ASSC 22 Keynote:

Evaluating Production Robustness of Deep Learning Models

Dr Kelvin Ross

26 May 2022

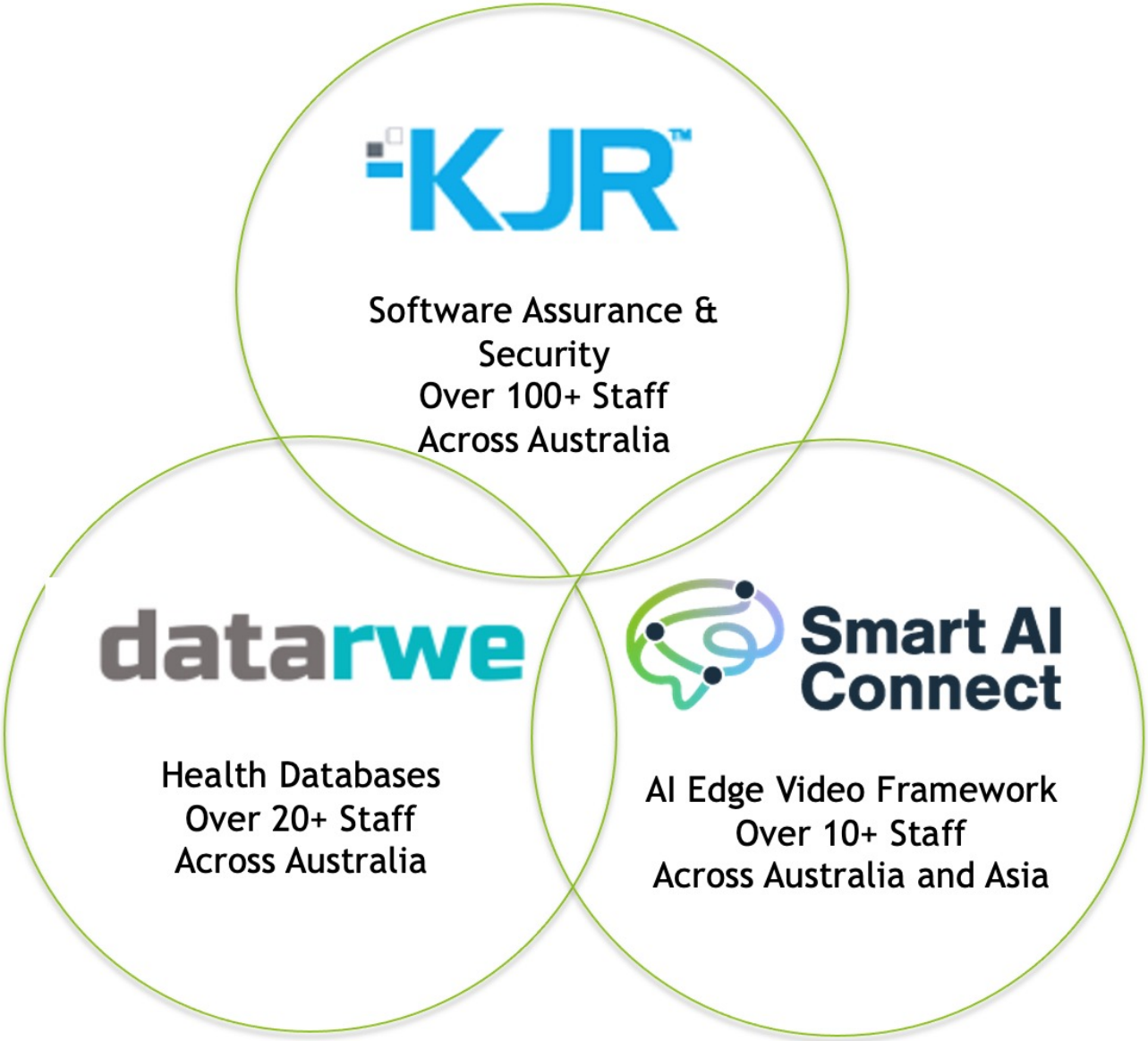
kelvin.ross@datarwe.com

@kelvinjross

Source: https://ihganalytics.sharepoint.com/:p:/s/Technology/EVqxXD0W7I1JrWzaF1WWrskB9z2hs3-aoihawjOLcoeMzw?e=waPX3H

datarwe

Real world data for AI enabled medical innovation


KJR
Software Assurance &
Security
Over 100+ Staff
Across Australia
datarwe
Health Databases
Over 20+ Staff
Across Australia
Smart AI
Connect
AI Edge Video Framework
Over 10+ Staff
Across Australia and Asia

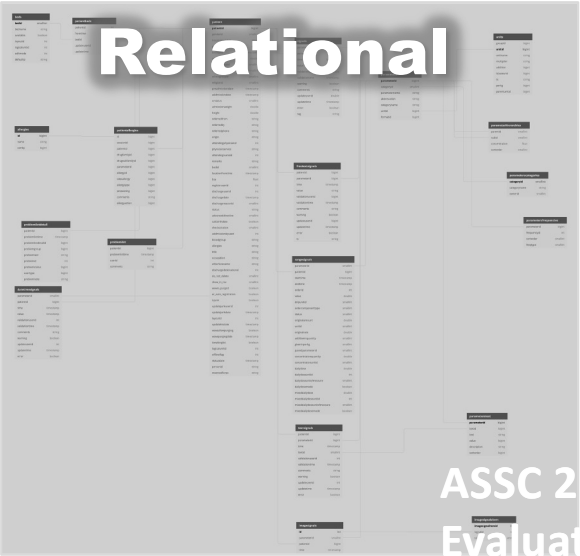
Sectors

HEALTHCARE

DEFENCE
AUSTRALIA
AUSTRALIA

INDIGENOUS

PUBLIC SAFETY


Relational
ASSC 2
Evaluat

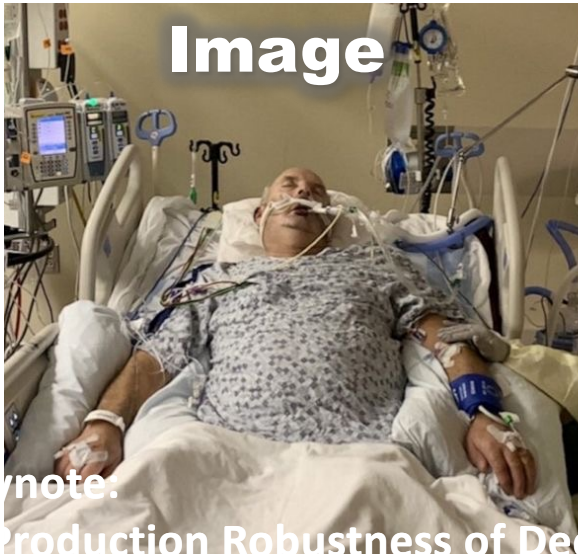
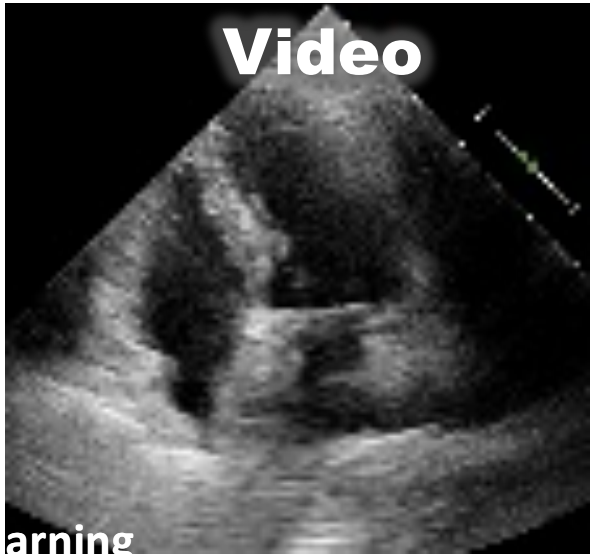
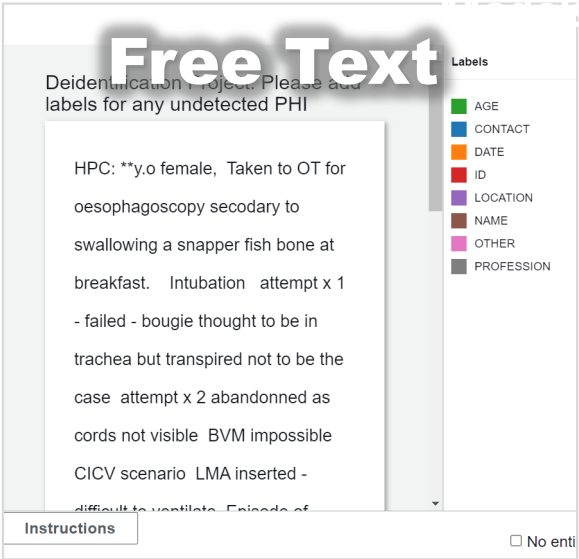
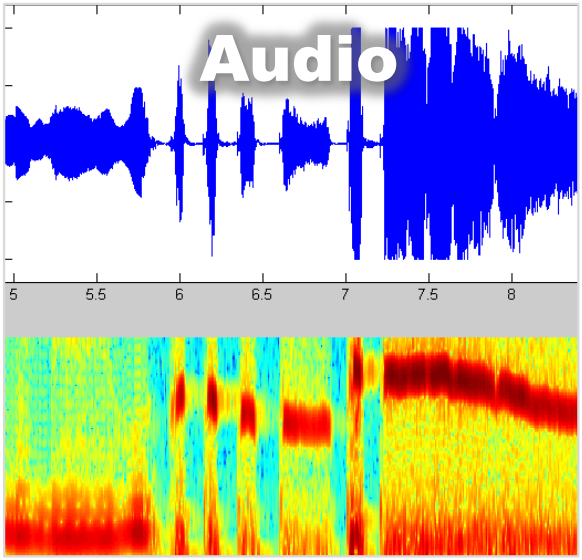
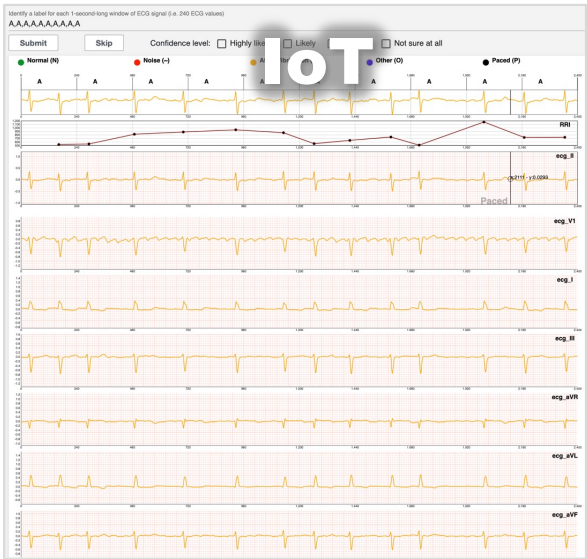


