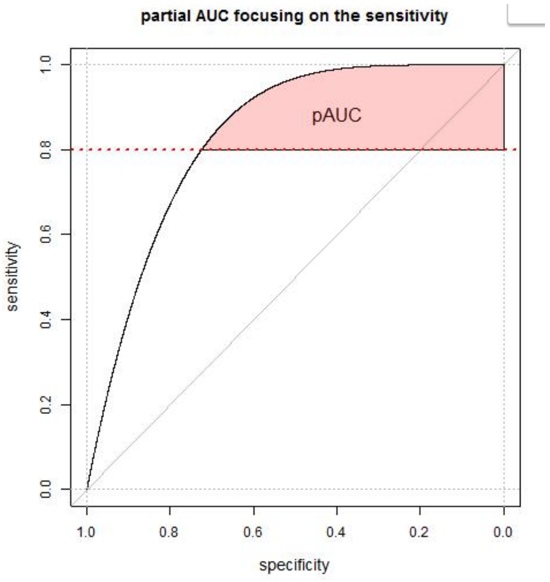
Andrew Ng, 2016

The Digital Mammography DREAM Challenge

Challenge setting:

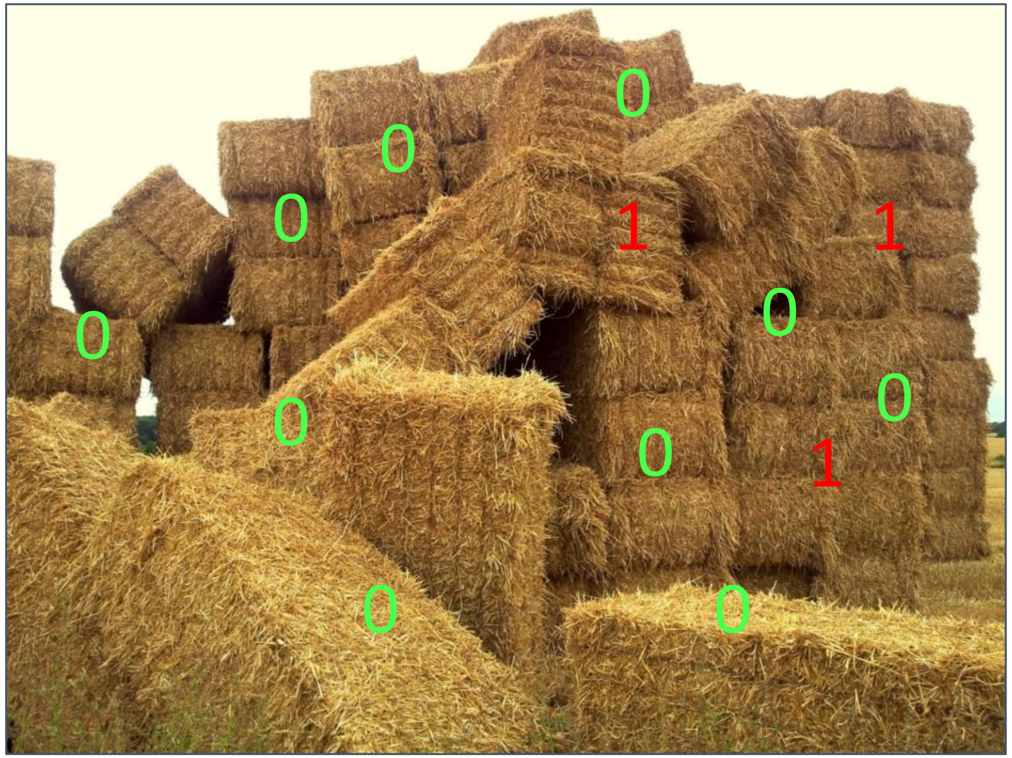
- Completely in the cloud
- 22 CPU cores + 2 GPUs
- 14 days / per team
- Performance measure: **AUC** and **partial AUC**





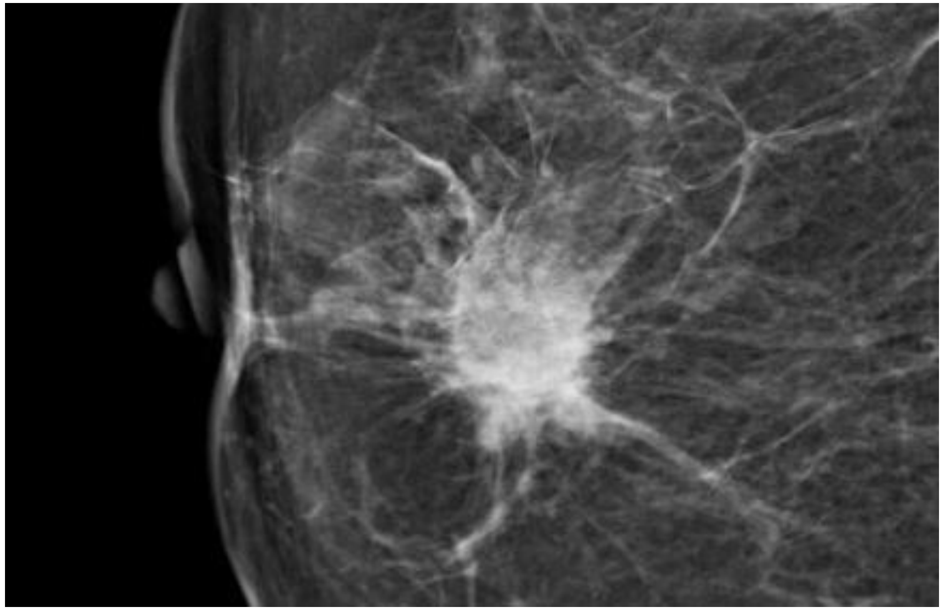
Why it is difficult - challenges of the Challenge

- 320k images
- Only 1548 (**0.47%!**) positive examples
- High resolution: from **3328x2560 to 5928x4728**.
- One single label per image: 0 or 1

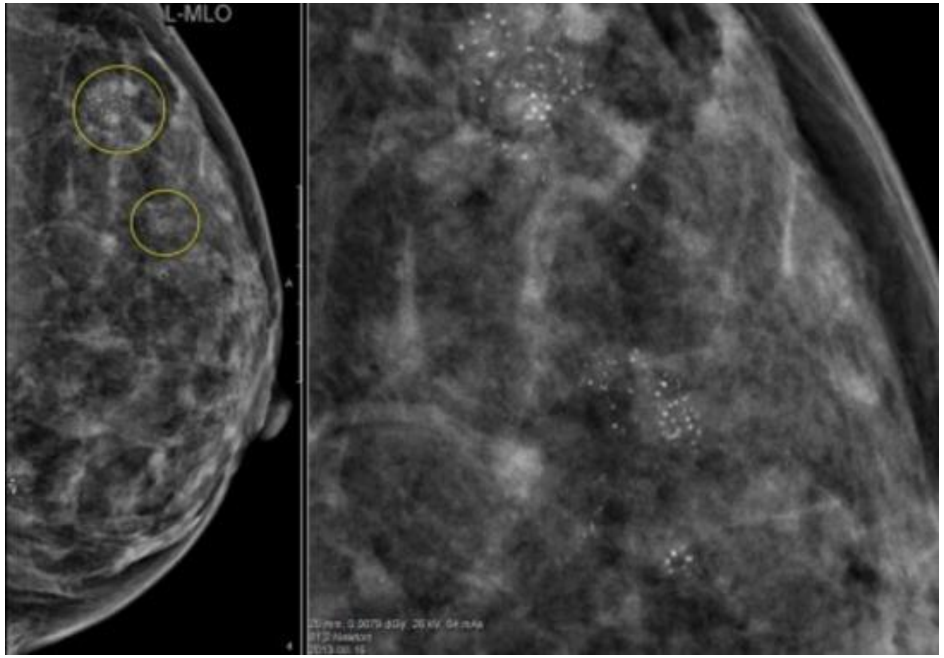


Now look for a needle in them…

Why it is difficult - challenges of the Challenge



- Different kinds of anomalies: calcifications, masses, distorsions



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

Can be malignant OR benign!

Part II: Winning solution dream_net

- Data specificity
- Dense annotations
- Patch model
- Image model
- Visualization

Why is it very different from ImageNet?

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

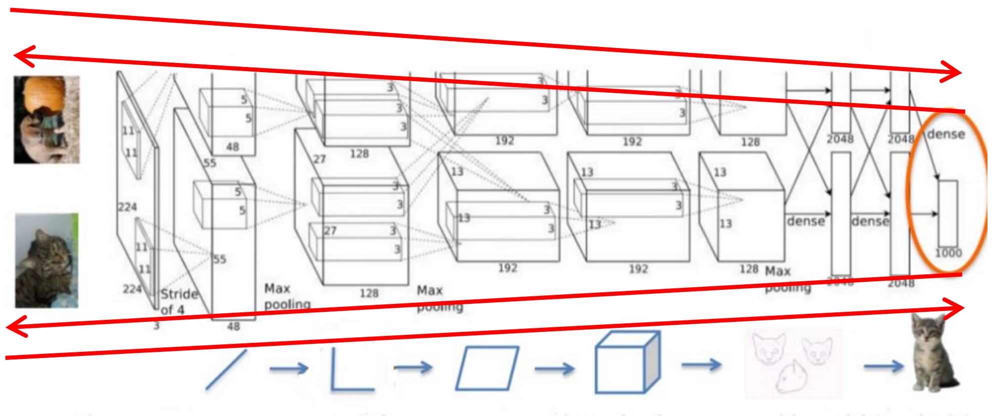
- Resolution: 1200x800 **vs** 224x224
- Zone of Interest : < 1% **vs** > 50%
- Number of classes : 2 **vs** 1000
- Highly imbalanced **vs** roughly balanced

