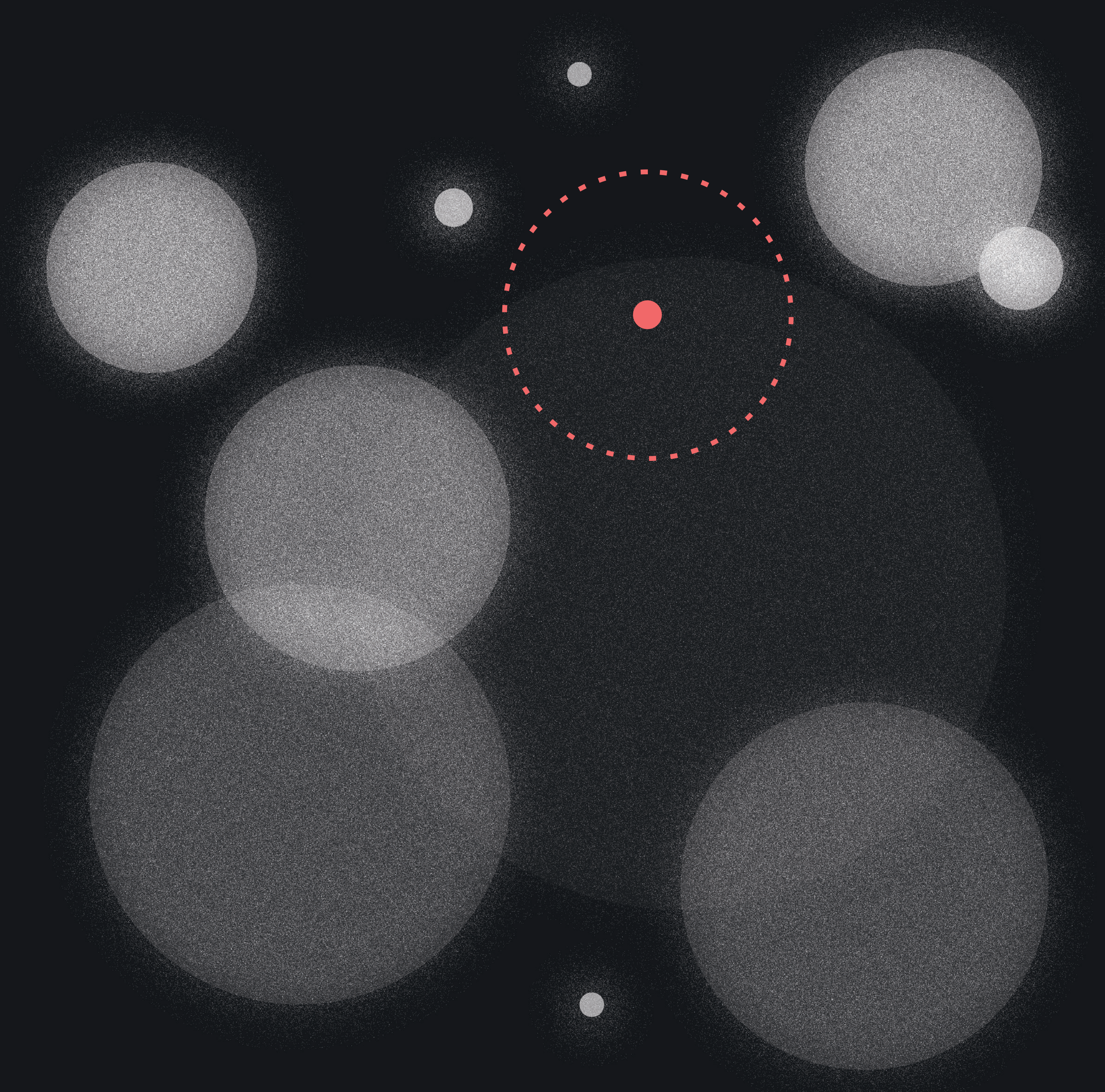
Radiobotics

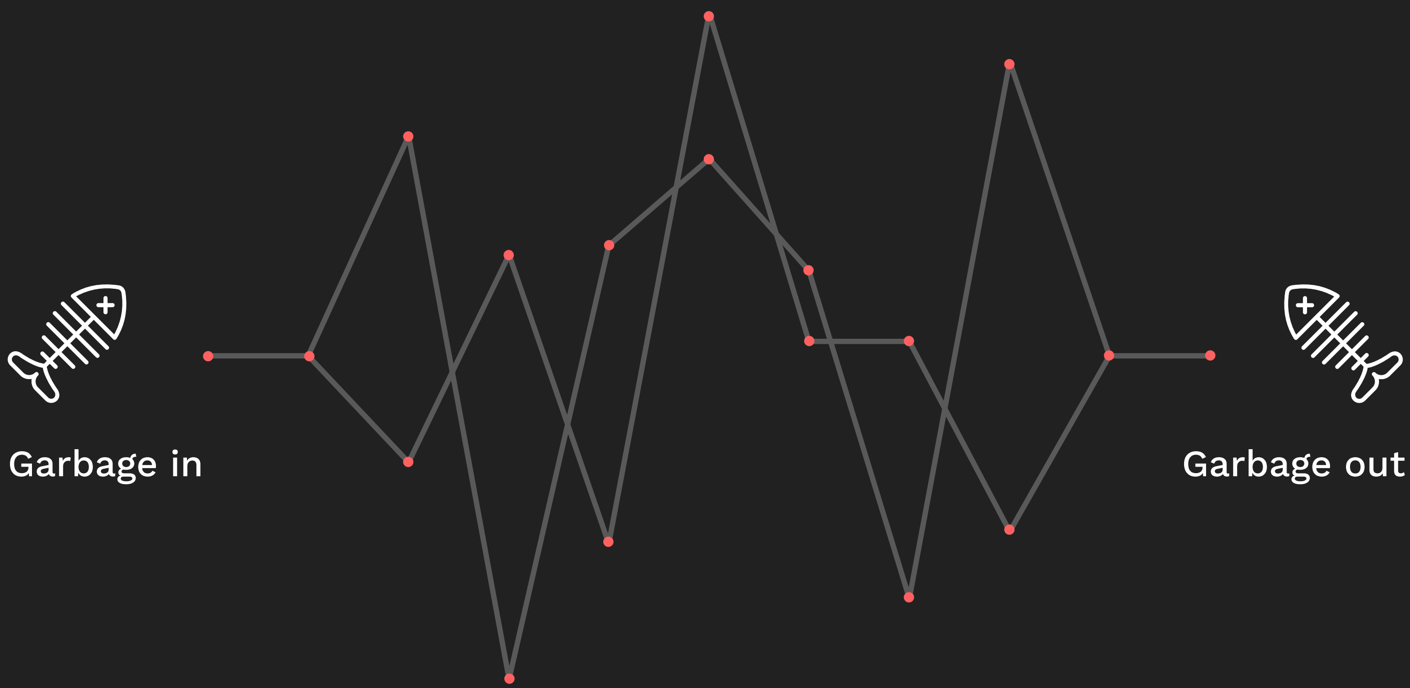


The best AI needs the best data:
How we ensure this at Radiobotics

A publication by Radiobotics
Written by Pavel Lisouski, Kasper Djernæs, Michael Lundemann

5/10/2021

Development of AI/ML/DL systems has two main components; the data and the code. Today, a lot of the development efforts and publications focus on the model and architecture improvements (code) - the so-called model-centric approach - which might be a more exciting component and a more standard approach since a lot of progress in deep learning has been on fixed datasets, where the focus of researchers has been to optimize the model. However, to obtain stellar results, the data component has to be equally (if not higher) prioritized.
Part of this is to ensure that you have enough data and that your data is of the highest quality possible, since the mantra “garbage in, garbage out” prevails.



In medical imaging, the size of data available is often relatively sparse (~ thousands of images) compared to other applications and domains where millions of images are used to train deep learning systems. This makes data quality even more paramount within the medical domain.