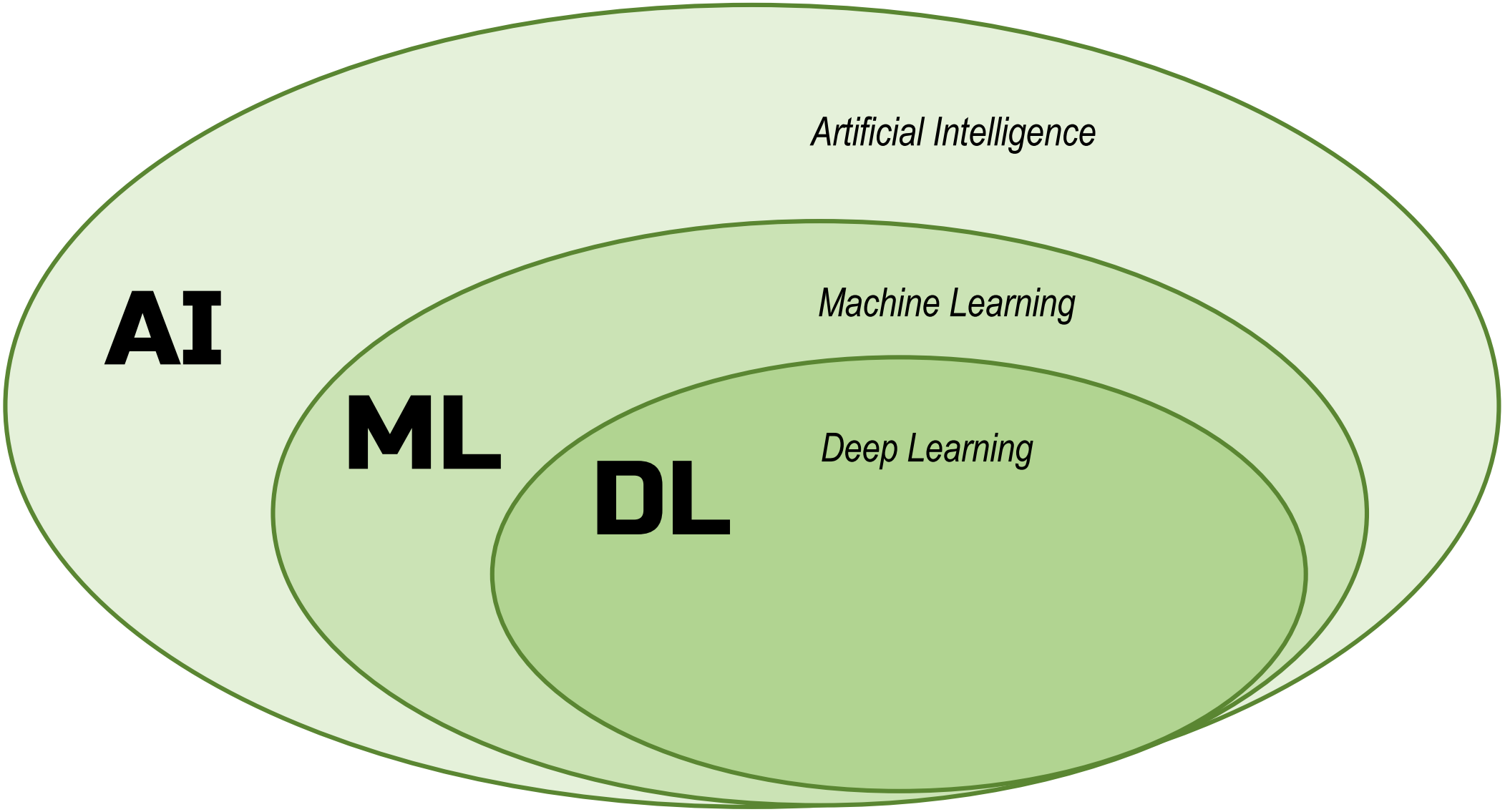
Image
ynote:
Production Robustness of De


Video
arning


Free Text
Deidentification Project. Please add
labels for any undetected PHI
HPC: **y.o female, Taken to OT for
oesophagoscopy secodary to
swallowing a snapper fish bone at
breakfast. Intubation attempt x 1
- failed - bougie thought to be in
trachea but transpired not to be the
case attempt x 2 abandonned as
cords not visible BVM impossible
CICV scenario LMA inserted -
Labels
AGE
CONTACT
DATE
ID
LOCATION
NAME
OTHER
PROFESSION
Instructions
No enti

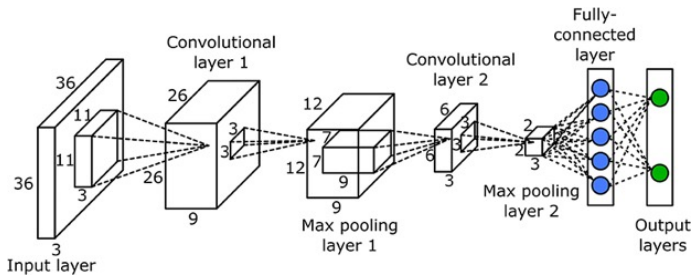

Audio
5
5.5
6
6.5
7
7.5
8


IoT
Submit
Skip
Confidence level:
Not sure at all
Normal (N)
Noise (~)
Other (O)
Paced (P)
RRI
ecg_II
Paced
ecg_V1
ecg_I
ecg_III
ecg_aVR
ecg_aVL
ecg_aVF


Artificial Intelligence
AI
Machine Learning
ML
Deep Learning
DL

DL Training

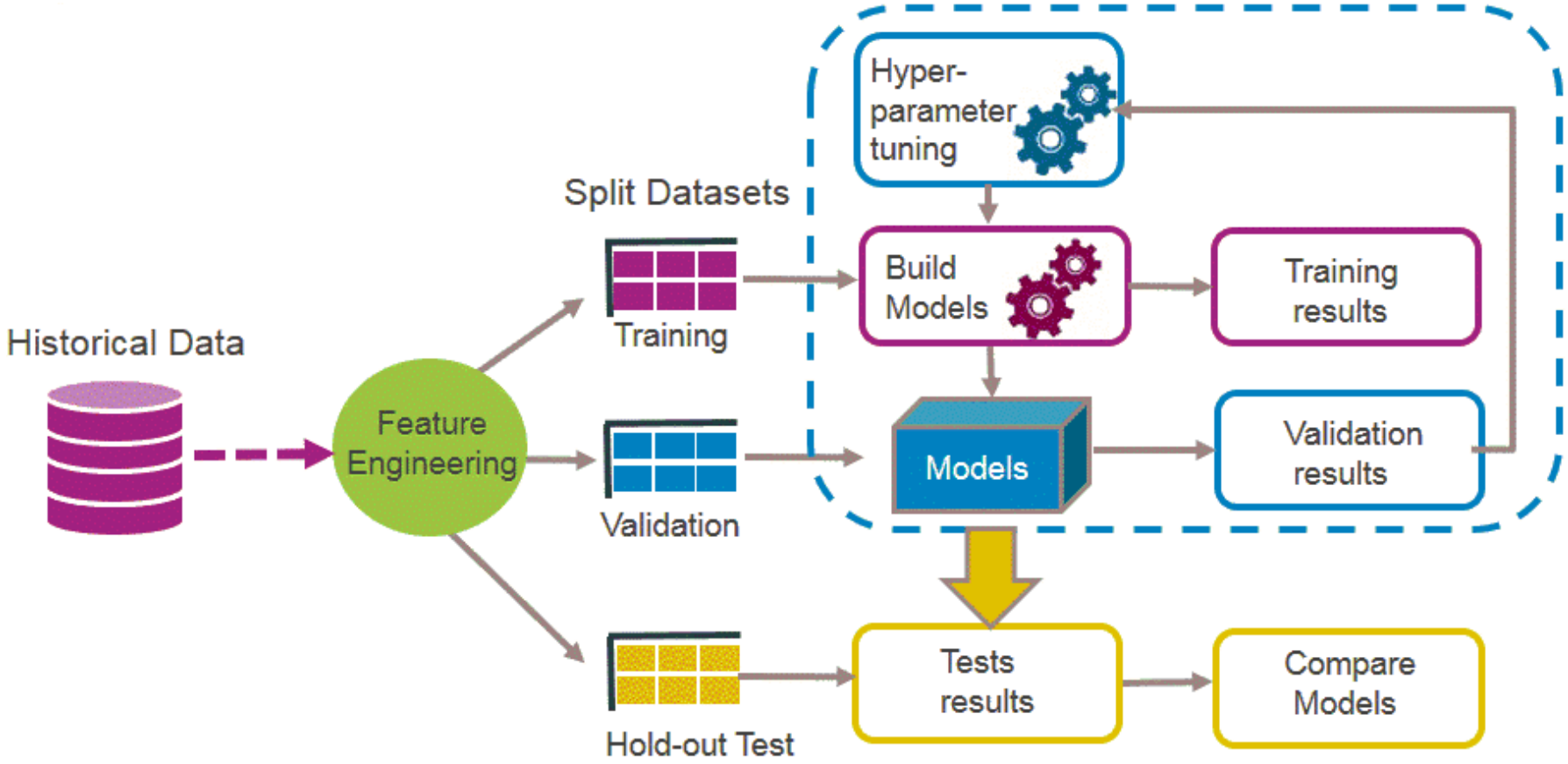
inference



back propagation

Class	Truth	Infer	Loss
Chihuahua	1.0	0.214	+0.786
Labrador	0.0	0.769	-0.769

ML Process – a.k.a. Kaggle Approach



Blinge

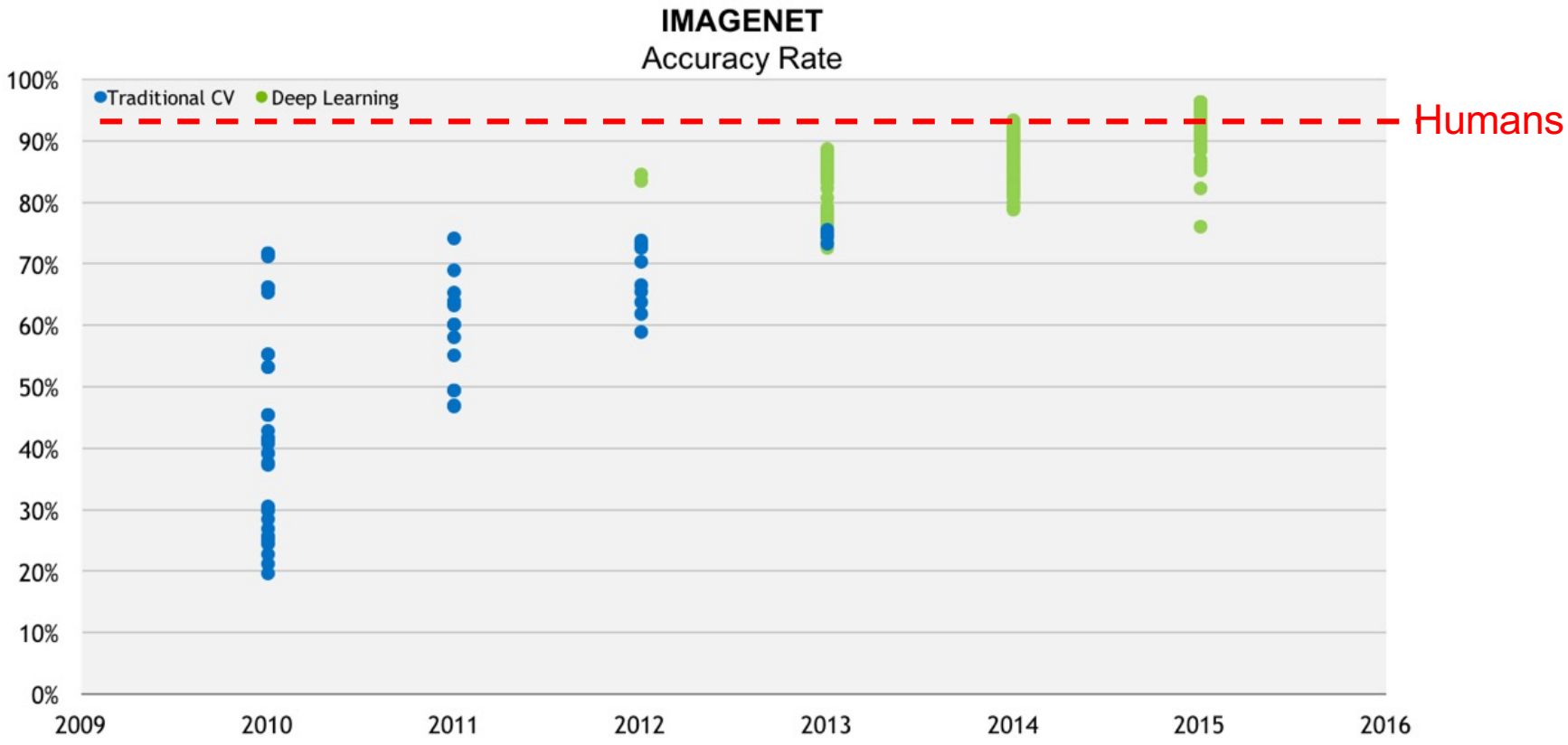
Chihuahua or Muffin

Should we use Deep Learning in Trusted Applications?