Zone of Interest

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

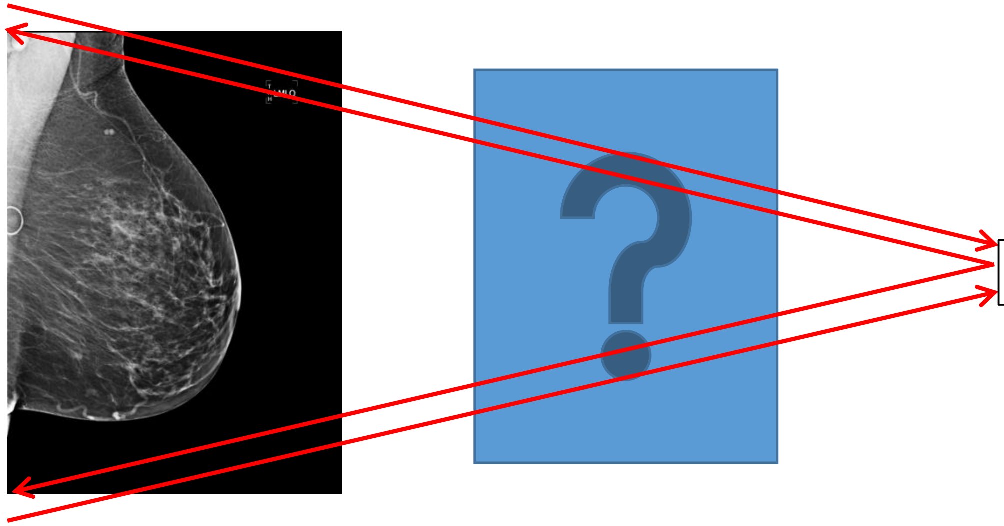
AlexNet (Krizhevsky et al. 2012)

Input size:
224x224



Output : 1 out of 1000
~ 10 bits of information

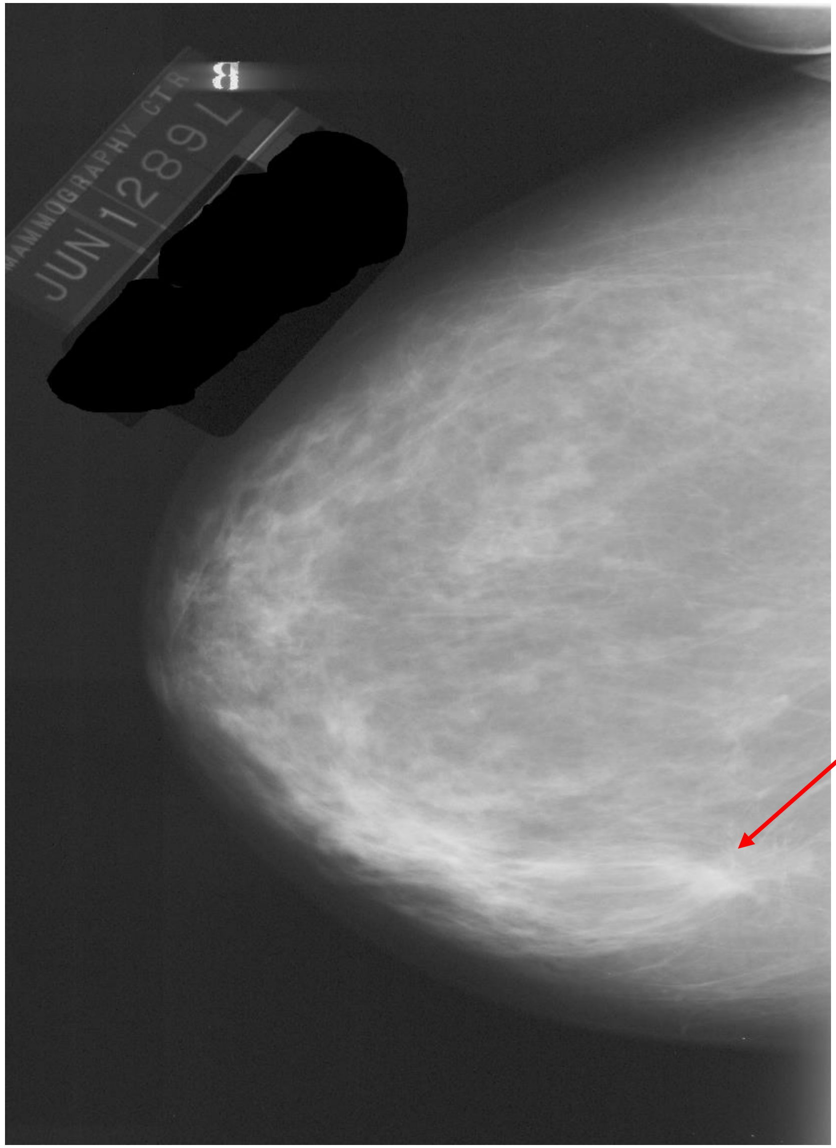
Input size:
~3500x2500



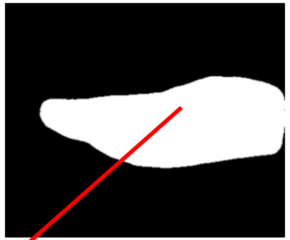
0 or 1

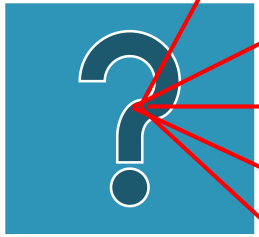
Output : 1 out of 2
= 1 bit of information

DDSM – bridge towards solution



	DDSM	DREAM
Total im	10k	320k
Positives	1807	1548
Info	mask&type	0 or 1

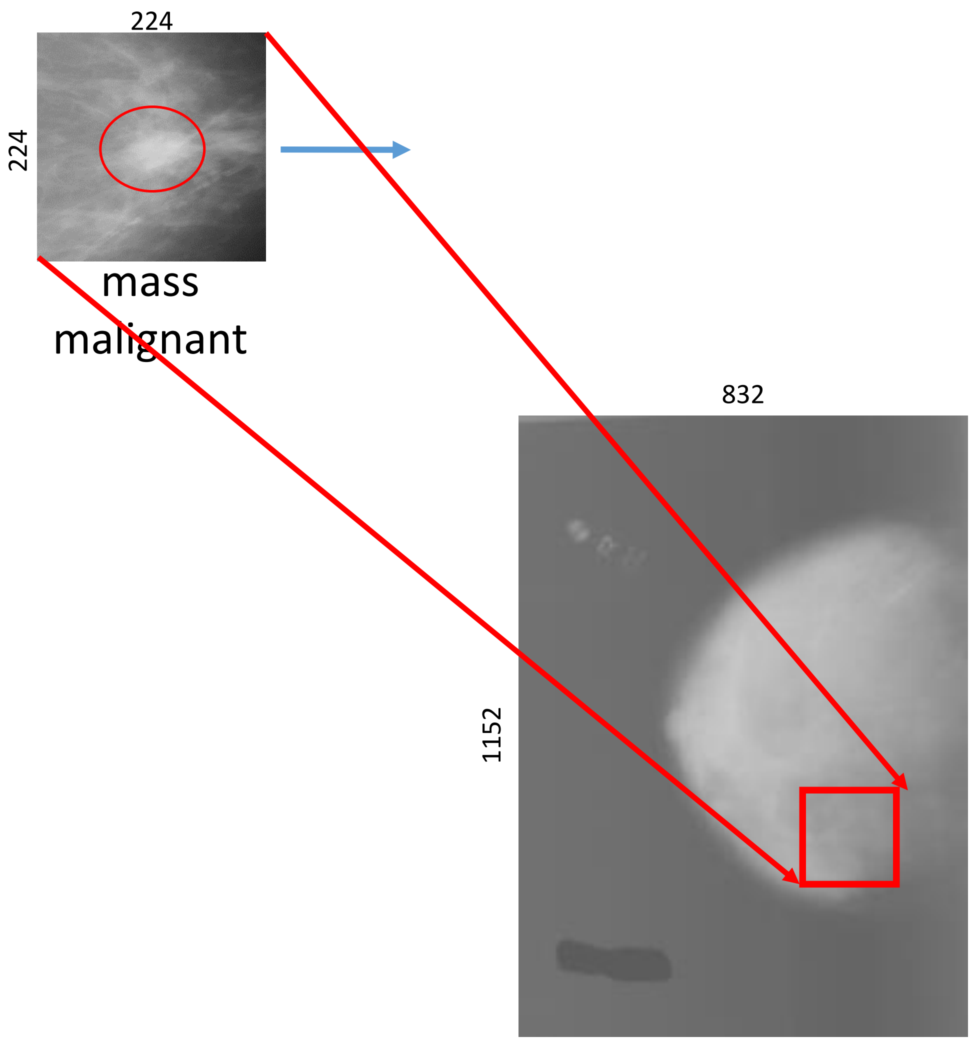




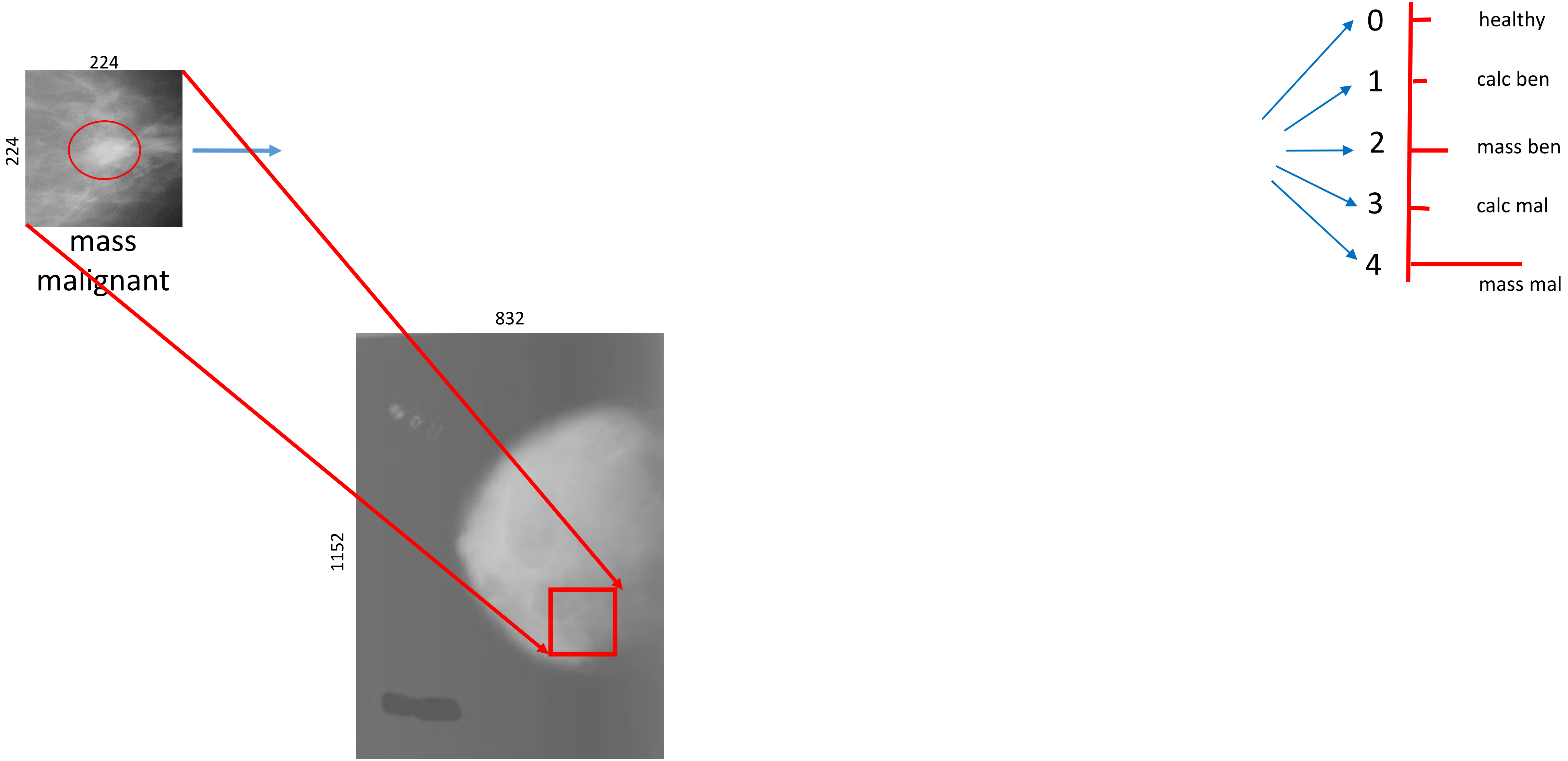
It would be great to:

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

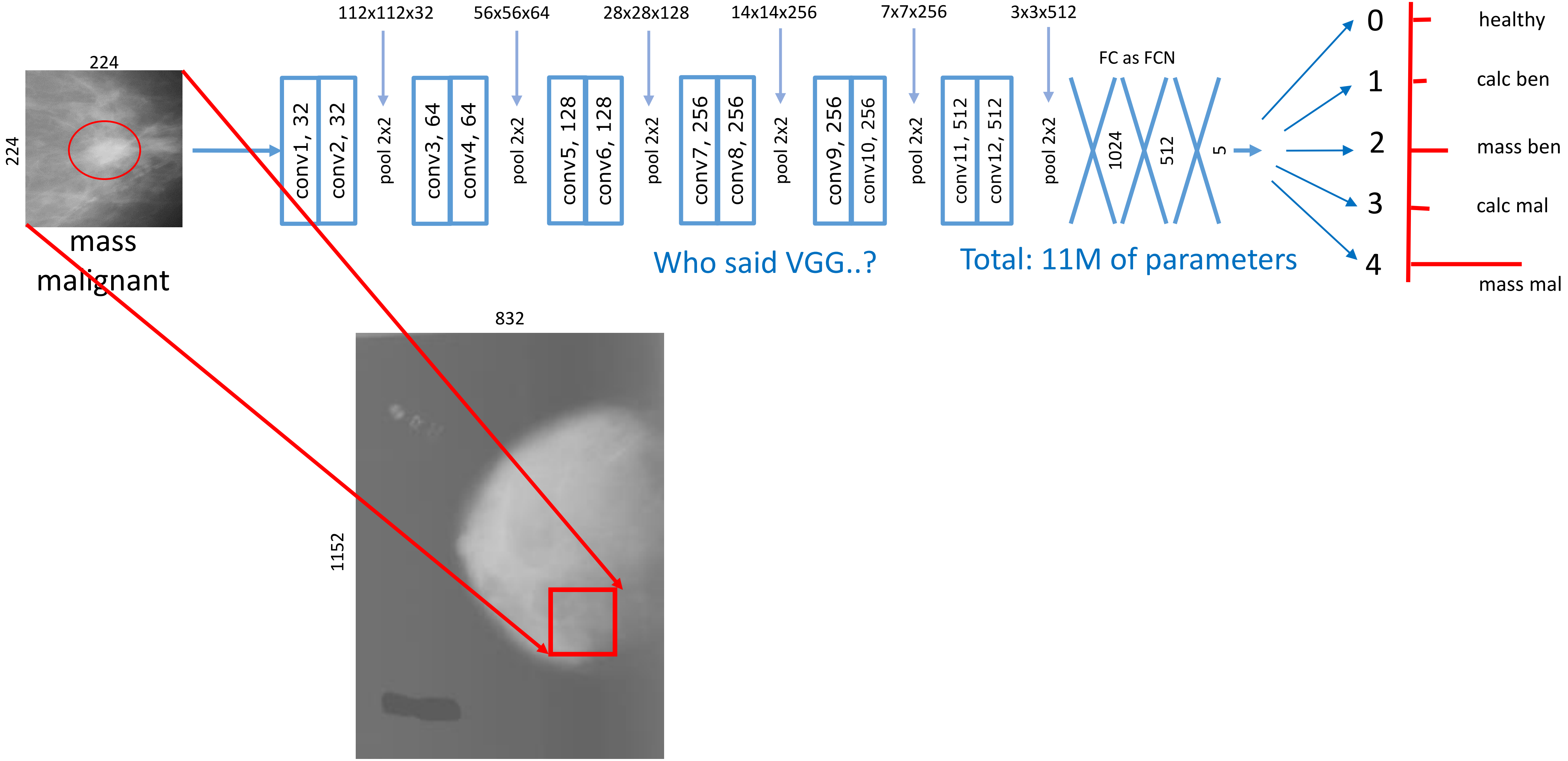
Patch model: Fully Convolutional Network



Patch model: Fully Convolutional Network



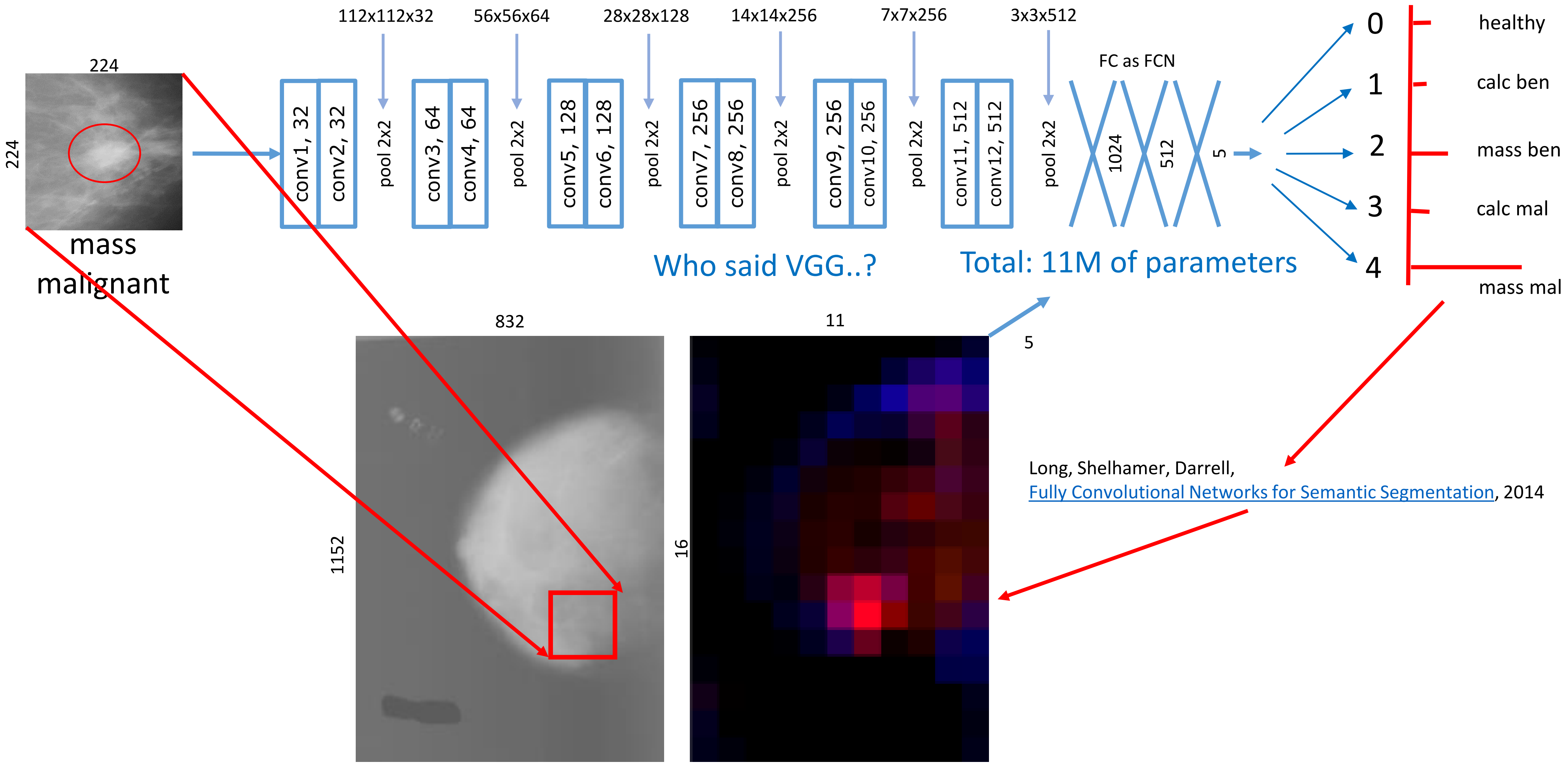
Patch model: Fully Convolutional Network



Who said VGG..?

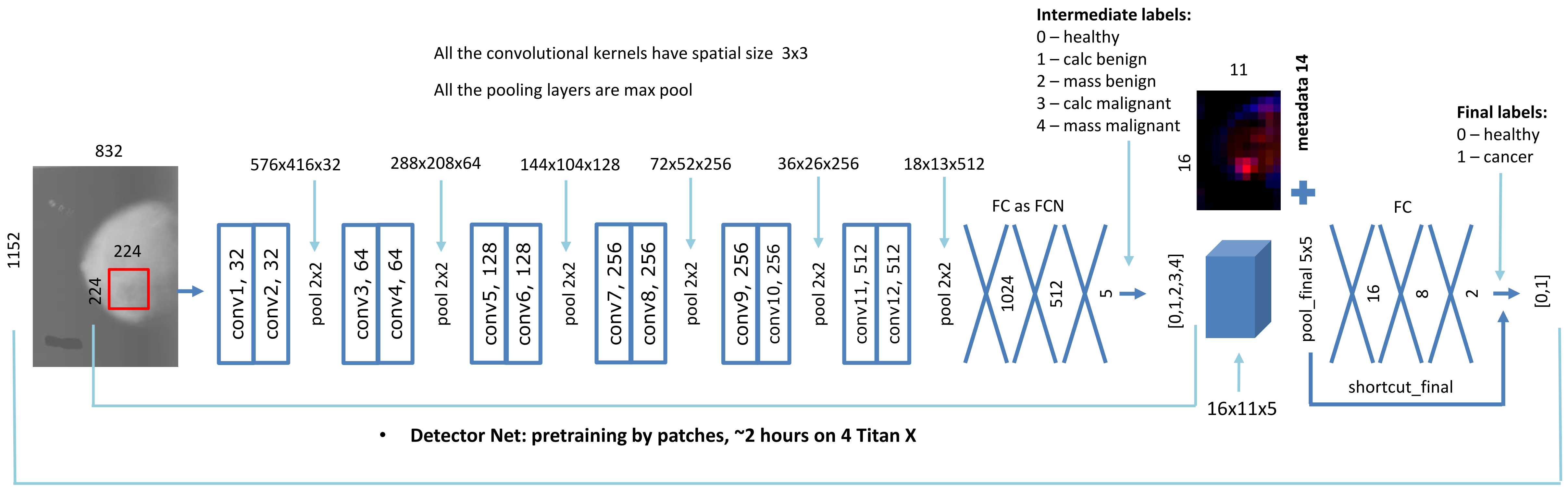
Total: 11M of parameters

Patch model: Fully Convolutional Network



Information Bottleneck..?

