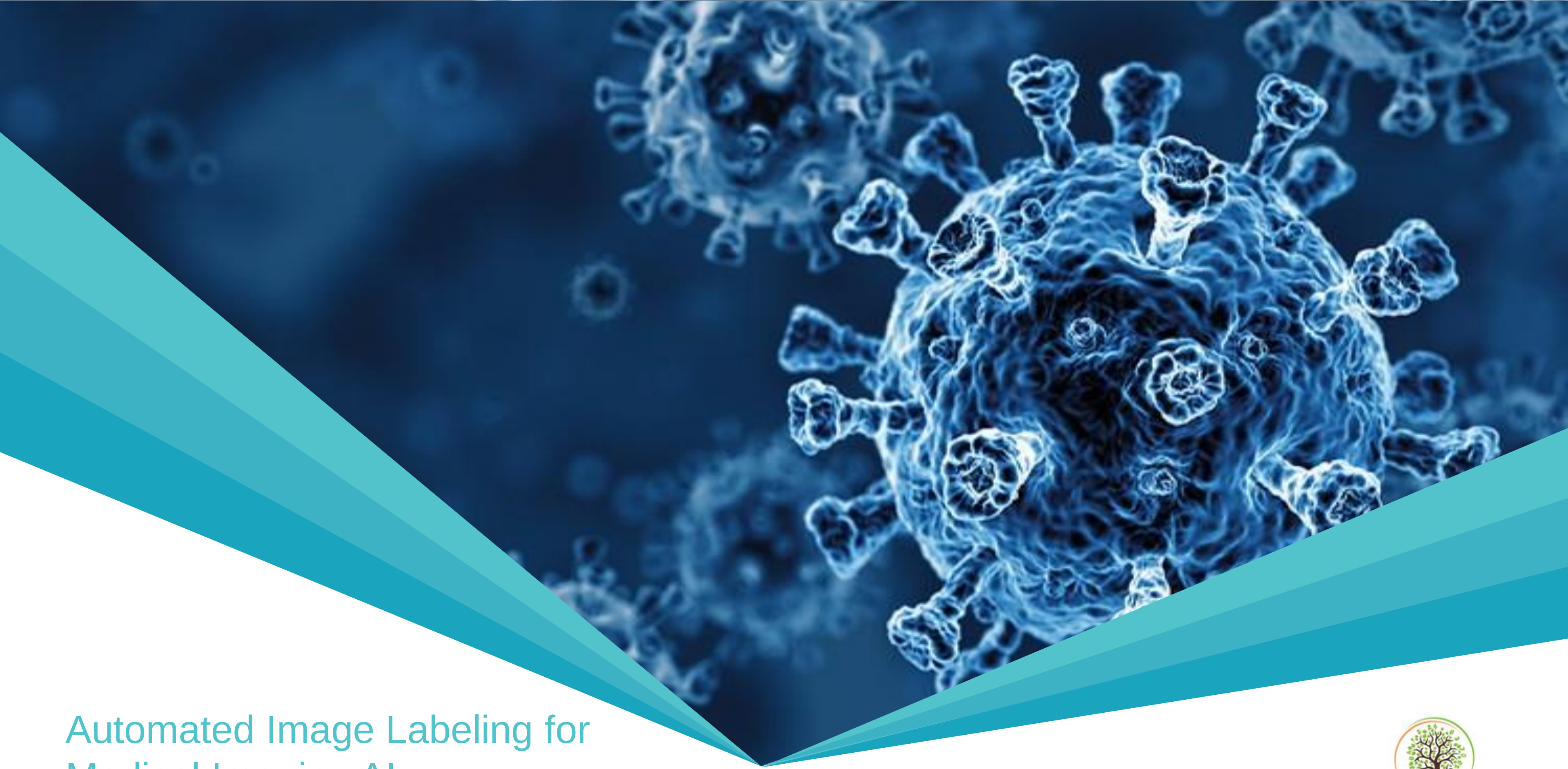


Automated Image Labeling for Medical Imaging AI

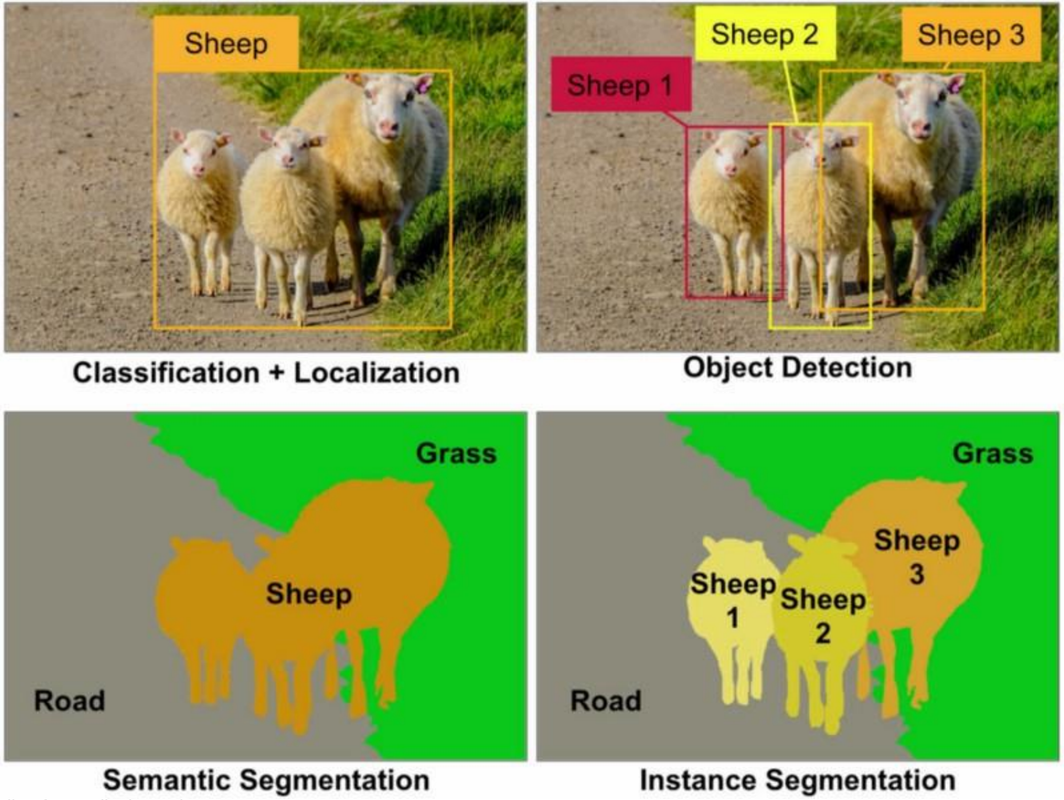
Patrick Bangert, Samsung SDS

May 14, 2021 for DahShu

PROMOTE RESEARCH AND EDUCATION IN DATA SCIENCE

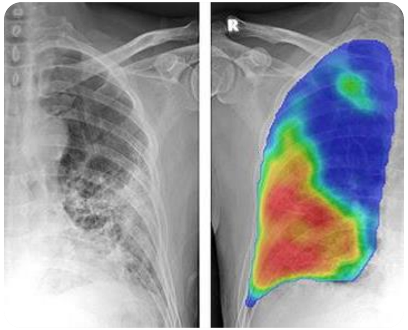
Computer Vision Applications

There are four standard use cases for computer vision requiring the relevant annotations.



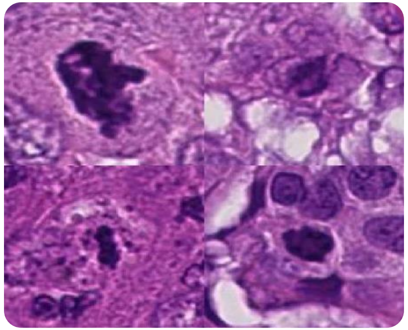
Medical Imaging AI Examples

AI detects diseases on images and thus helps doctors by pre-screening for normality vs. pathology.