YES , but...

DEEP LEARNING FOR VISUAL PERCEPTION

Going from strength to strength



More info: https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence

Corner Cases

https://www.youtube.com/watch?v=Ucp0TTmvqOE&t=1h51m05s

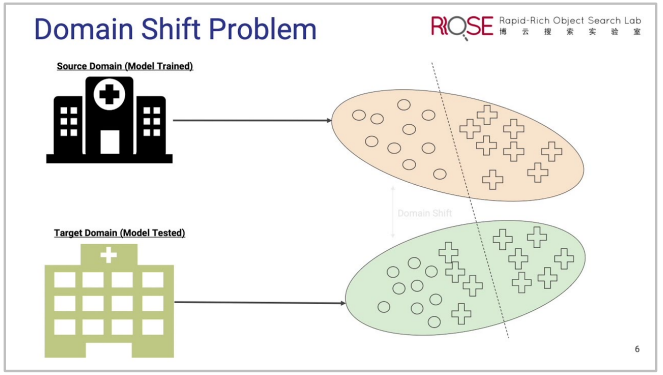
Robustness Issues

Unintended Signal Correlation

Hidden Stratification

https://youtu.be/_4gn7ibByAc

Domain Shift



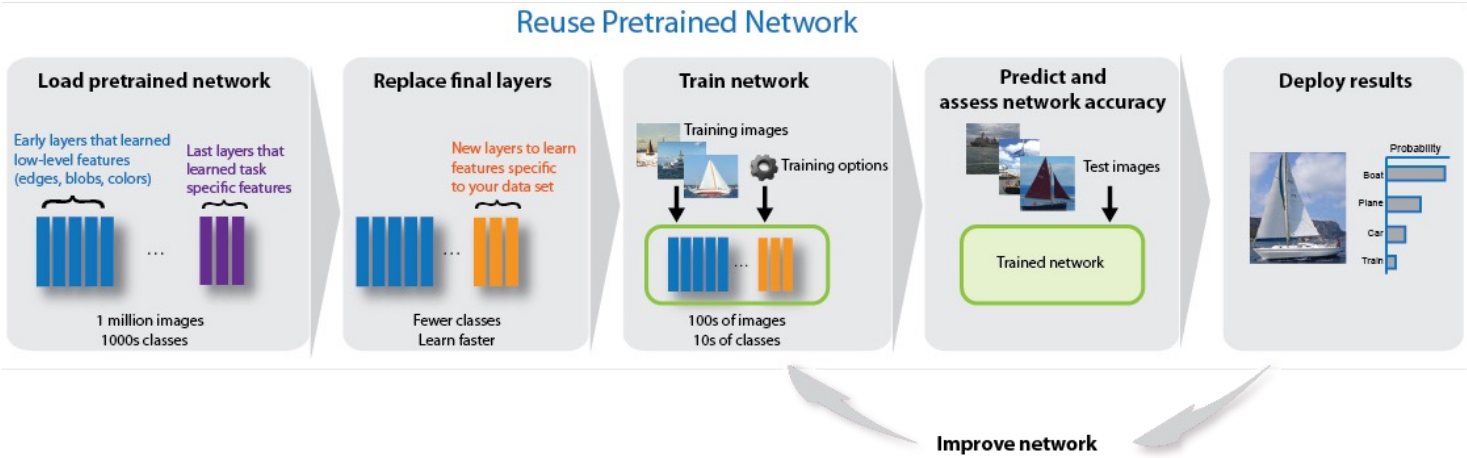
https://youtu.be/diJAM-Z6u0Y

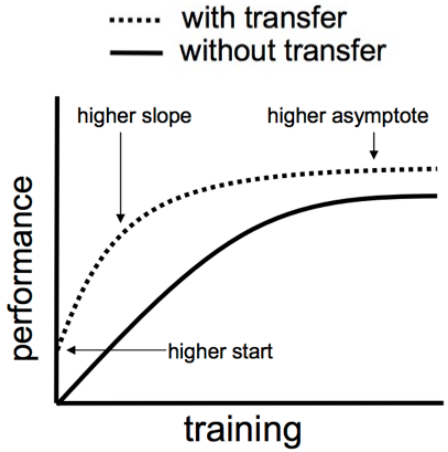
Super
Model
Sampling
Bias

Overfitted Data

Training optimised with sample biased data

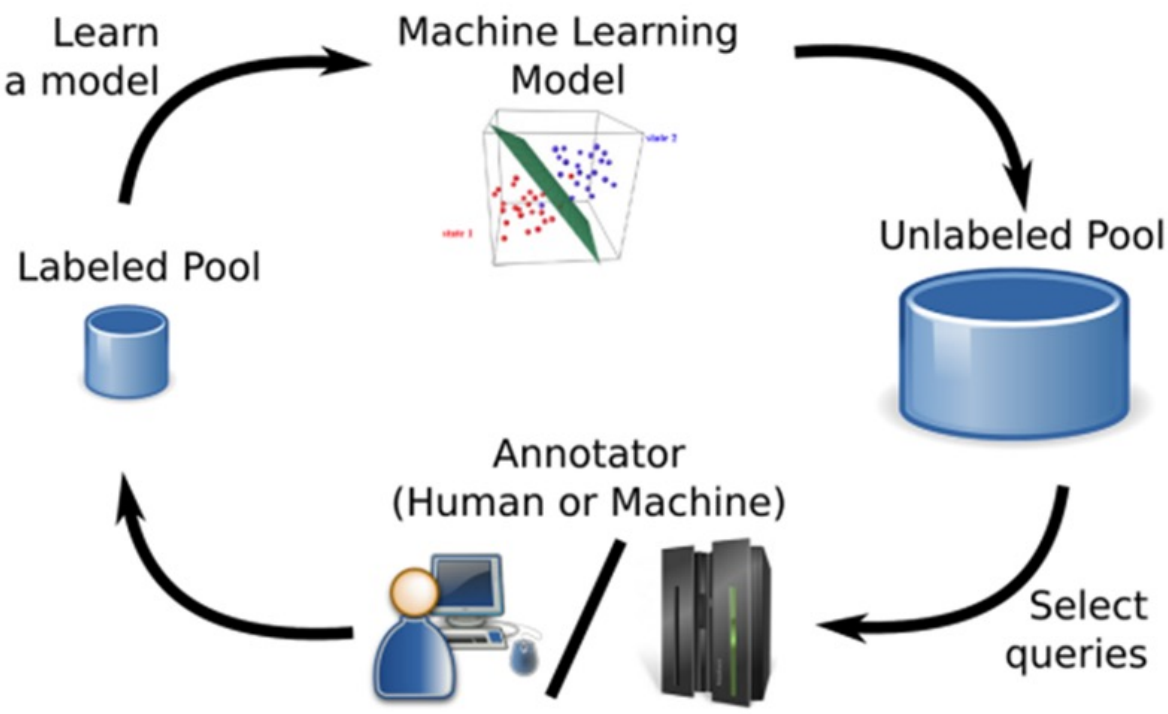
Transfer Learning

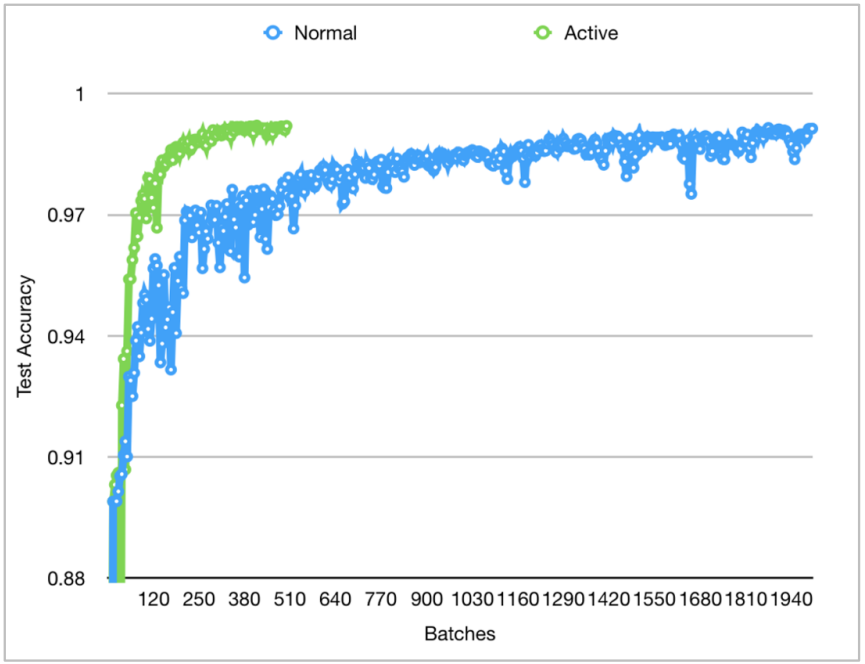




Source: https://it.mathworks.com/help/deeplearning/ug/transfer-learning-using-alexnet.html

Active Learning



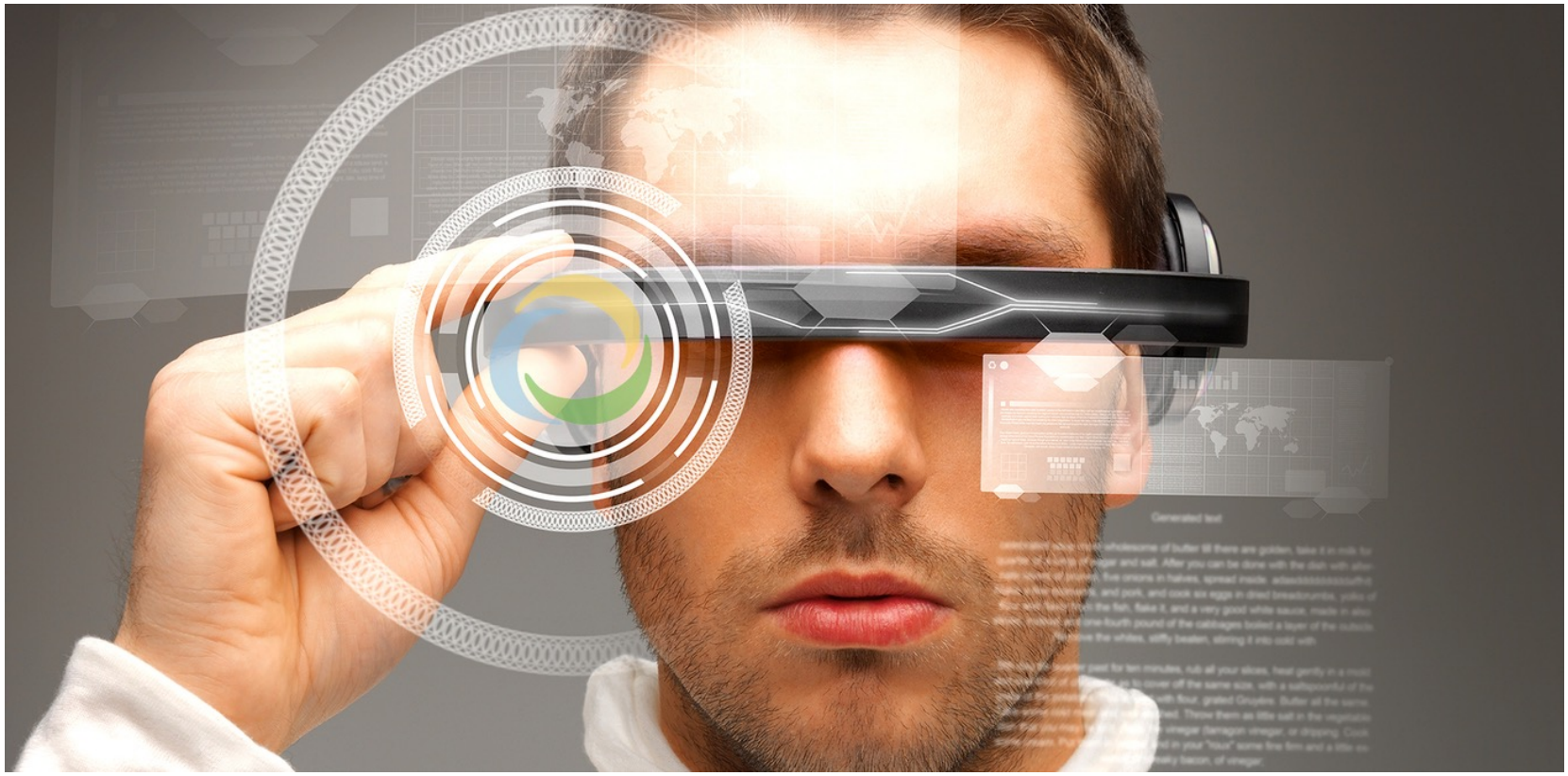


https://becominghuman.ai/accelerate-machine-learning-with-active-learning-96cea4b72fdb