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



- **Detector Net: pretraining by patches, ~2 hours on 4 Titan X**

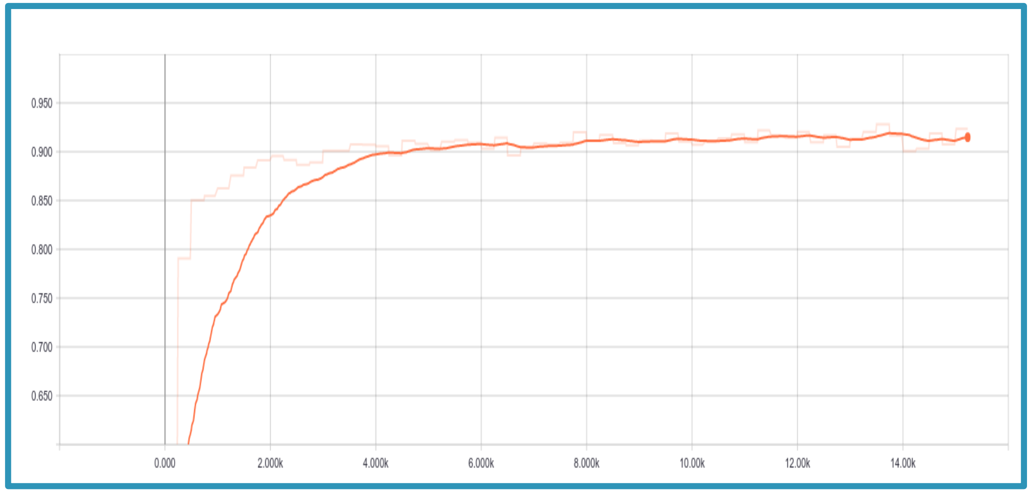
- **End-to-end finetuning by images, ~20 hours on 4 Titan X**

Important to train on images:

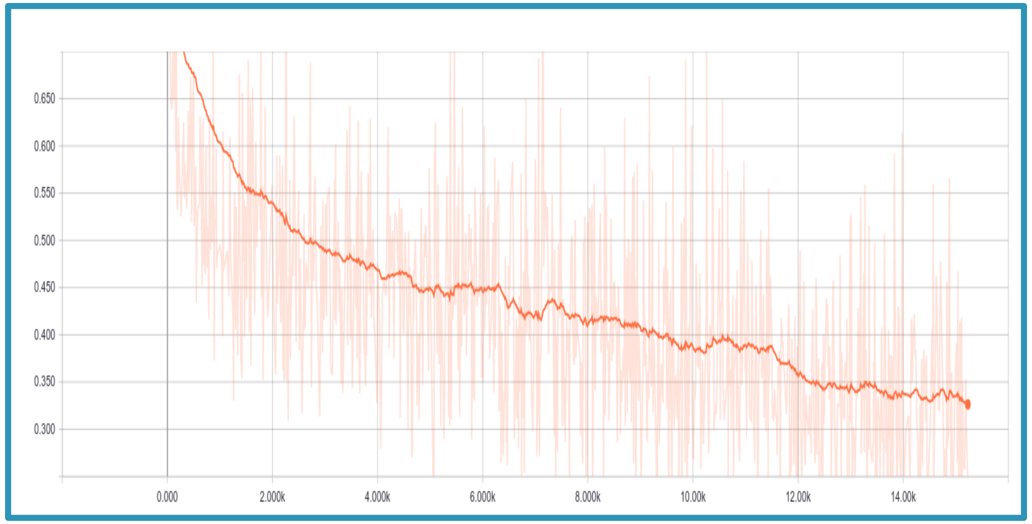
- Final pool 5x5
- Adjust learning rate
- Linear shortcut

Some technical details: training procedure and EMA

AUC per breast (DDSM)



Loss



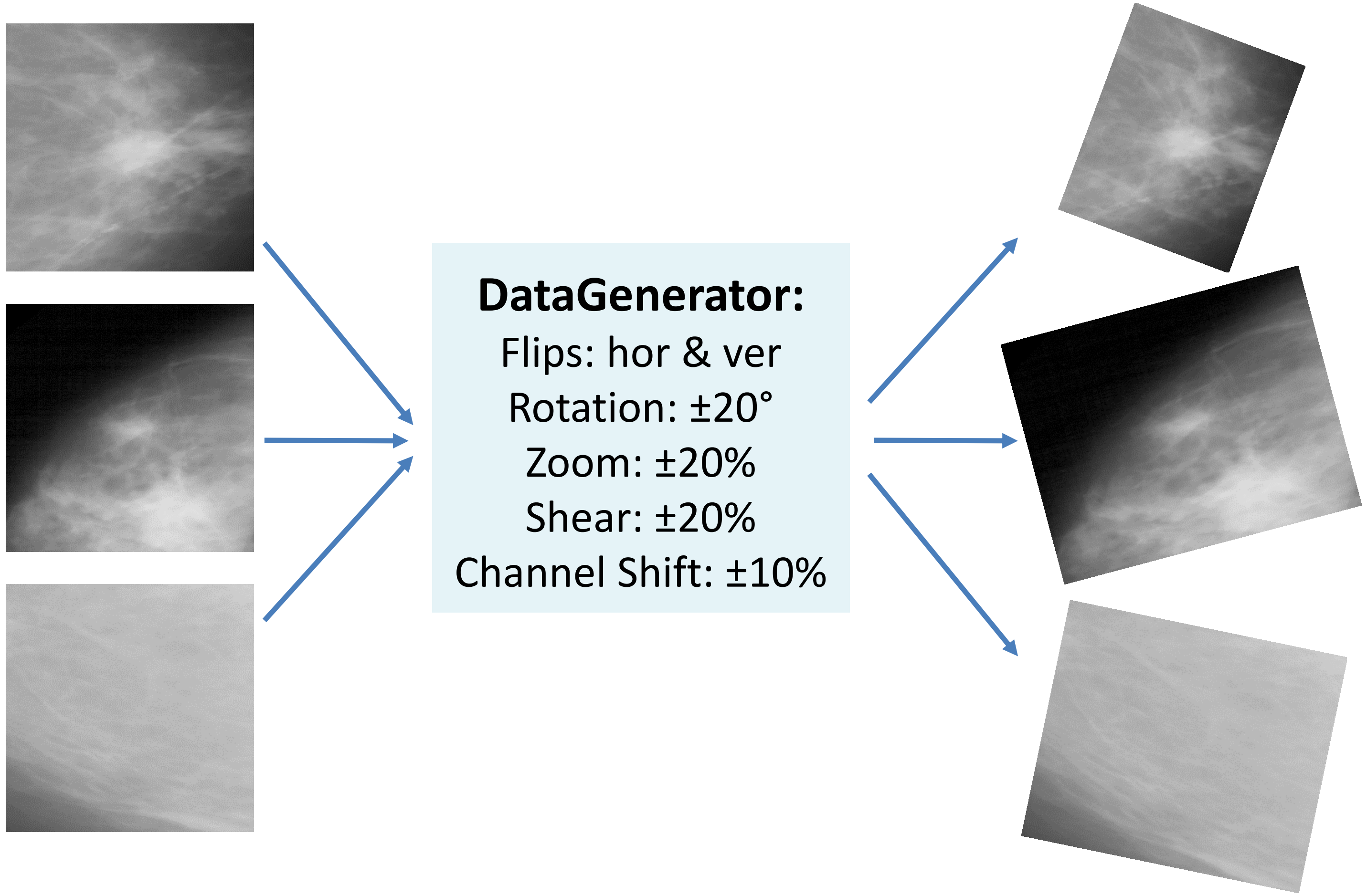
default

- 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

Why 0.9? Seems to be near optimal for AUC optimization (~+1%) given the number of positives divided by batch size.