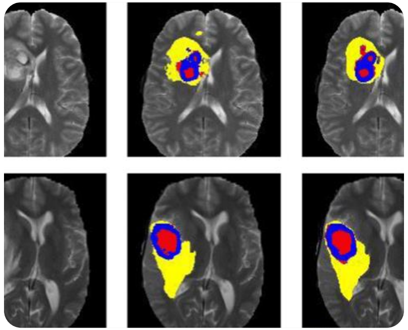
X-Ray

Detect Covid-19, pneumonia, or normality



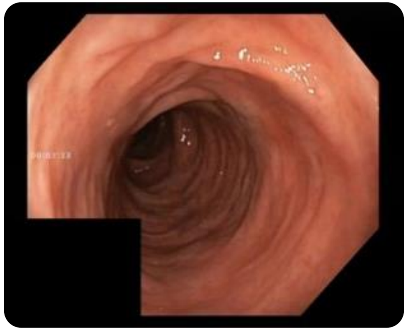
Whole Slide Image

Determine the presence or absence of breast cancer.



CT

Find the brain tumor, if any.



Photograph

Detect presence of absence of colon cancer.

AI Project Workflow

Before we can do AI, images must obtained and labeled. Then AI involves several steps as well.

AutoLabel → AutoML →

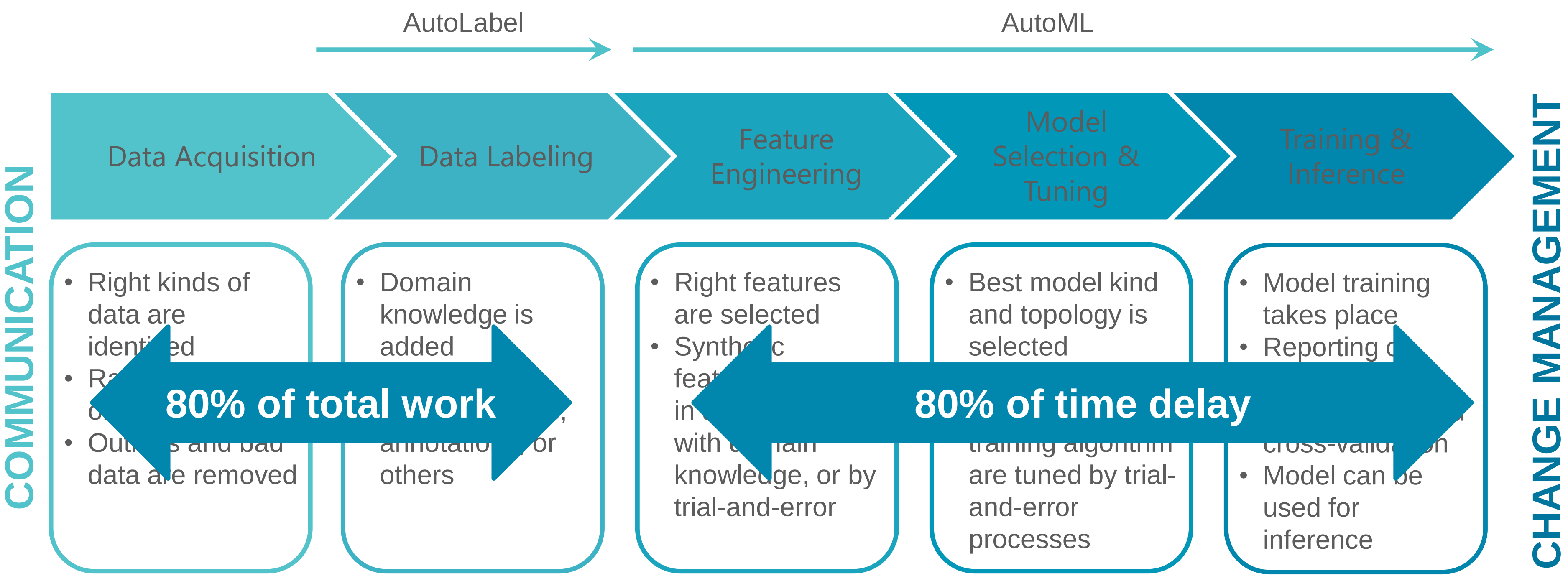
COMMUNICATION

Data Acquisition	Data Labeling	Feature Engineering	Model Selection & Tuning	Training & Inference
• Right kinds of data are identified • Raw data are obtained • Outliers and bad data are removed	• Domain knowledge is added • Could be