- Uncertainty
- Decision Boundaries
- Similarity

https://www.datacamp.com/community/tutorials/active-learning

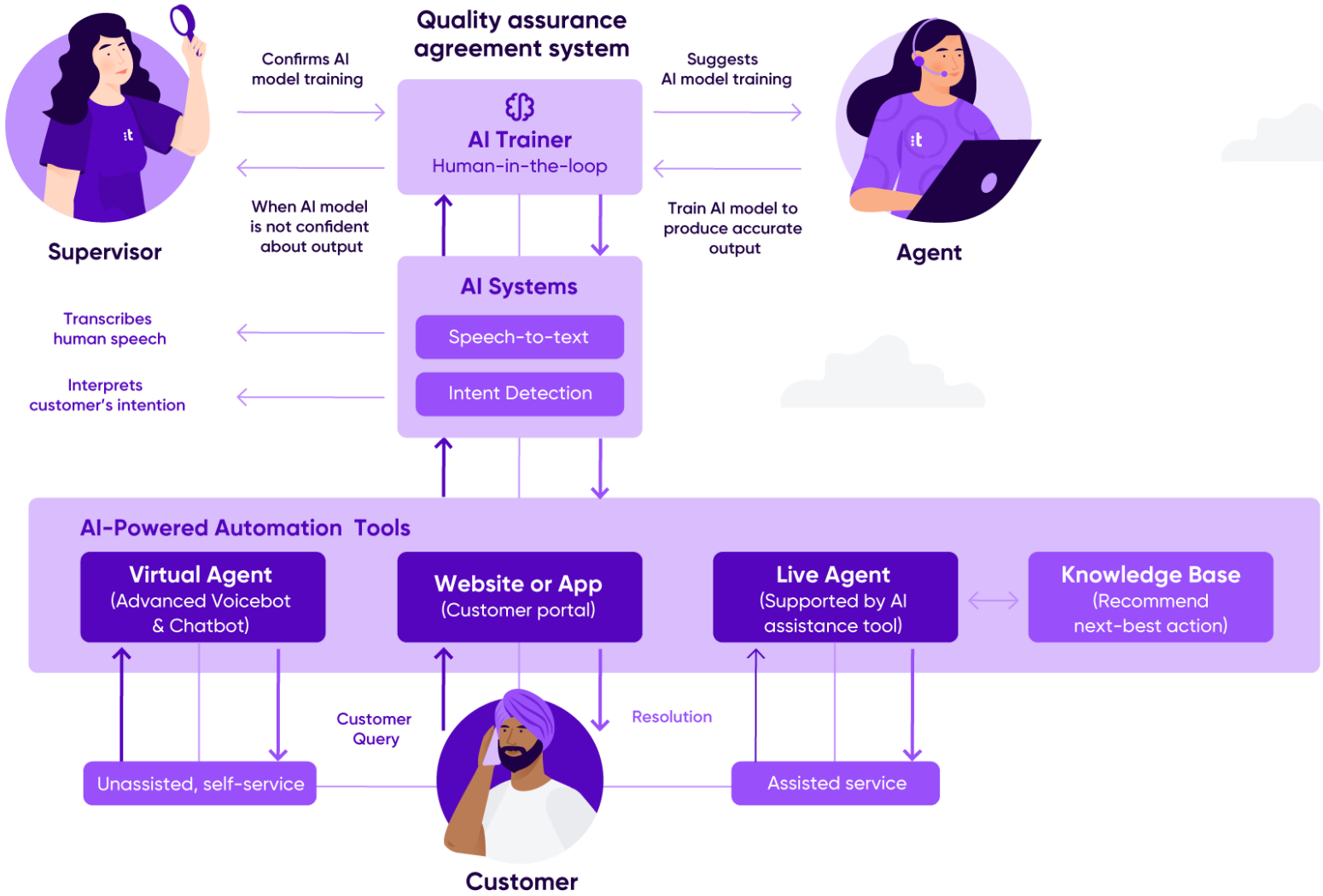
Human Augmentation



Error Rate in detection of cancer in lymph node cells	Human Pathologist	AI	Human Pathologist + AI
	3.5%	**7.5%**	**0.5%**

Source: Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Andrew H. Beck, “Deep Learning for Identifying Metastatic Breast Cancer,” June 18, 2016, https://arxiv.org/pdf/1606.05718v1.pdf.

Human-In-The-Loop



FDA: Artificial Intelligence and Machine Learning in Software as a Medical Device

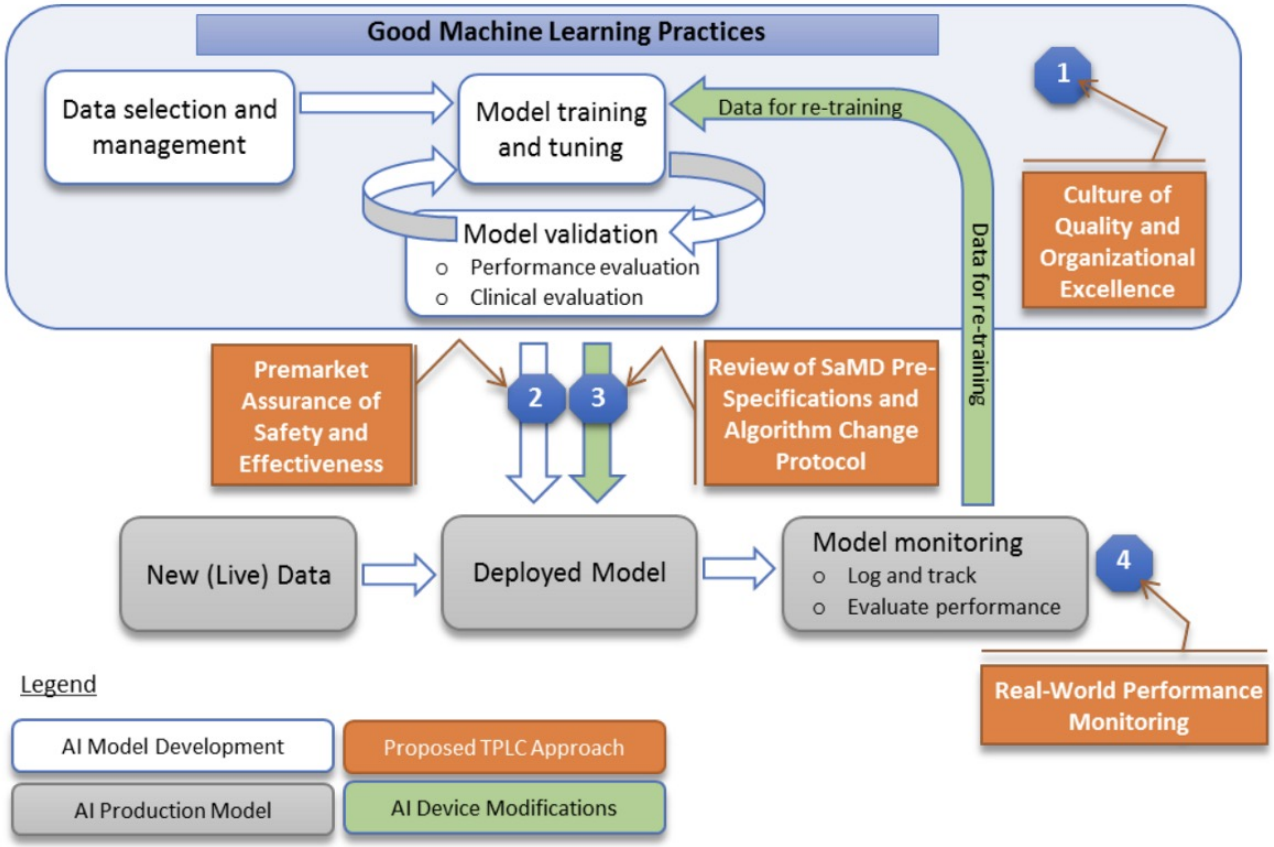
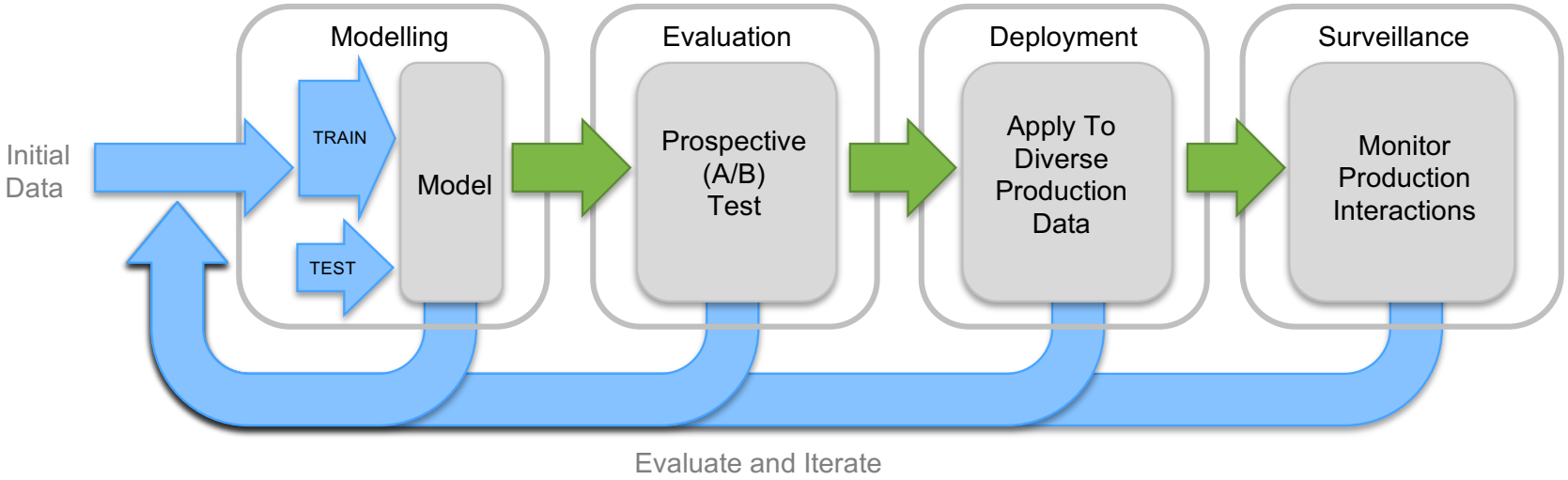