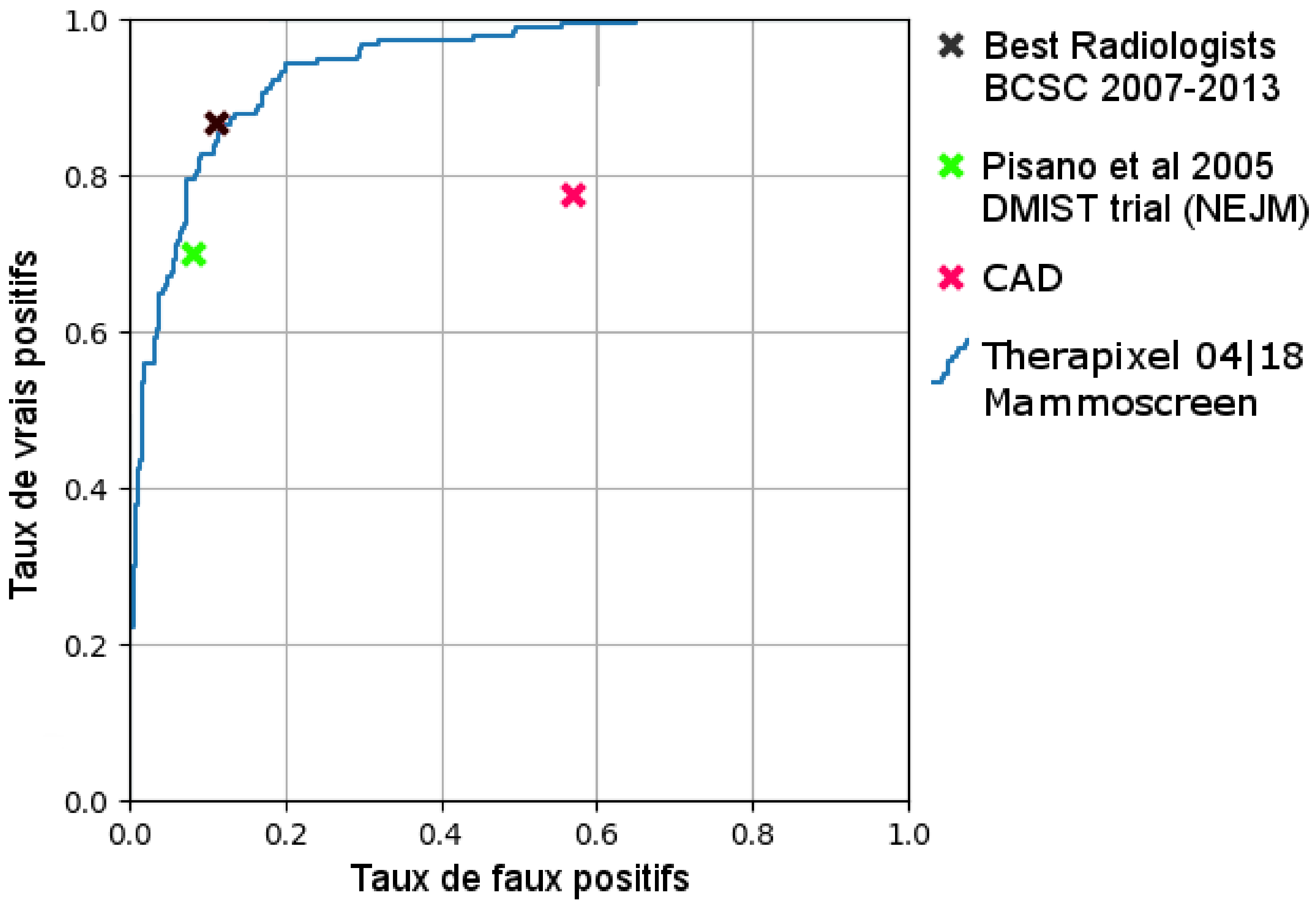
$$0.9^{125} = 2 \cdot 10^{-6}$$
$$0.99^{125} = 0.28$$

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

Current results on DDSM

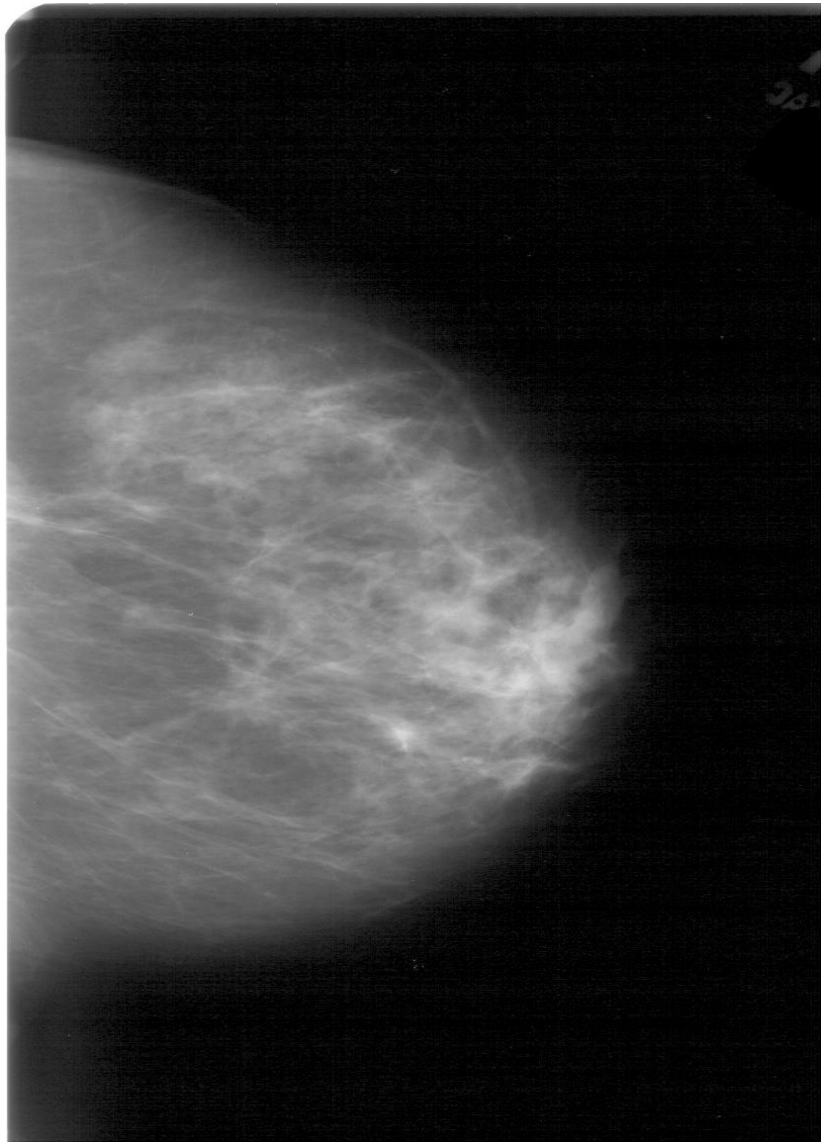


A note on overfitting and "advertising" stats:

- Overfitting happens on several levels:
    1. training data
    2. validation data
    3. test data = overfit dataset
    4. overfit a particular problem
    5. overfit a particular domain (?)
    6. overfit human style of thinking (??)

- In particular, performance of DL model on mammographies depends on:
    1. Device used for mammography
    2. Skills of technician
    3. Screening period (1-1.5-2 years)
    4. Positive/negative ratio, closely linked to
    5. Fraction of truly difficult cases
    6. Population (country)
    7. …

Saliency maps for weak detection

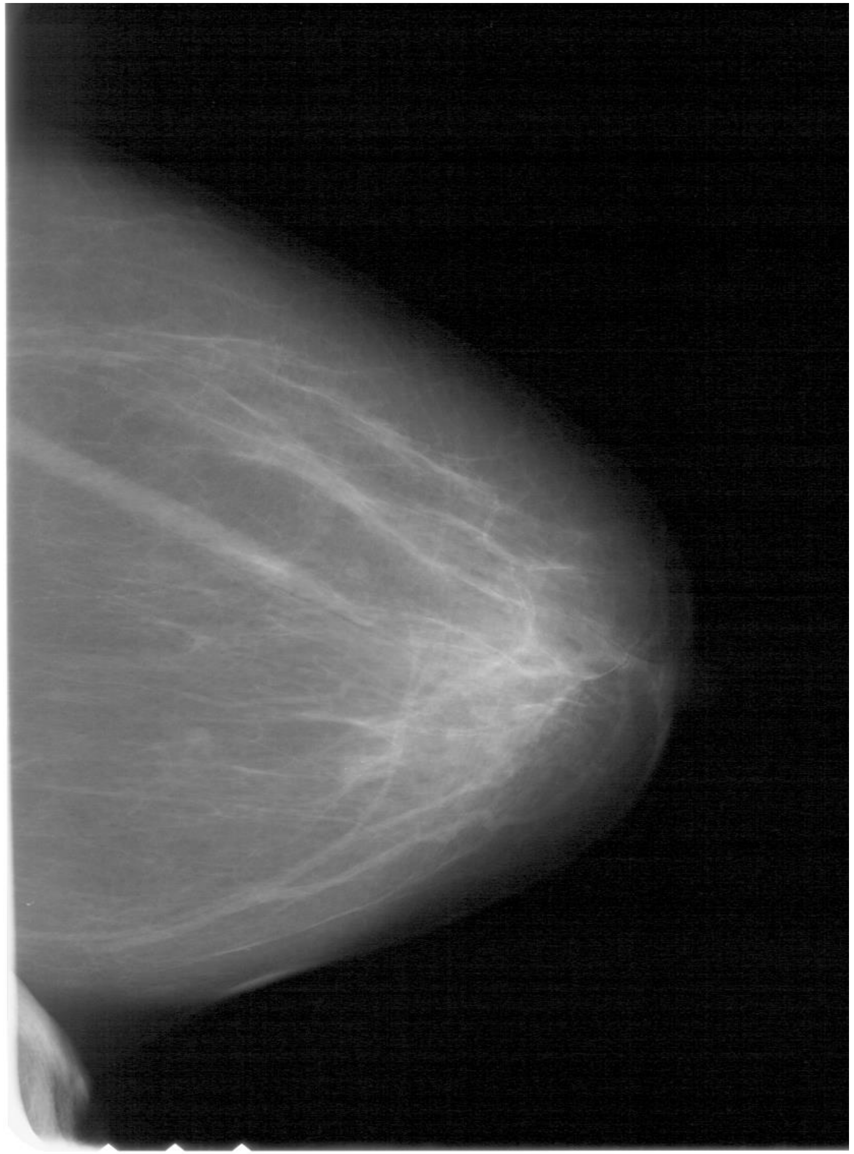


label = 1

Cancer. But where?