categorizations, annotations, or others	• Right features are selected • Synthetic features created in accordance with domain knowledge, or by trial-and-error	• Best model kind and topology is selected • Hyper-Parameters of training algorithm are tuned by trial-and-error processes	• Model training takes place • Reporting on model performance and cross-validation • Model can be used for inference

CHANGE MANAGEMENT

AI Project Workflow

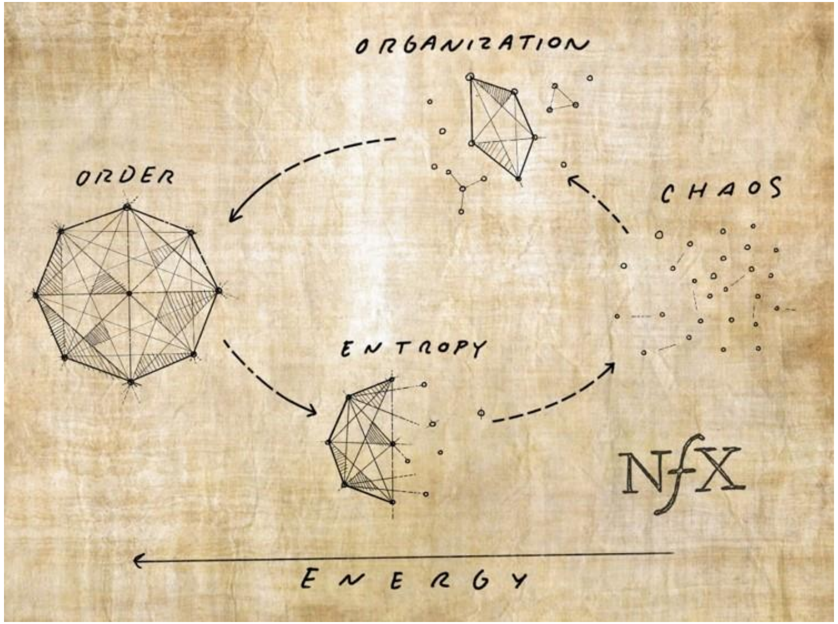
Before we can do AI, images must obtained and labeled. Then AI involves several steps as well.



Labeling is Effortful

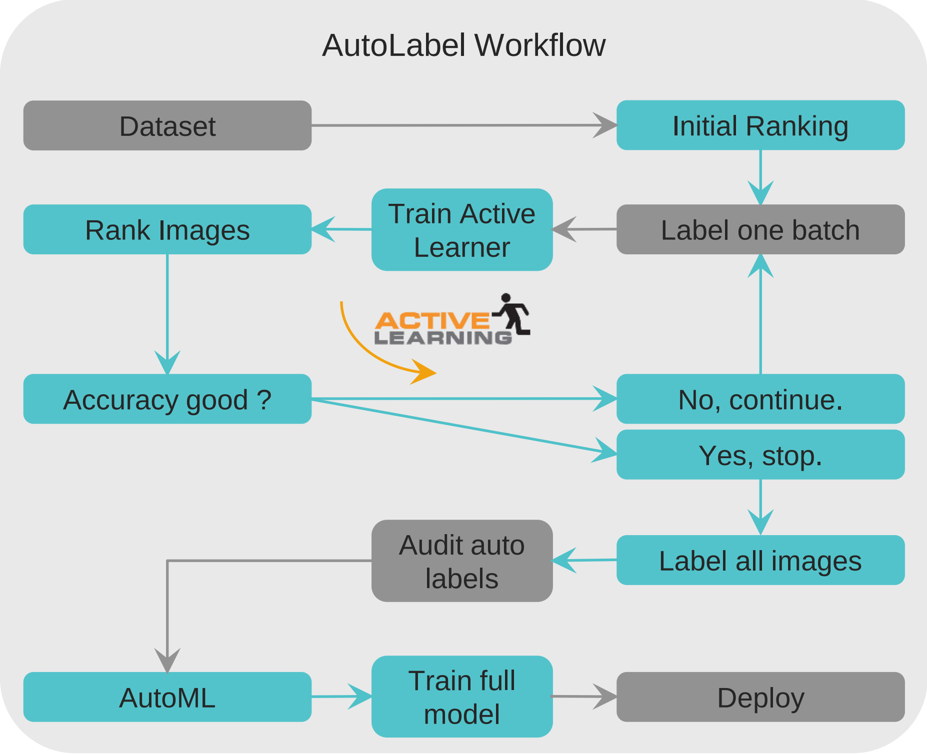
Humans must label each data point in the training dataset and generally do so in arbitrary order. But order matters!