Figure 2: Overlay of FDA's TPLC approach on AI/ML workflow

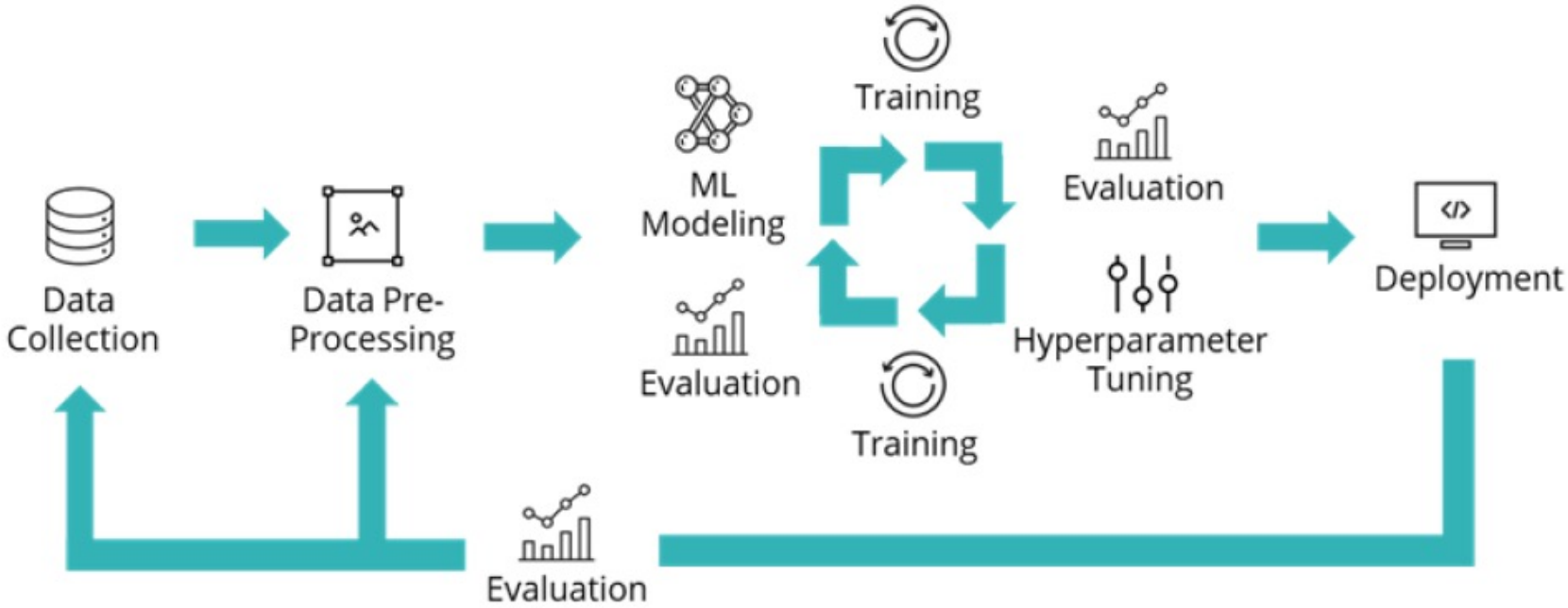
https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Clinical Trial Approach

Phase 1	Phase 2	Phase 3	Phase 4
Lab	Single Site	Multi Site	Market Surveillance

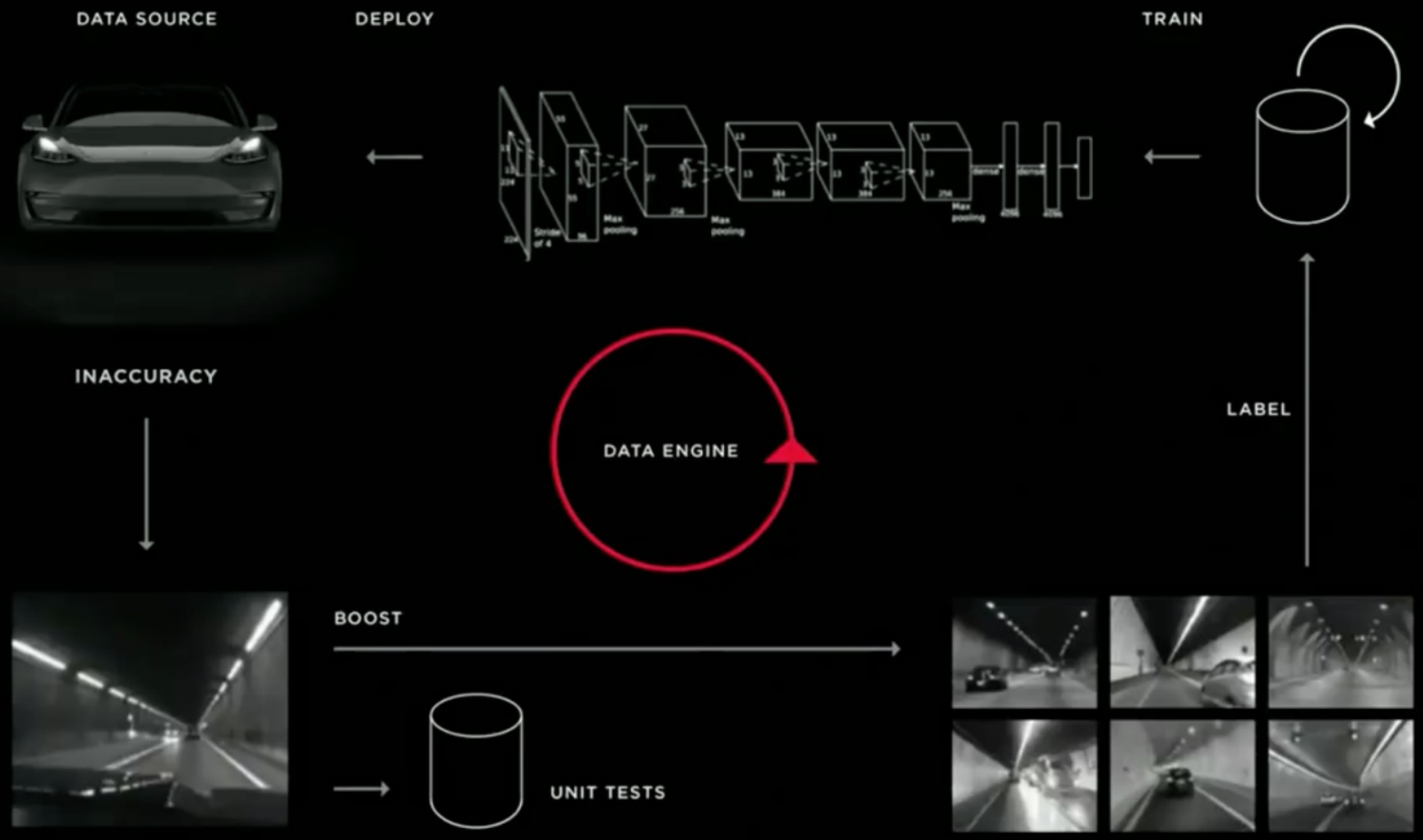


Data-Centric AI



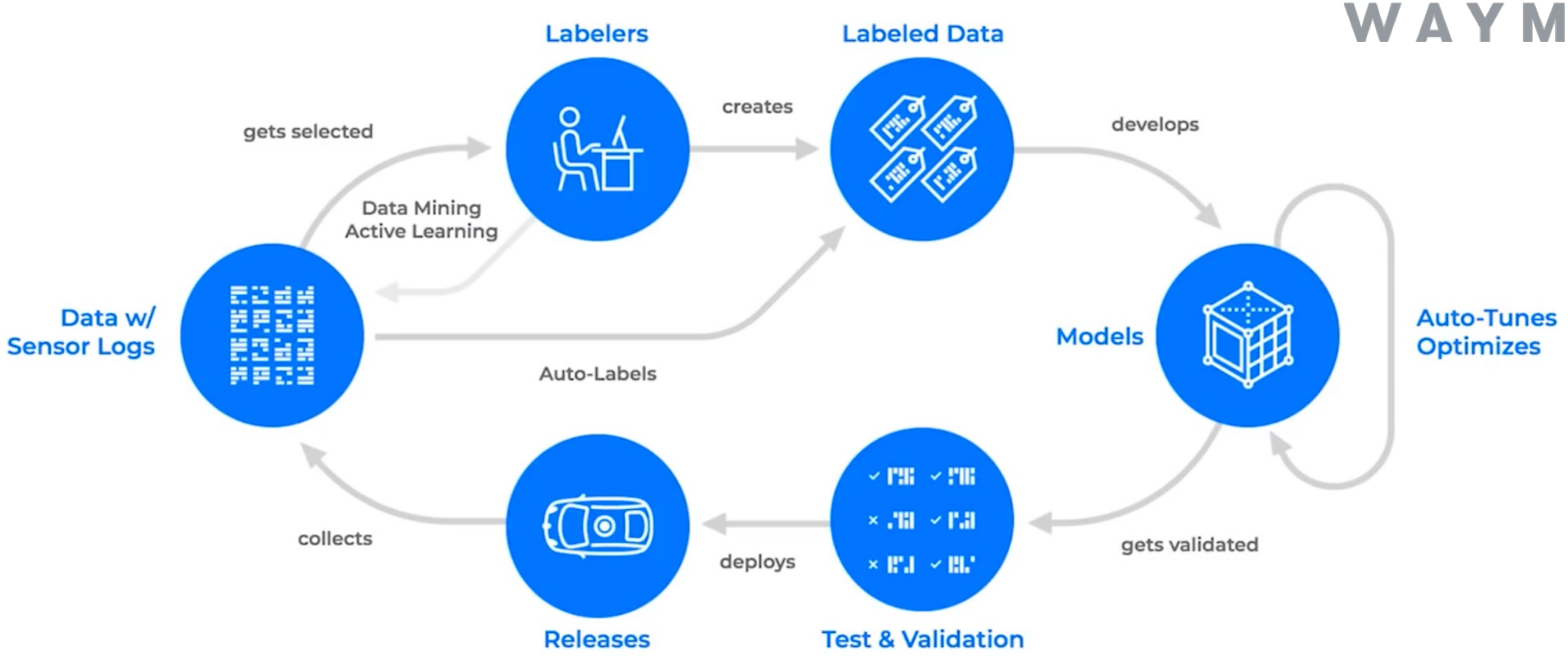
https://dida.do/blog/data-centric-machine-learning

Tesla Data Engine



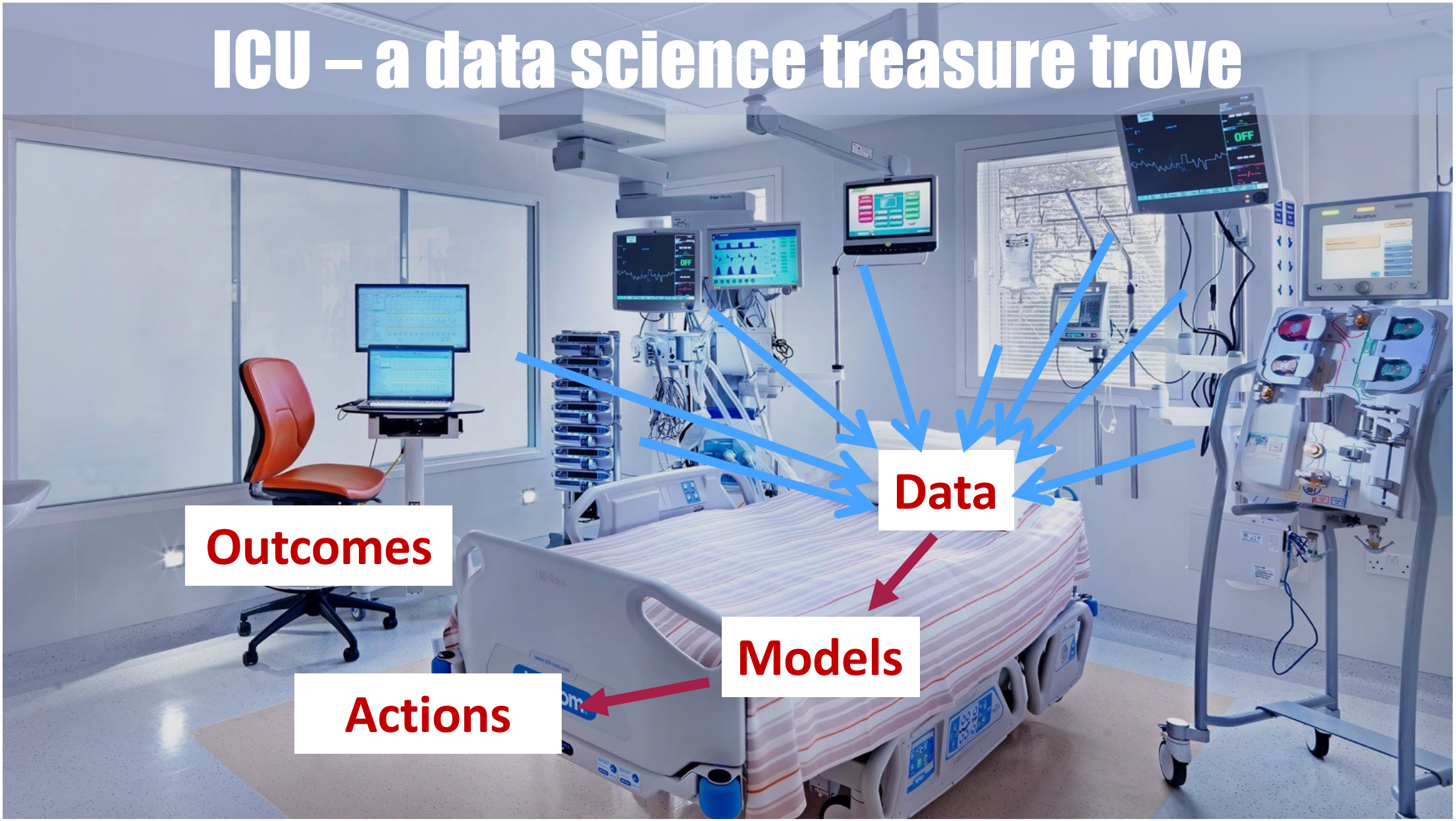
https://youtu.be/j0z4FweCy4M

ML Factory For Self Driving Models



https://www.youtube.com/watch?v=Q0nGo2-y0xY

ICU – a data science treasure trove



Unlocking Secondary Use of Medical Data

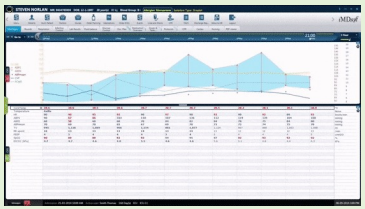
17 Hospitals **155k+** patient episodes **780m** minutes of observations