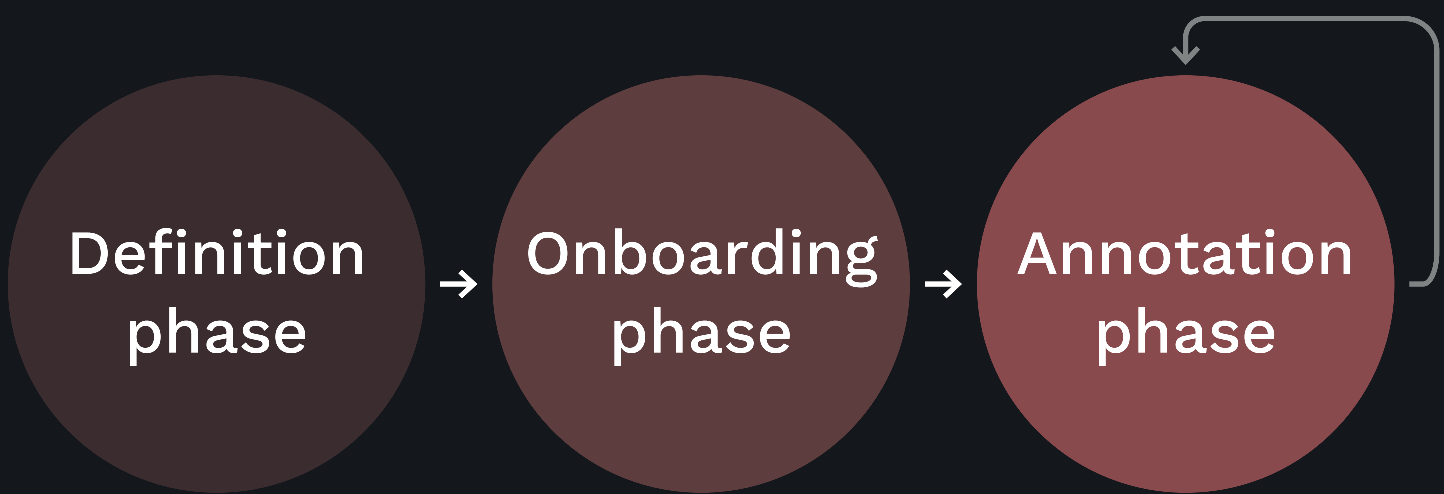
In Radiobotics, we develop AI systems that are able to detect different radiological findings or objects in medical images. Most often, we are both in charge of getting access to the data used for training and testing of the systems, but also creating the labels or "annotations" for the data. This can be grading of osteoarthritis of the knee from 0 to 4 or highlighting fractures using bounding boxes.

This publication will give a brief insight on our internal framework for ensuring that the quality of our data.

Radiobotics Annotation Process

It is tempting to annotate a data set as fast as possible to be able to move on to the more exciting part - the model development and training. However, this initial part of the process will in the end have a huge impact on the system performance no matter how deep your neural network is.

We've defined a generic data annotation process that consists of the following three phases;



01
Define metrics & annotation goals

02
Recruit and train annotators

03
Annotate data and track quality of the annotations

01

Definition - Quality Objectives & Requirements

The first phase of the annotation process is the definition phase where the goal is to define the requirements of the data & annotations, metrics to measure the performance and quality of the annotations. Some examples that could be discussed during this phase are:

Data size
How much data is needed? Sometimes this can be hard to answer if the data size is predefined. However, looking into publications of similar systems one can get a good estimate of the data size. This will have an impact on how long time it will take to annotate the data and in the end the cost of the project

Data diversity
do you have a diverse data set in terms of gender, age, racial identity, data from multiple hospitals etc.?