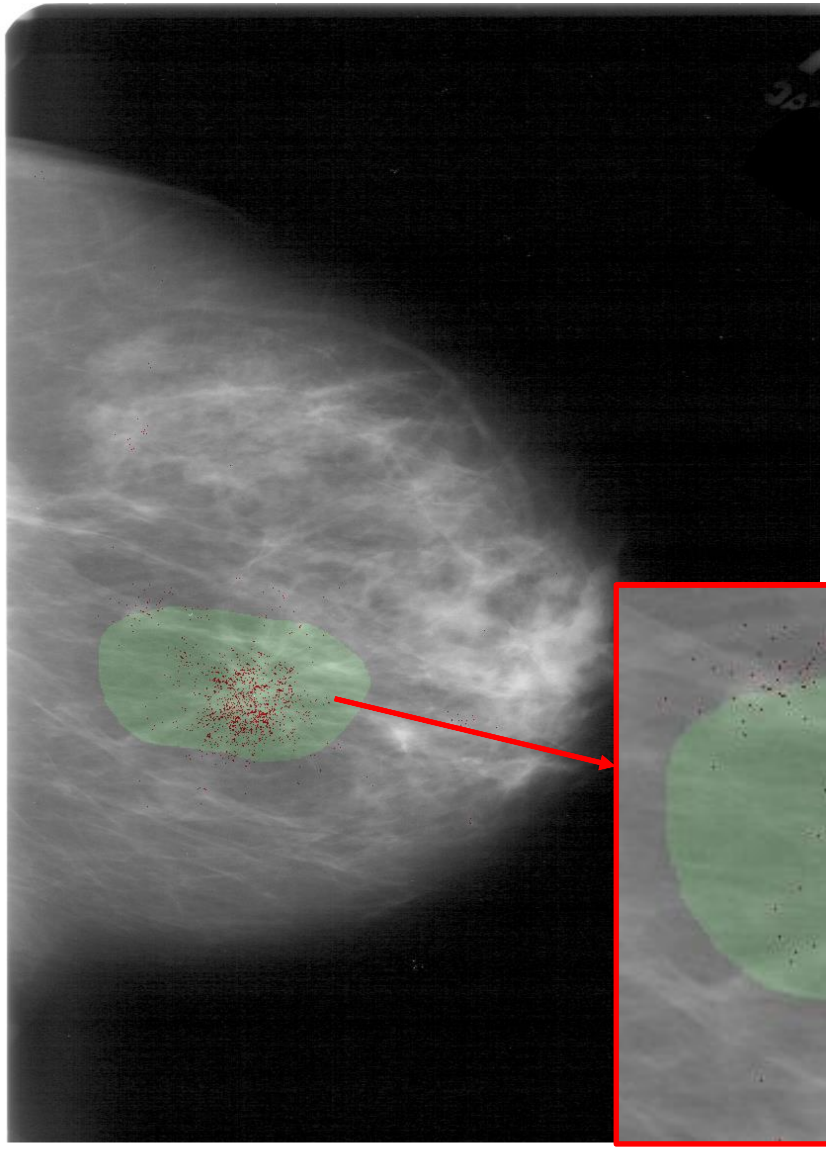
$$\frac{\partial O_1}{\partial Im}$$

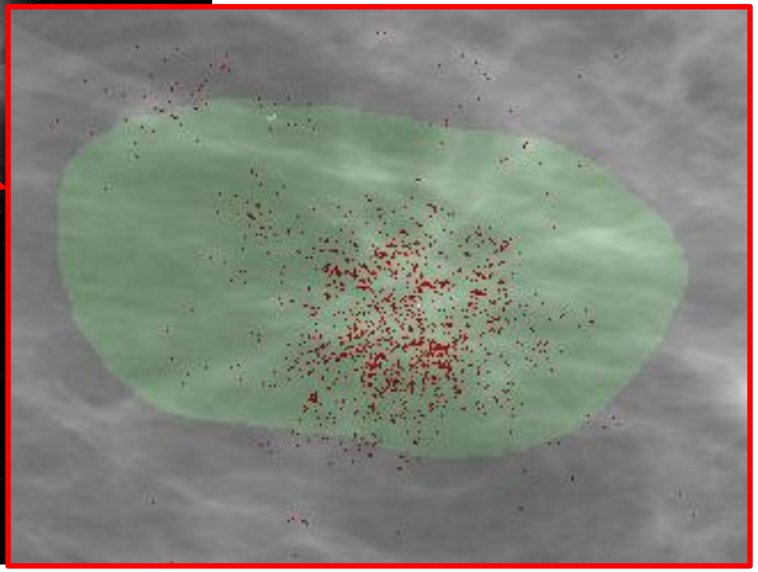
gradient of the output "1" with respect to the input



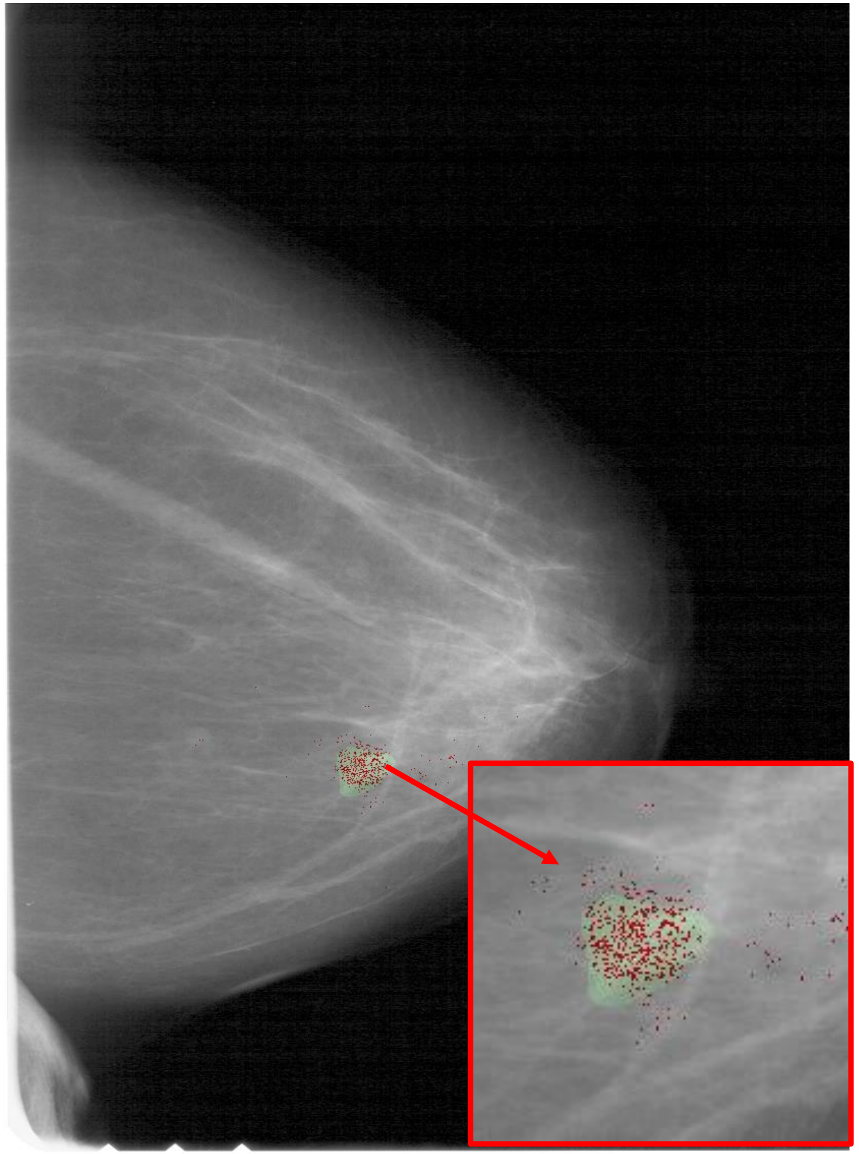
Idea credit: Simonyan, Vedaldi, Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2013