- As each data point is labeled, we gain information
- However, some data points are worth more than others
- Let's sort them and label the good ones first
- Law of diminishing returns
    - At some point, the additional information gain is so small, that we can stop labeling before getting to the end of the training dataset



AutoLabel / AutoML Technology

Active learning can be implemented efficiently backed by distributed training to run in realistic time.



… with a little help from our friends …

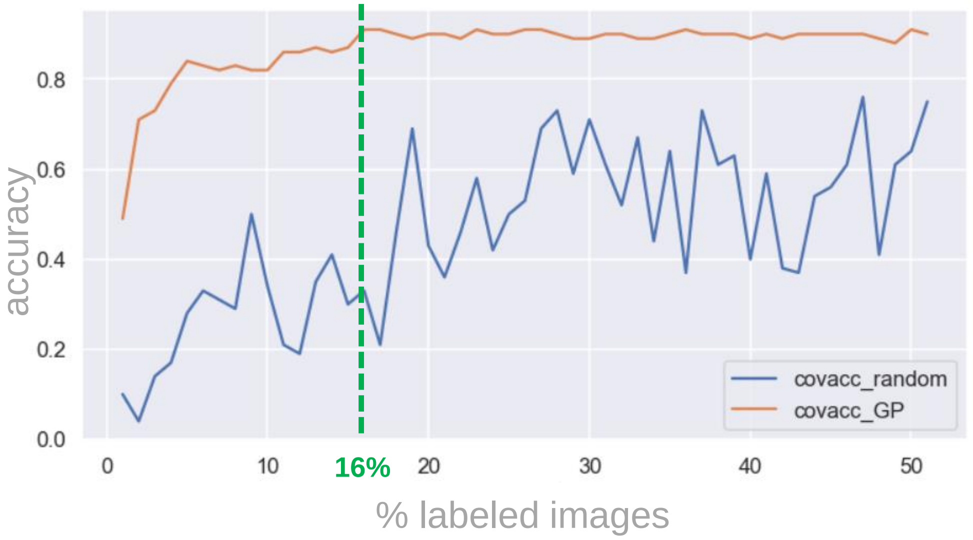
PyTorch, OpenAI, Keras, mxnet, scikit learn, theano, jupyter, TensorFlow, HOROVOD, HELM, Chainer, Caffe, Terraform, aws, Azure

manual activities

Active Learning – Order Matters!

Data points are sorted by going through a loop of a little labeling, training, and evaluating the confusion of the model.

- Accuracy grows faster with active learning than compared to a random order.
- Nearly all information is contained in a small subset of the data.
- Labeling can stop at this point.
- The point at which this happens depends upon the dataset but typically occurs between 5 – 15%.



Labeling Medical Images is Costly

Active learning helps experts label only the most informative images manually – providing ~80% automation.

01

Segmentation Labeling

Requires between 2 – 5 minutes per image per labeler. Between 3 – 6 people label each image.

02

Checking Labels

Requires between 10 – 20 seconds per image per auditor. Between 1 – 2 people audit an image.