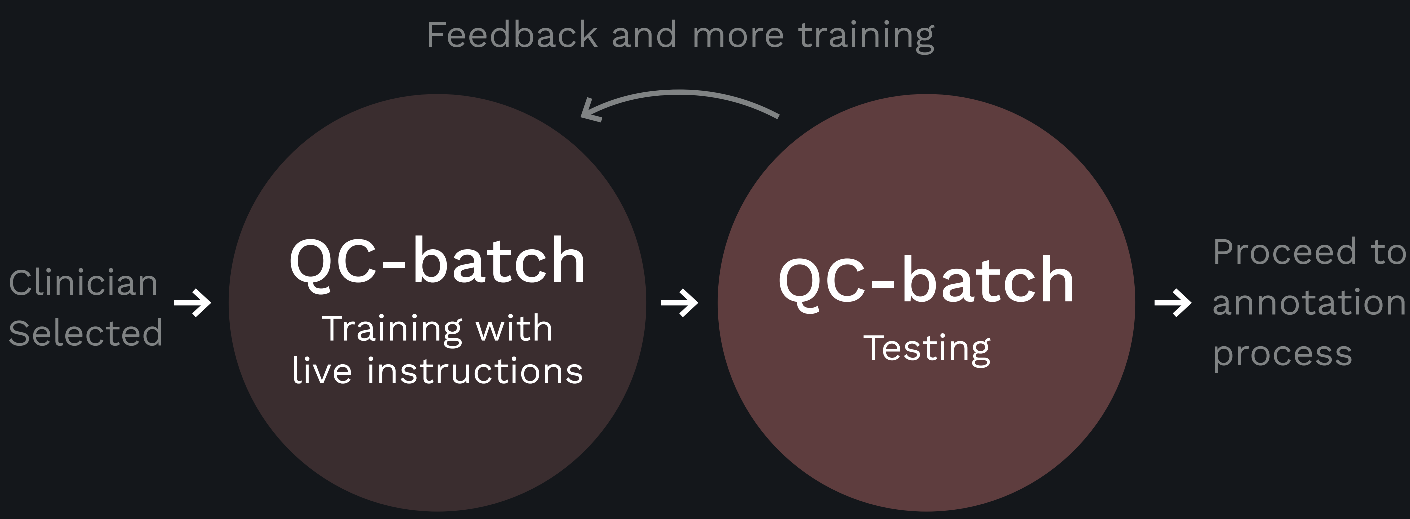
Representation of classes
Medical data can be very unbalanced, therefore it might be important to ensure in advance that all classes are well represented in the data set.

System performance metrics
In the end, it is the final system performance that will matter, therefore it might be important to align the way the annotations are being created with the system performance requirement. If your system ought to detect a specific object with pixel-precision, you should ensure that the annotations are also performed with this high precision.

Data Quality metrics
An important aspect of this is to ensure that the ground truth is unambiguous, i.e. the annotators agree with each other. A good approach is to use multiple annotators on a subset of the data and calculate the inter-annotator agreement or use a majority vote as the ground truth.

02 Onboarding

During the on-boarding phase, the task is to find and recruit the right clinicians for the annotation tasks, ensure they have the right qualifications and get enough training and test their ability to perform the task. Clinical personnel (radiologists and radiographers) might have specialized in different modalities or even pathologies within the radiological domain, therefore it is important to ensure that the clinicians have enough experience working with exactly that specific problem the system is ought to solve.



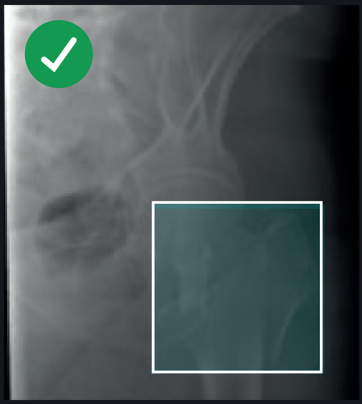
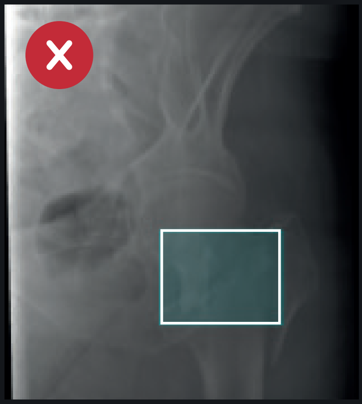
Some important steps during on-boarding process:

- **Recruit** the right clinicians and ensure their qualifications
- **Train the clinicians to perform the task** - labeling and annotation of deep learning data is far from the same as writing radiological description in clinical practice. One good approach is to create an annotation atlas that can be shared with clinicians
- **Perform tests on on-boarding images ("dry-run")** and provide feedback. Iterate until you have a satisfying performance

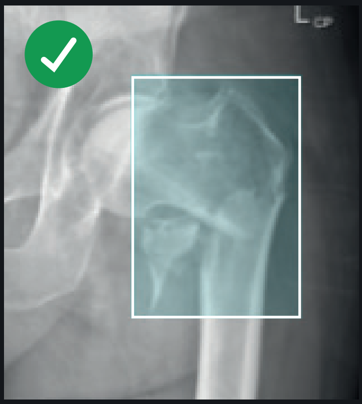
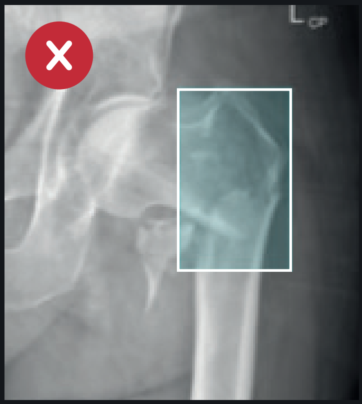
It's important to train the clinicians in performing the annotation task. Annotation of bounding boxes around fractures or segmentation of different pathologies in the medical image is different from clinical practice. Before training, we always create an "annotation atlas" that provides clear instructions and examples of how to, and how not to annotate the data.