Digital Twin

Metavision

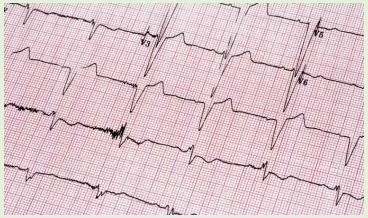
Streaming Device data

- Demographics
- Pre-Admission Data
- Time-Series Physiologic Data
- Lab Results
- Orders
- Medications & Fluids
- Nurses Notes & Observations
- Images
- Outcomes
- High Frequency Waveform Data



Metavision (GCUH)

- 50 beds – avg 25 beds occupied
- 20k episodes - 8 years
- 2.5k episodes p.a
- Median 10k signals p.episode
- Storage: 0.5 Mb p.episode

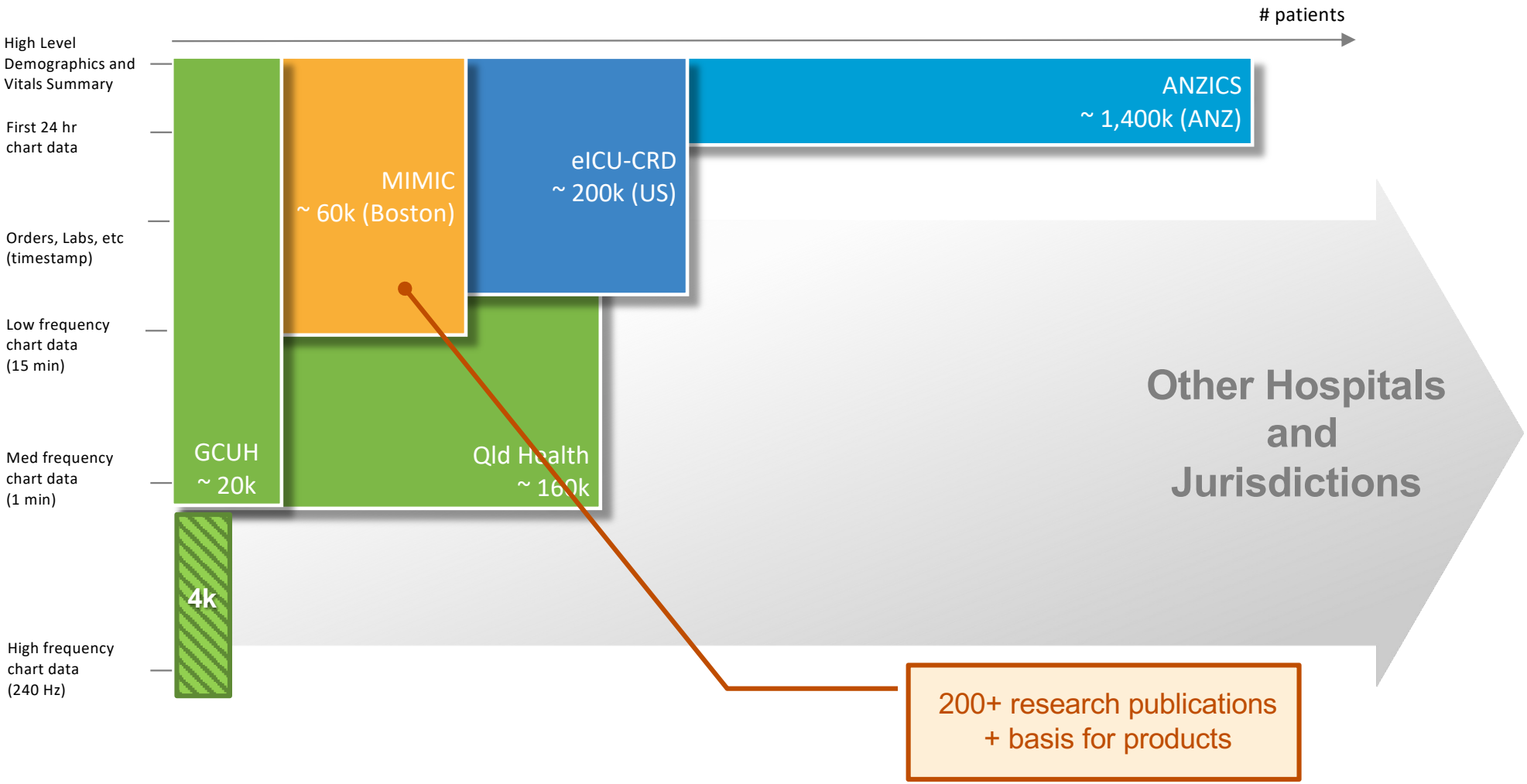


Bedside Monitors (GCUH)

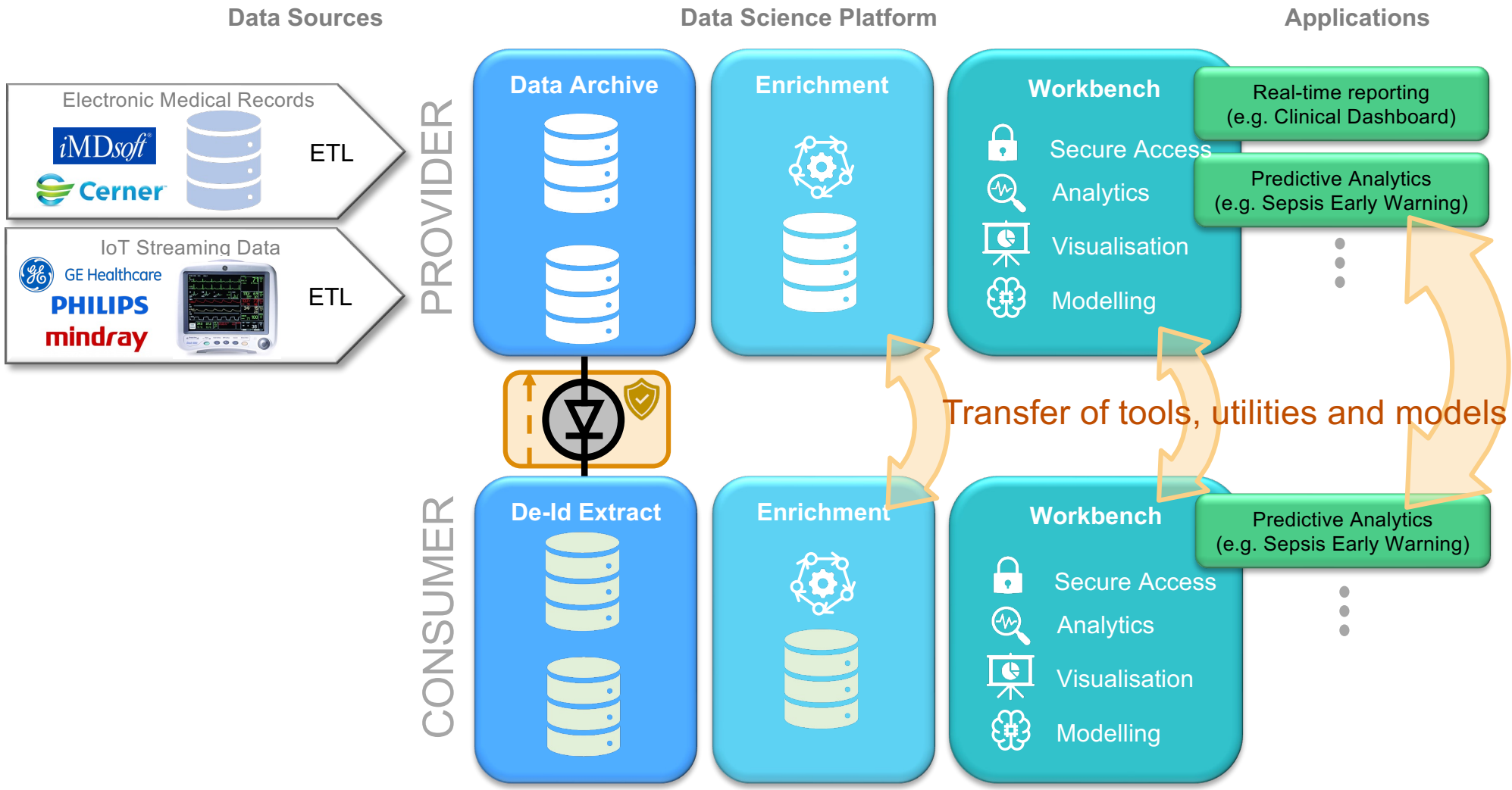
- 2k episodes - 1 year
- 140k hrs of ECG
- 800b data points
- + BP, SpO2
- Storage: 2.2 Gb p.bed/p.month