03

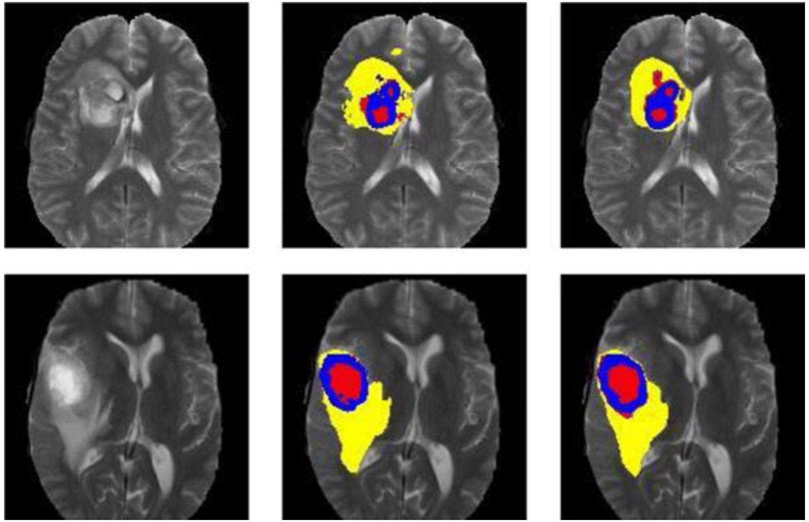
Data Set Size

Typical datasets are between 100k – 1m images.

04

Cost (example)

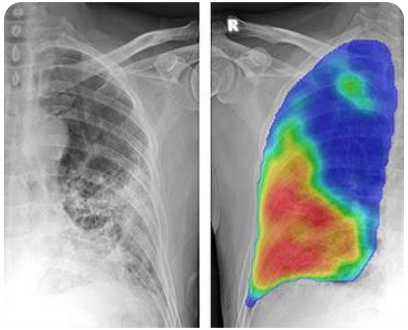
Labeling everything manually costs **$12m** and 19 person-years. Using active learning reduces cost to **$1.35m** and 2 person-years.



Assuming 4 labelers, 2 minutes per image and labeler, 1 auditor, 10 seconds for auditing, 250k images, and an hourly cost of $350 typical in medicine.

Medical Imaging AI Examples

AI detects diseases on images and thus helps doctors by pre-screening for normality vs. pathology.



X-Ray: Covid-19

Of 15254 images, only 915 had to be labeled. Automation of **94%**.

AutoLabel Accuracy: **95%**

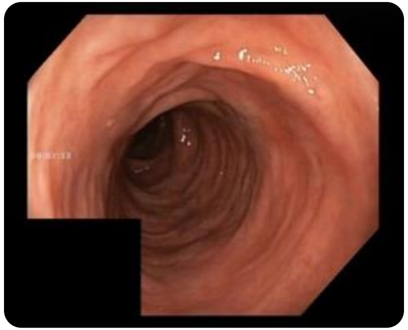


Photo: Colon Cancer

Of 6000 images, only 500 were labeled. Automation of **92%**.

AutoLabel Accuracy: **100%**

Automation: This many images were labeled by the AI as opposed to a human.
Accuracy: This many automatic labels were correct, as determined by audit.

Active learning helps humans to label efficiently. It is not a substitute for the AI training of the diagnostic model once a fully labeled dataset is obtained.

Market for Medical Image Labeling

The market for medical image AI is large and growing – labeling is the main obstacle for growth.

"The global **medical imaging** market size was $33.69 billion in 2019 and is projected to reach $43.33 billion by 2027."

FORTUNE™ BUSINESS INSIGHTS

"**AI-enabled medical imaging** solutions market size to be worth USD 4,720.6 million by 2027."

EMERGEN RESEARCH

"The global **data collection and labeling** market size was valued at USD 1,307.7 million in 2020. It is expected to expand at a compound annual growth rate (CAGR) of 25.6% from 2021 to 2028."