Saliency maps for weak detection

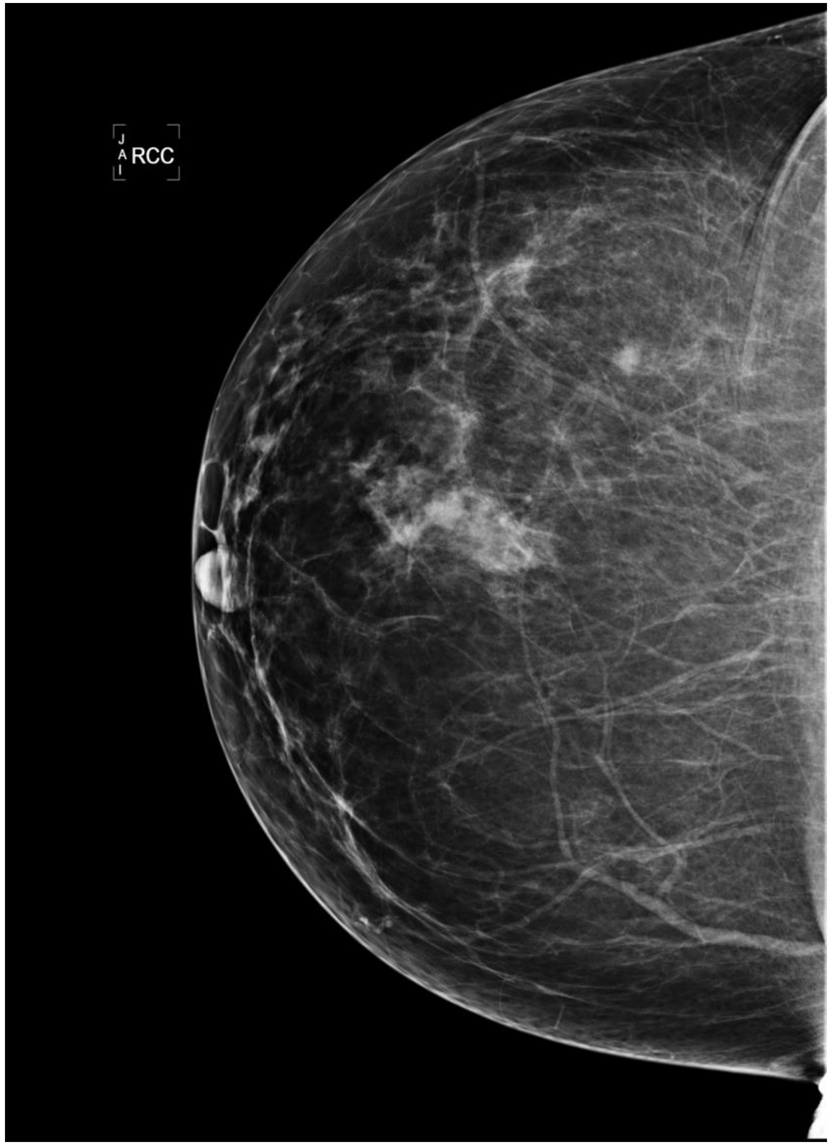




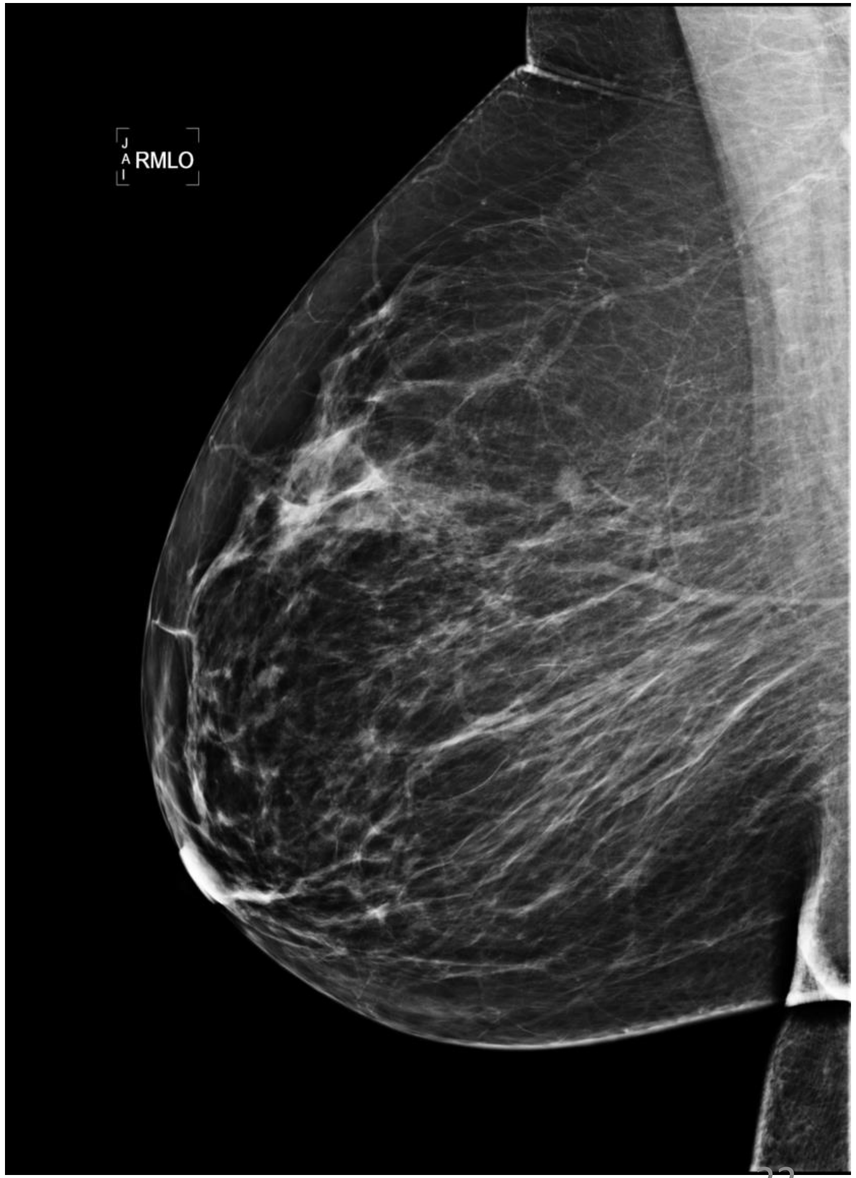
- Red dots – slightly post-processed saliency maps
- Green area – mask suggested by radiologist



Saliency maps for weak detection

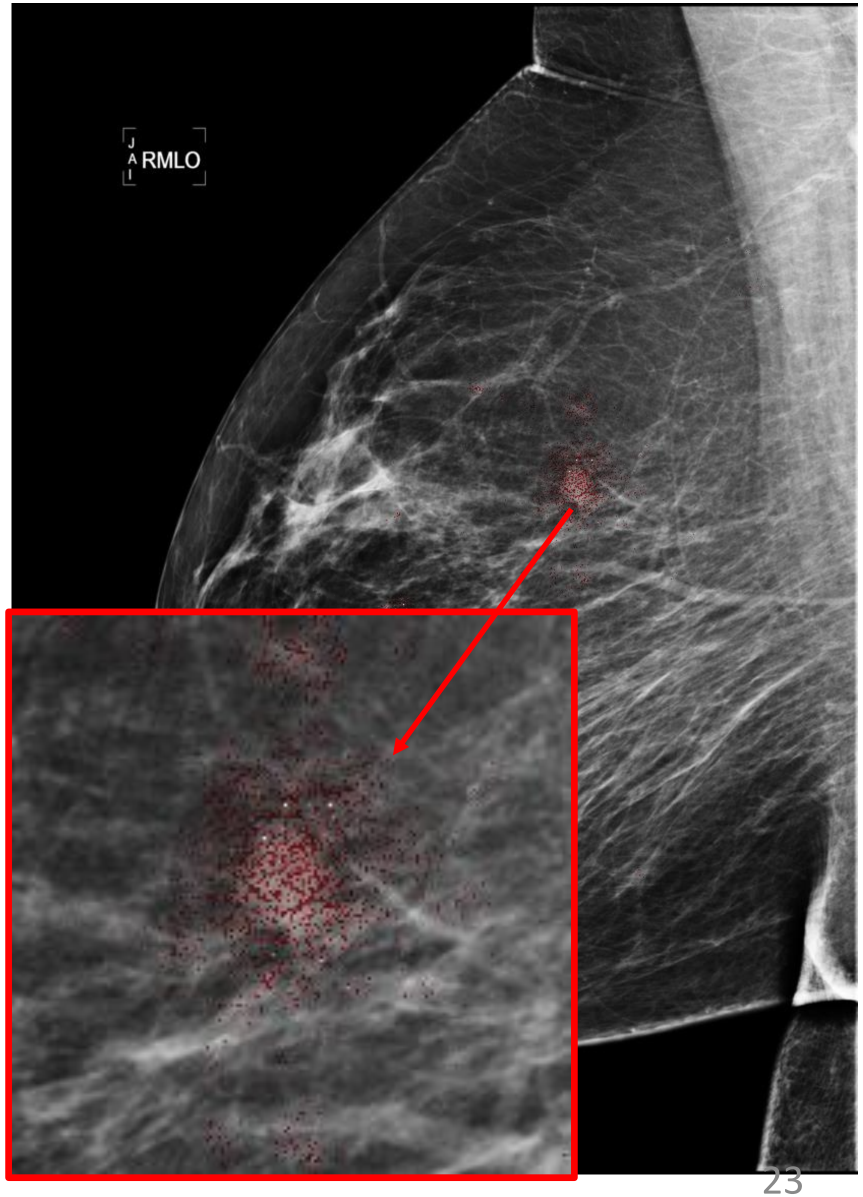


- DREAM data is much sharper

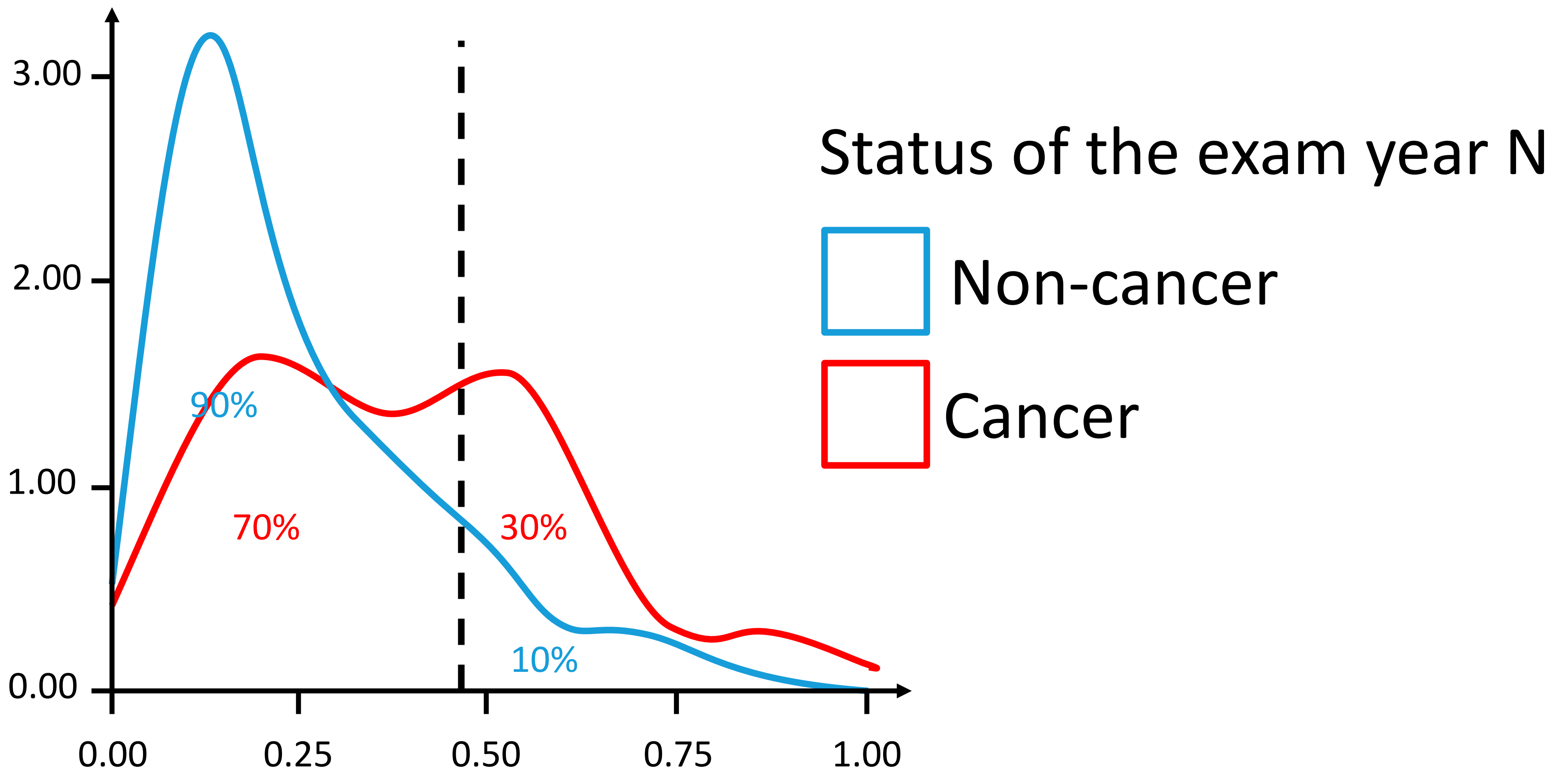


Saliency maps for weak detection

- Red dots – slightly post-processed saliency maps
- The same lesion is highlighted on both views