as of May 2022

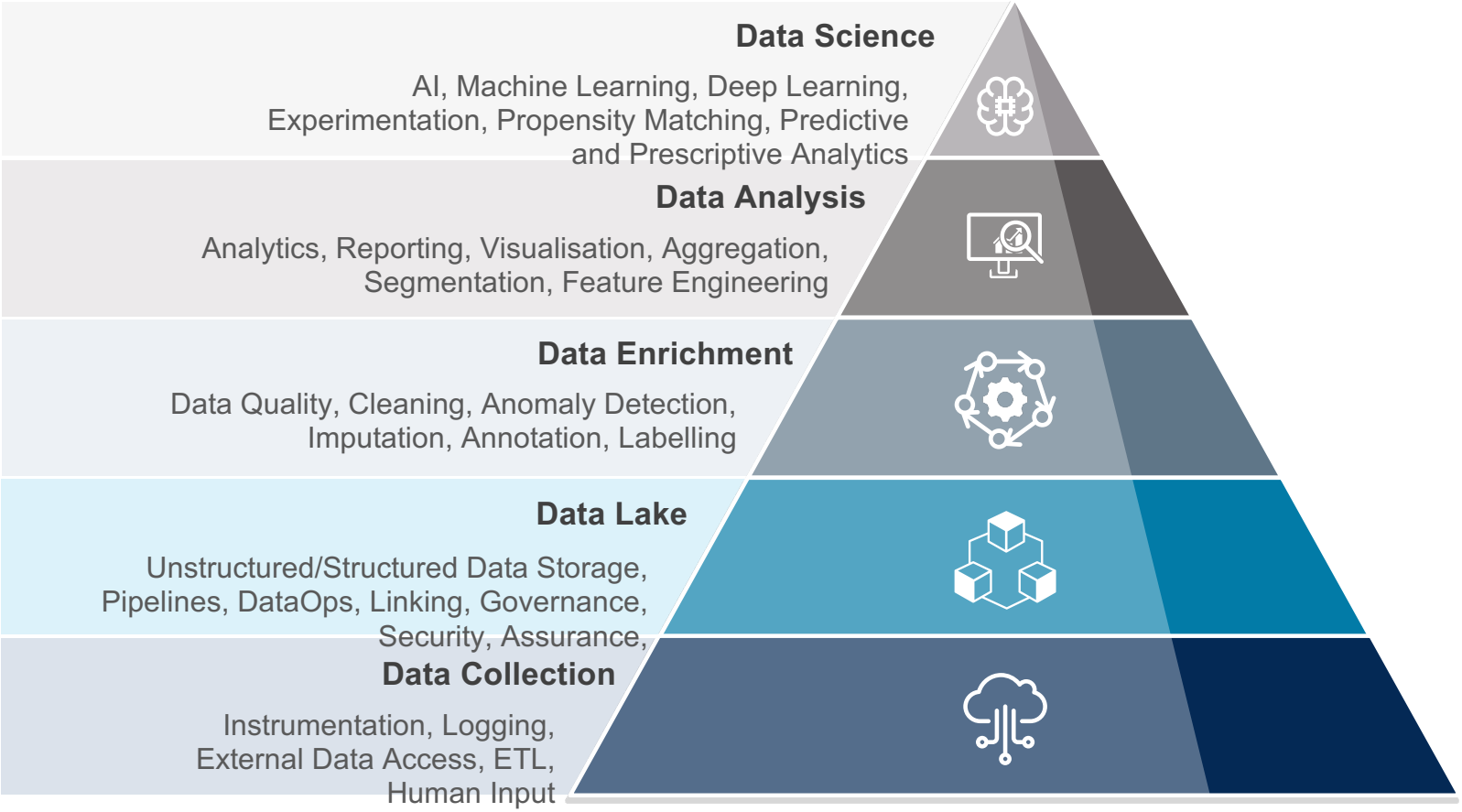
Global ICU Data Sets



Platform Architecture



Data Science Hierarchy



powered by aws

Kubeflow

Sagemaker
Sagemaker Ground Truth
Comprehend [Medical]

Athena

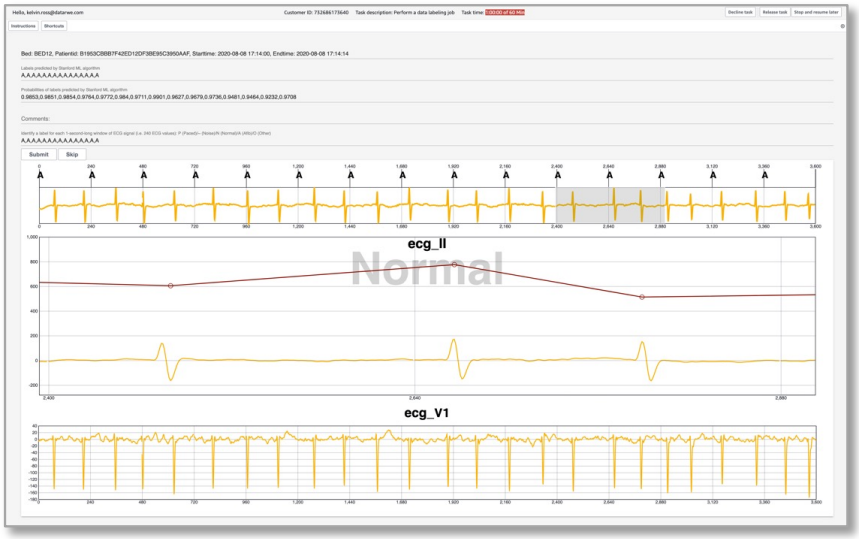
Lambda

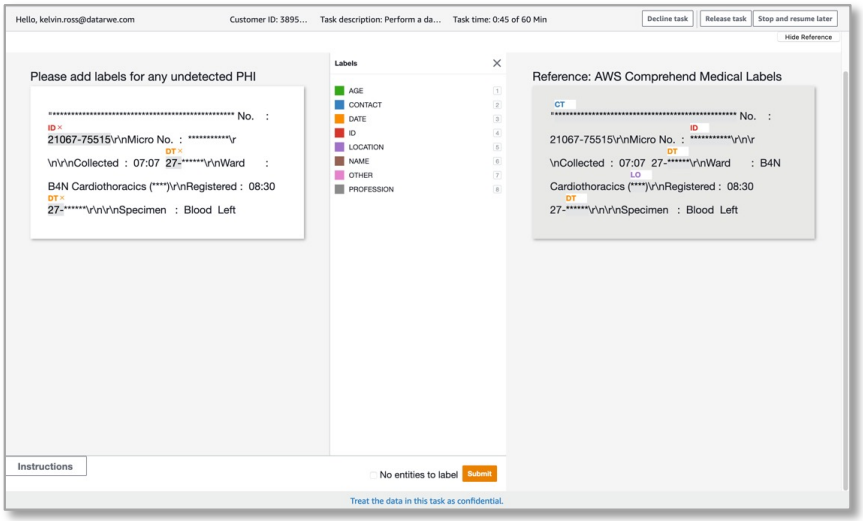
S3 (CSV, Parquet)

Ground Truth Labeling

Labeling tasks undertaken with different levels of clinical expertise:

Task	Generic Skill	Nurse, Medical Technician	Doctor	Consultant
BP Anomalies	●			
PHI Testing	●			
RR intervals	●			
ECG Arrythmias		●	●	●
Named Entity Recognition		●		
Event Time Verification		●		
Phenotype Classification		●	●	



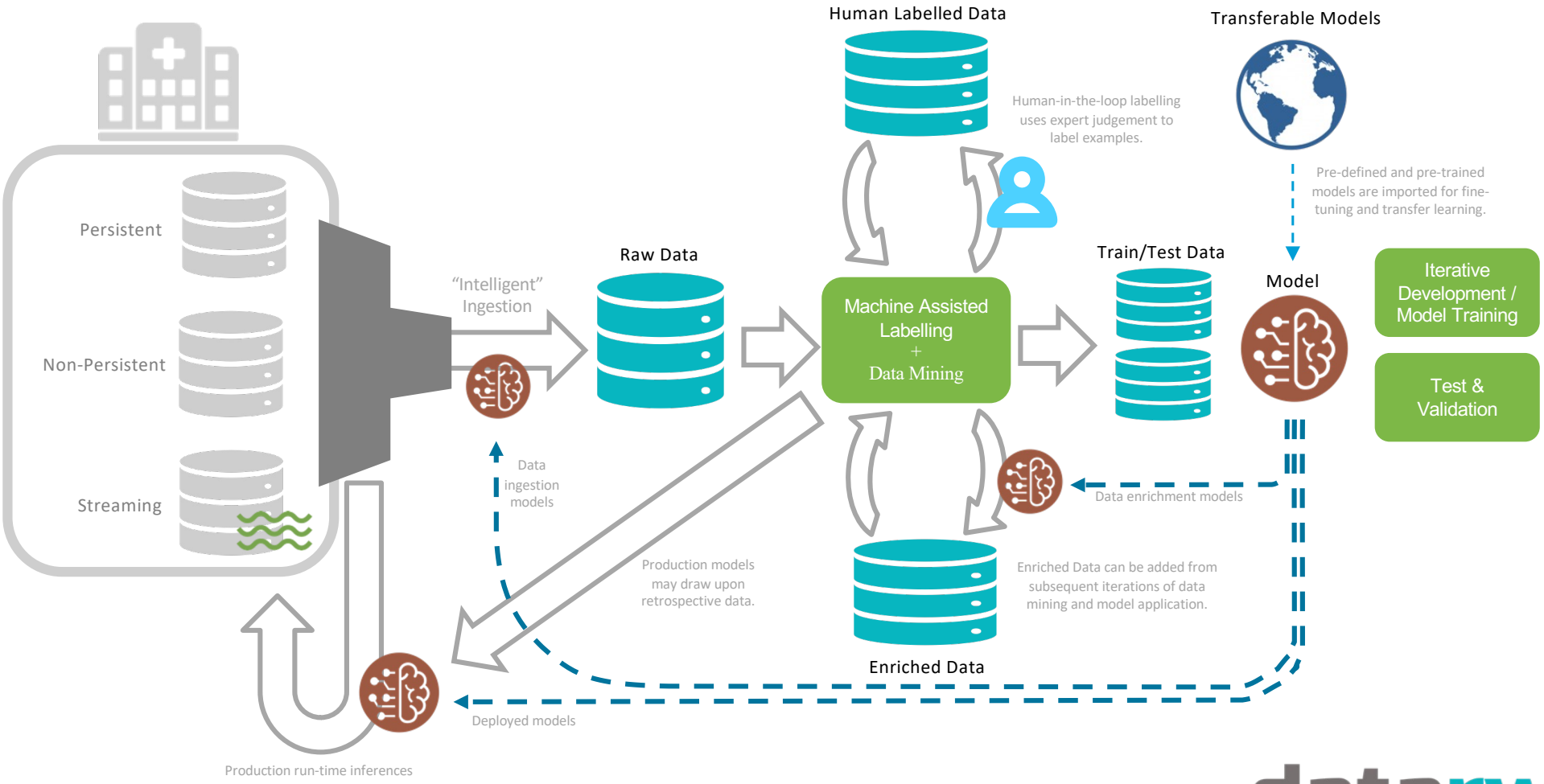


Sagemaker
Sagemaker Ground Truth
Comprehend [Medical]

Cognito

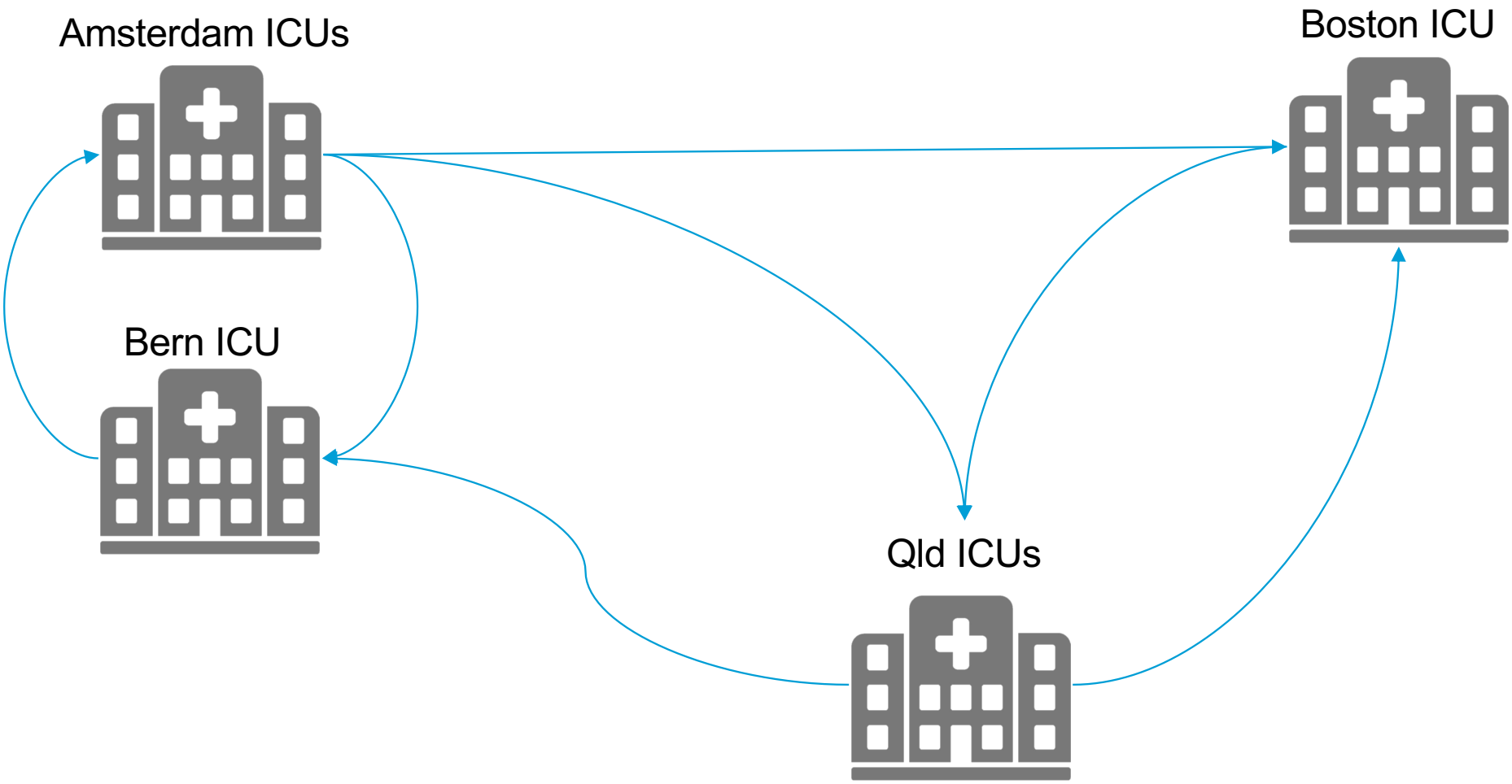
https://ihqanalytics.sharepoint.com/sites/Technology/Shared Documents/AnnotationPlatforms/Annotation With AWS Sagemaker Ground Truth.docx