Here's an example of how to annotate the hip fractures in specific cases:

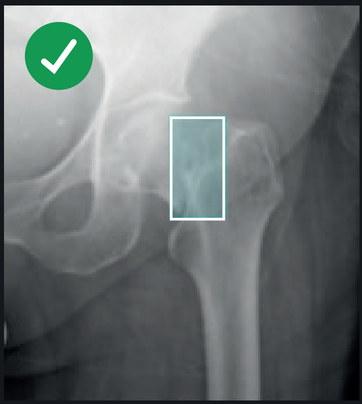
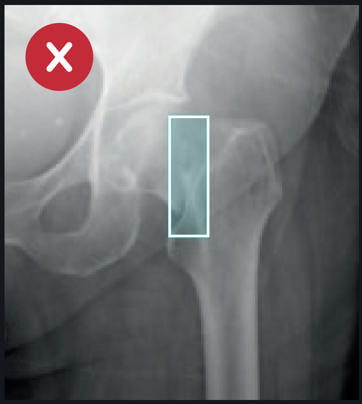
Case 1
Include the entire fracture site

Case 2
Include trochanter minor/major if detached

Case 3
Do not include too much in top or bottom outside fracture

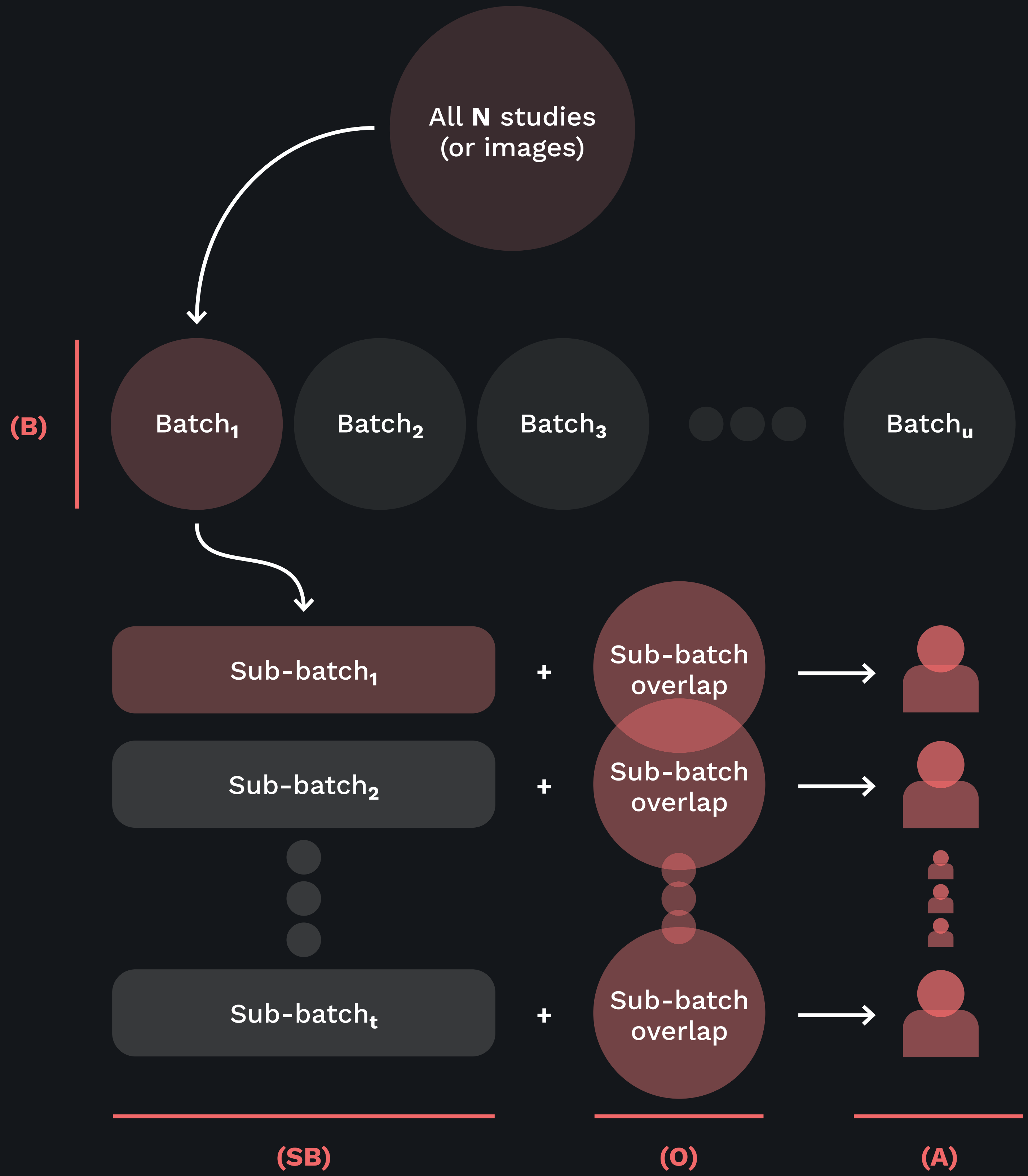



03 Annotation & Continuous Quality Inspection

During the annotation process, the goal is to effectively annotate the dataset while continuously monitoring the quality of the “annotators”. At Radiobotics, we have implemented a schema to monitor agreement between the annotators, which gives us the ability to monitor a few things; if the inter-reader/-annotator variability is low - this is an indication that the task is “understood” and that the ground truth is unambiguous. Also, if one of the annotators' performance is “drifting”, we are able to detect it and correct it by providing more training.

A simple overview is illustrated to the right →

1. The data set is split into batches **(B)**
2. Batches are split into sub-batches **(SB)**, where a sub-batch overlap **(O)** is defined (e.g. 10% if the data is annotated by all three clinicians)
3. Each sub-batch + the overlap is assigned to each annotator **(A)**
4. When all annotations have been completed, quality metrics are calculated (e.g. average IOU or benchmark agreement between the annotators)