GRAND VIEW RESEARCH

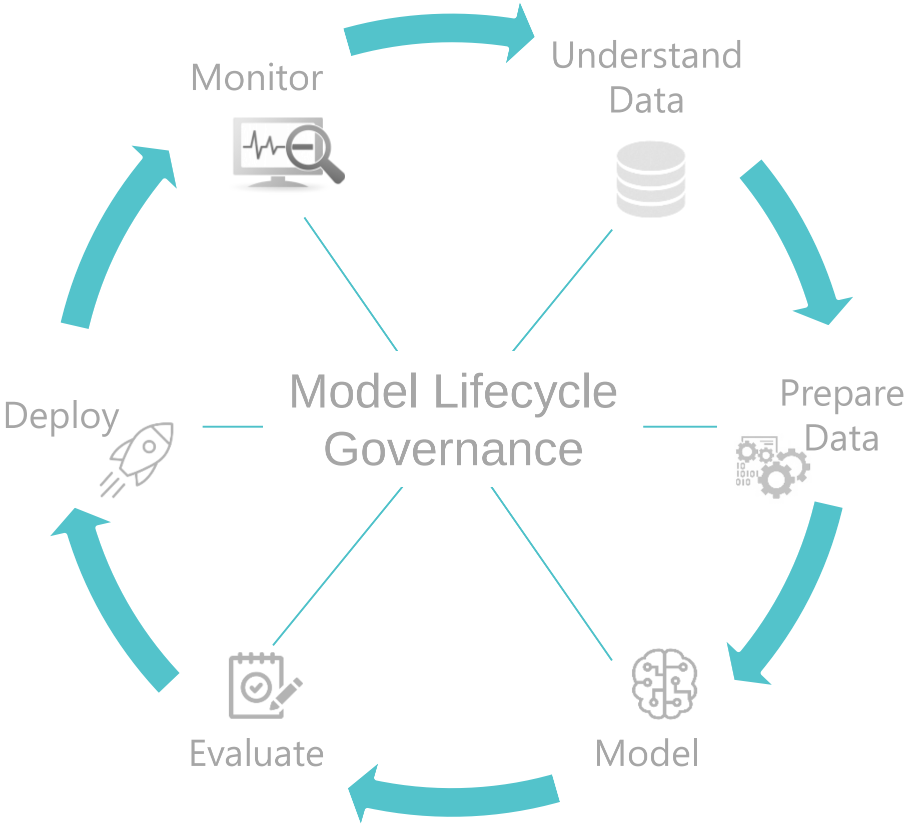
"The global **data annotation tools** market was valued at $320 million in 2018 and is expected to reach $1,820 million by the year 2026."

SDMR International

Invitation: Want to try this out?

Be part of the journey to validate the science.

- Get in touch if you are interested in validating this on your dataset.
- Scientific paper: https://arxiv.org/abs/2103.05109
- Two-part popular article
    - https://www.linkedin.com/pulse/artificial-intelligence-covid-19-screening-covid-using-bangert/
    - https://www.linkedin.com/pulse/teach-computer-vision-training-covid-scans-part-2-patrick-bangert/
- Demo video of the technology: https://youtu.be/wcP1fRPKXSU



Thank you

Thank you for your attention.

화이팅 !

Patrick Bangert
p.bangert@samsung.com

Abstract

Computer vision application in the medical domain have recently become quite popular. Detecting, localizing, and diagnosing diseases and recognizing structures on MRI, CT, PET, X-ray, ultrasound and photographic images is efficiently done by AI. What is rarely discussed, is how these AI systems are made. They require copious amounts of training data consisting of images and human-supplied labels. The labels are marked areas (either rectangles or free-form boundaries) that are attached to a word, e.g. “tumor.” Naturally, only highly qualified professionals are able to provide these labels making the process effortful and expensive. A technique called “active learning” from AI can help with this by reducing the manual effort by 90%. This allows the creation of state-of-the-art AI models for medicine using a significantly smaller budget of time and resources. This talk will present the method with several examples and will argue that this approach is a disruptive shift in medical AI.