Model's output distribution on exams year N-1

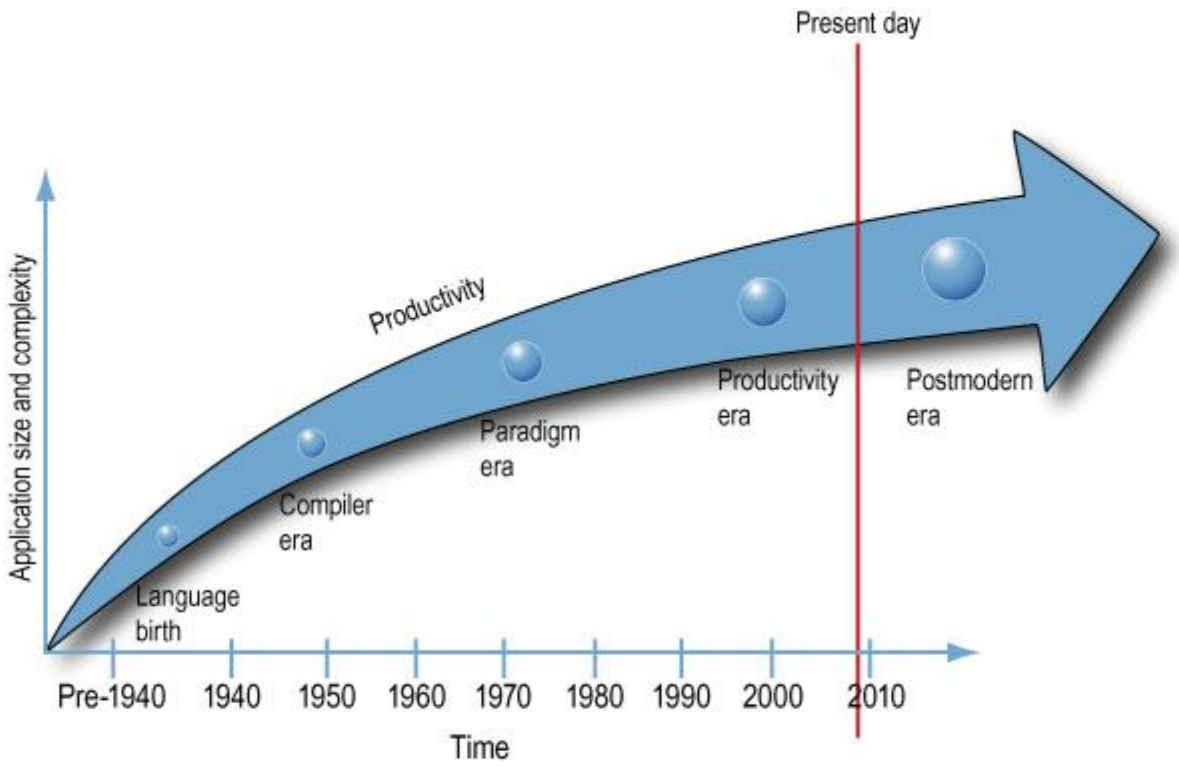


Part III: From research project to industrial product

- ML projects need new paradigm
- How we work at Therapixel
- Specific advices

Data Science 2019 = Software Engineering 1999

- 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?

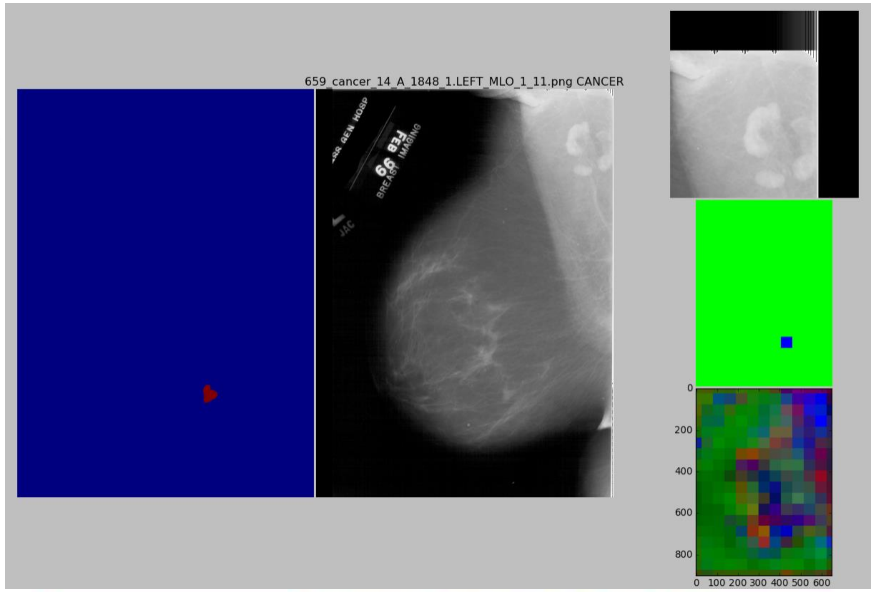


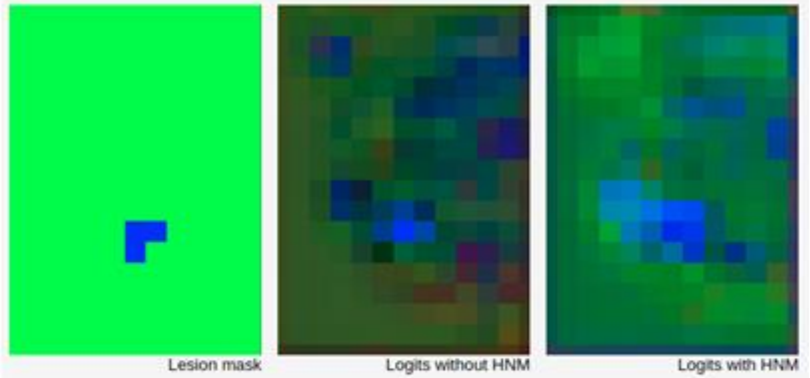
New paradigm – new development process and roles

- Therapixel:
  1. Development team
     - Cloud infrastructure
     - Integration with PACS in hospitals
     - Visualization & User Interface
  2. Data team
     - Partnerships with hospitals
     - Raw data extraction
     - Data clearing and structuring
  3. Research team
     - Interfacing of structured data
     - Running experiments, reporting errors
     - Testing new ideas and extensions

Some specific advices and practical moments

- Know your data
  - If you don’t understand your data – DL won’t either
  - Total nb of images, nb of images per class, typical resolution, RoI…
  - Regularly examine worst offenders, manually guide your model
  - Metadata is also under git (and dumped at each experiment)



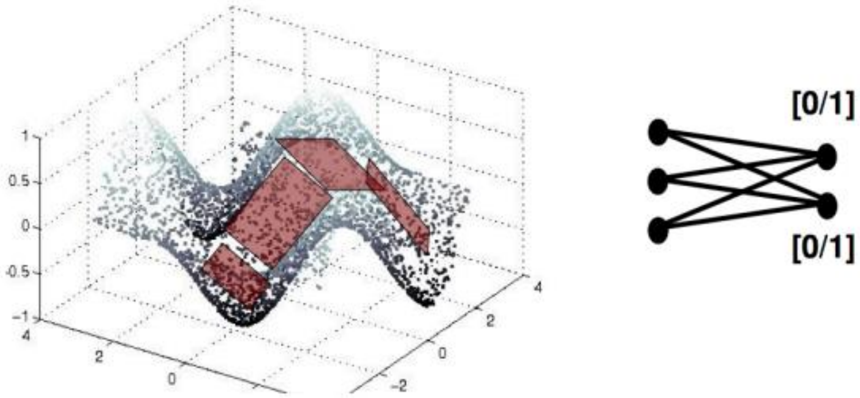


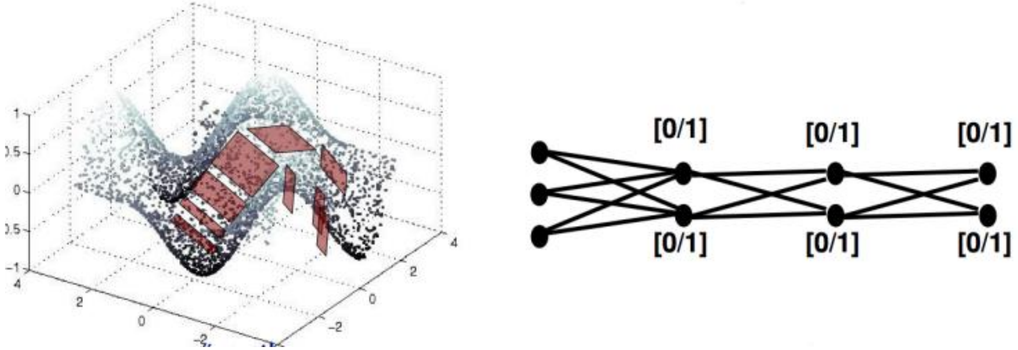
- Enforce reproducibility
  - No more binary reproducibility – GPUs
  - Each experiment has an output folder
  - For each experiment git hash and git diff are dumped
  - Unit tests where applicable (example: complex stats calculations)

- Work in team
  - Development cycles: 1-2 week
  - Regular meetings with discussions
  - Issue tracking tool
  - Code review

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



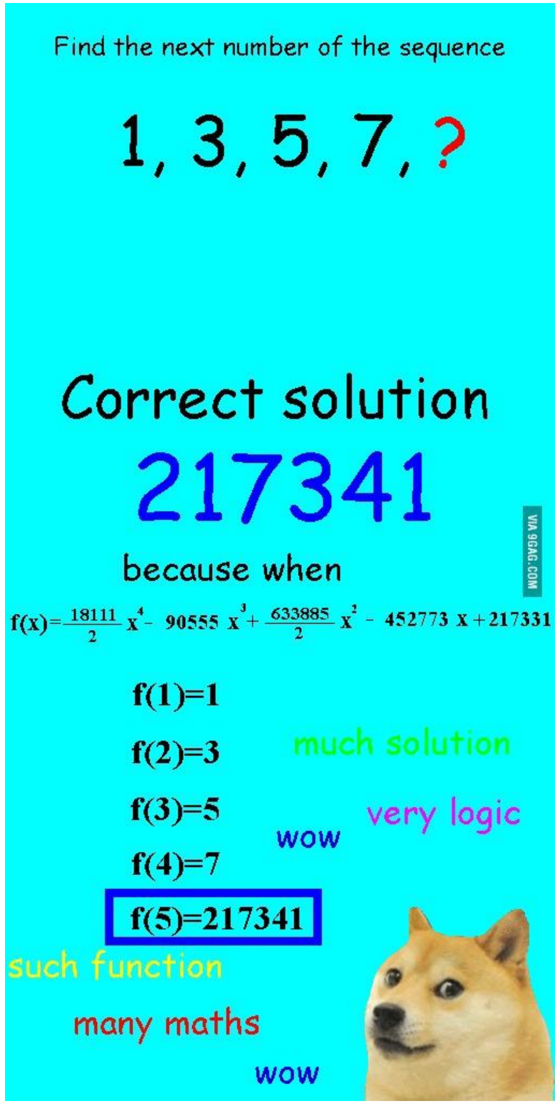


Slide credit: 1)G. Montúfar et al, On the Number of Linear Regions of Deep Neural Networks 2) Marc'Aurelio Ranzato slides 3) Introduction to Deep Learning by Iasonas Kokkinos

Some specific advices and practical moments

- ➢ Pretrain on balanced batches – make your network distinguish.

- ➢ Fight the overfitting: early stopping is simple but undesirable.
  - Smaller model
  - More data/data augmentation
  - Regularization



Thank you for your attention!

Q&A session

Therapixel