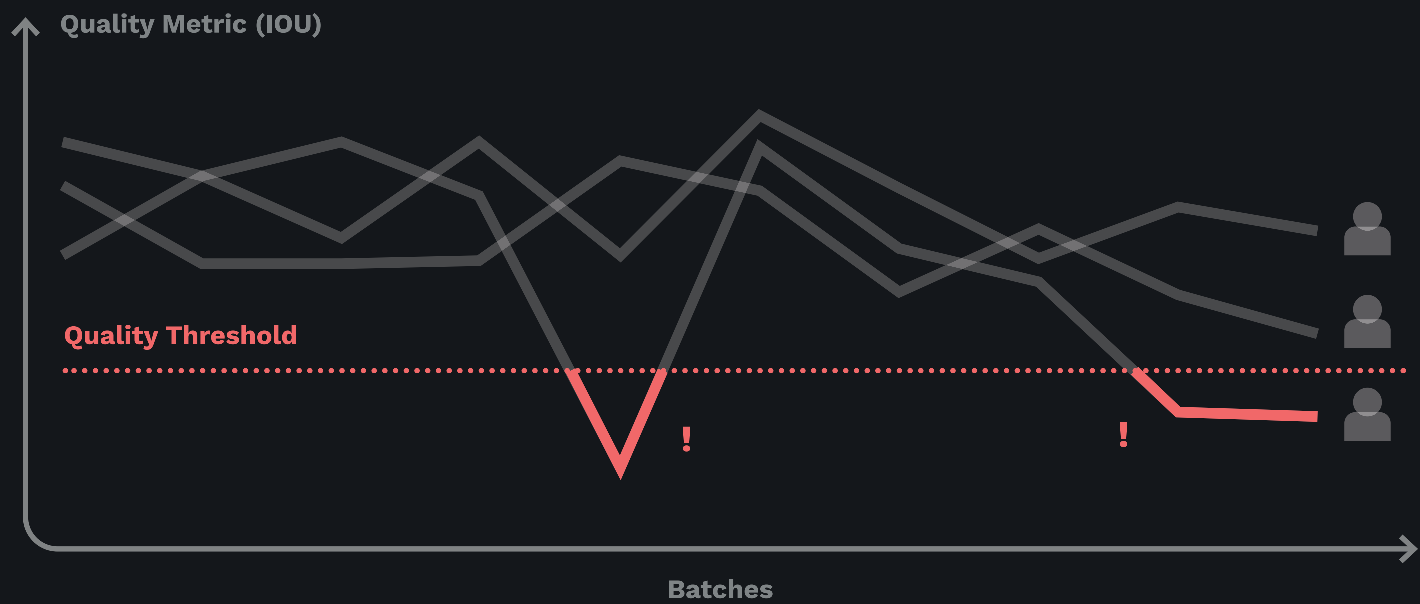


Continued on next page

Based on the quantitative (i.e average size of bounding boxes) and quality metrics (frequency of disagreement with majority), it is possible to identify annotators that are deviating from the norm.

By continuously tracking these metrics it is possible to identify which annotation samples need to be discarded, need more quality control or which annotators might need additional training for realignment.

To setup this kind of automatic process might require some time to setup and development effort. However, this will in the end pay-off in terms of reduced development time during model development and training.



An example could be that the average size of Annotator A's bounding boxes are consistently significantly smaller than Annotator B and C's across several overlaps. Using this framework, it is easy to detect this disagreement, pause Annotator A's annotation progress and show where the misalignment is largest. Then run quality control on different overlap of sub-batch to ensure correction and "re-release" Annotator A into the annotator pool.

Final Remarks

Working with AI requires a lot of focus on the data and data quality. It is very tempting to allocate most resources on development of AL/DL algorithms, however, your data will in the end define how good performance you can get.

In this publication we showed how we internally evaluate and track the quality of our data that in the end will be used for development of our products. To summarize:

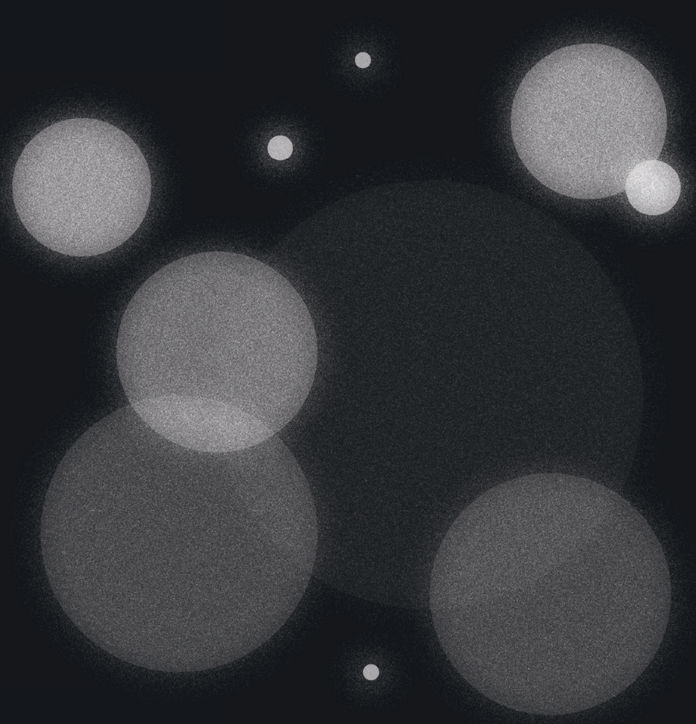
1) Define metrics and data quality KPIs

2) Evaluate and train annotators before annotating on production data

3) Continuously evaluate annotator performance (e.g. by having multiple clinicians annotating the same data and measure their agreement)

If you want to know a bit more about our data quality process, watch our QA Manager Astrid Ottosen presenting how we use this process as part of our Supplier Evaluation Process, on our website under publications.

Radiobotics



About Radiobotics

Radiobotics is a multiple award-winning health tech company having their HQ in Denmark and with an office in Texas, US. The company has build an innovative AI technology specialized in x-ray analysis with focus on musculo-skeletal radiology. Based on advanced computer vision and machine learning methods, Radiobotics' algorithms generate fully automated, objective text and visual reports.

Contact

info@radiobotics.com
www.radiobotics.com

A publication by Radiobotics
Written by Pavel Lisouski, Kasper Djernæs, Michael Lundemann