Datarwe Data Engine

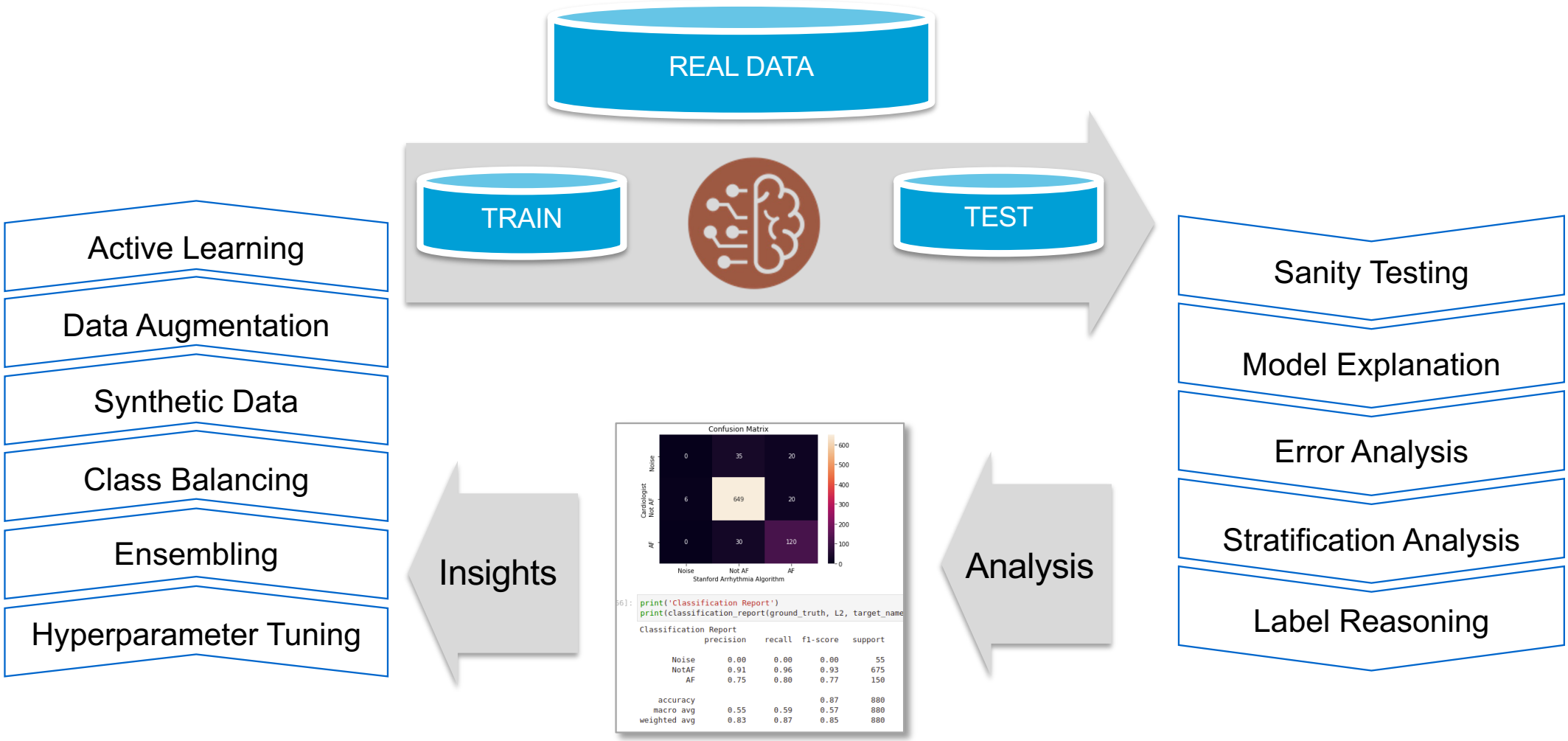


datarwe

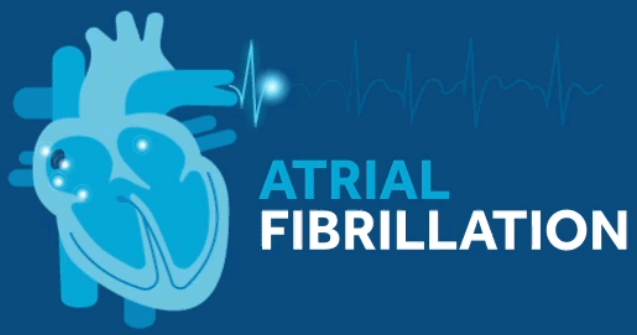
Transfer Learning



Validation Driven Machine Learning



Atrial Fibrillation



Post-Operative Atrial Fibrillation (POAF) complications[1]:

↑ 12-24hrs ICU ()

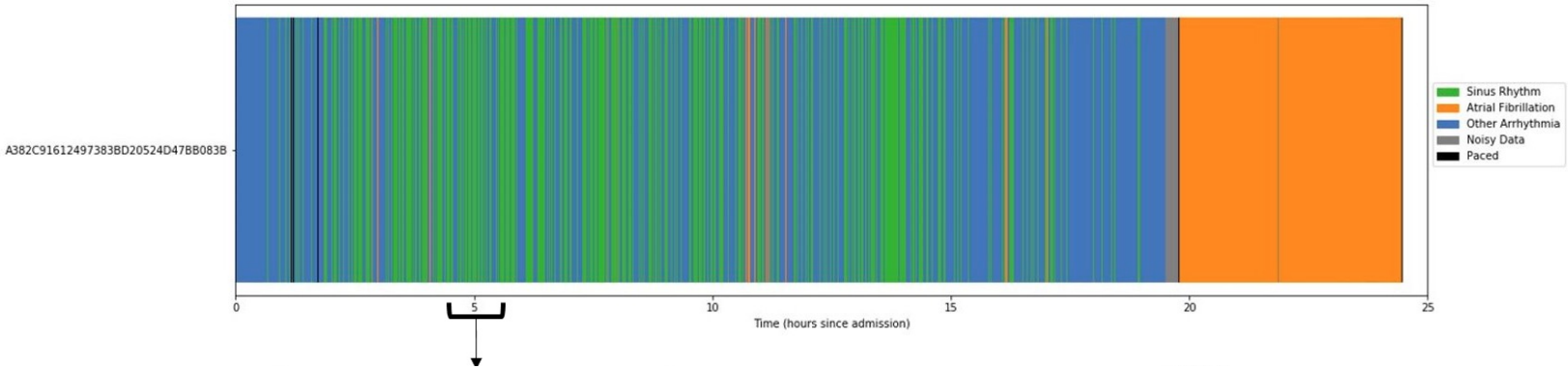
↑ 2-5 days in hospital

Increased risk of:

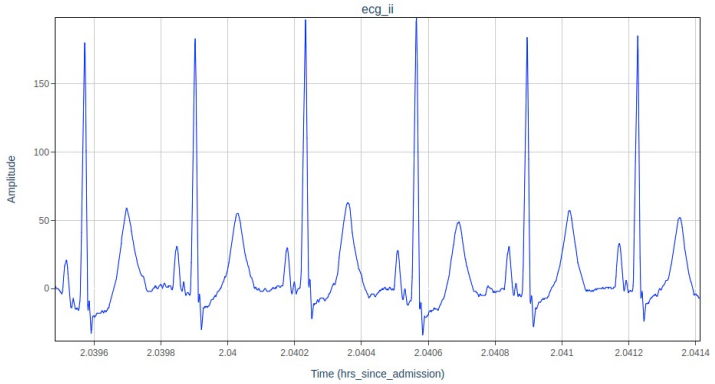
- Stroke
- Reoperation for bleeding
- Infection
- Renal or respiratory failure
- Cardiac arrest
- Cerebral complications

[1]Jason W. Greenberg, Timothy S. Lancaster, Richard B. Schuessler, Spencer J. Melby, Postoperative atrial fibrillation following cardiac surgery: a persistent complication, *European Journal of Cardio-Thoracic Surgery*, Volume 52, Issue 4, October 2017, Pages 665–672, https://doi.org/10.1093/ejcts/ezx039

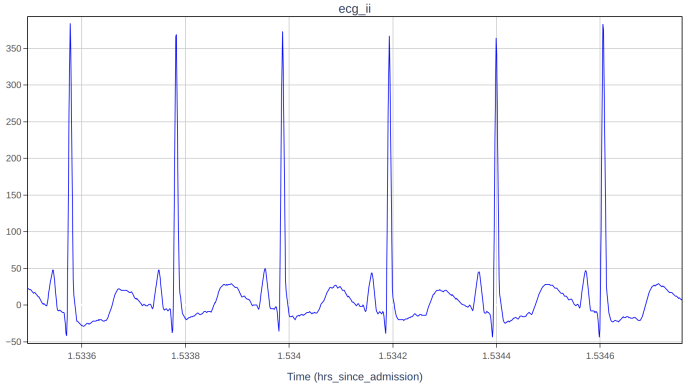
Which Patient will experience POAF?



Later AF

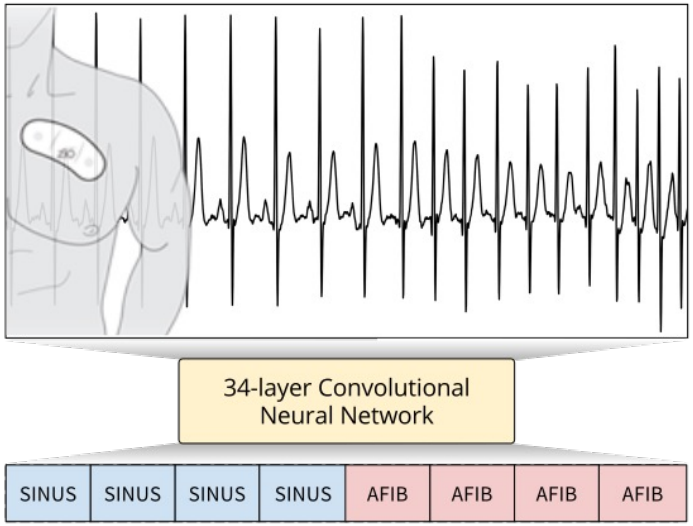


No Later AF



Cardiologist-level arrhythmia detection with CNNs

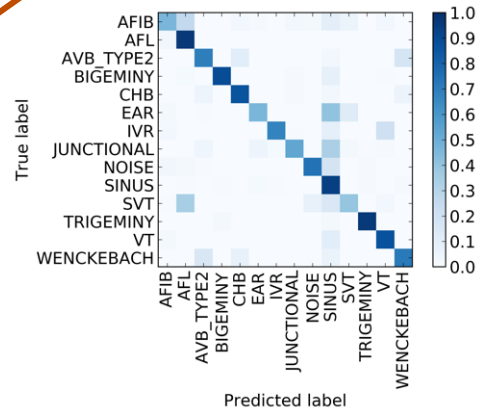
- Andrew Ng's Stanford group
- 34 layer CNN to diagnose arrhythmias on 64,000 30sec ECGs
- Model outperformed cardiologists



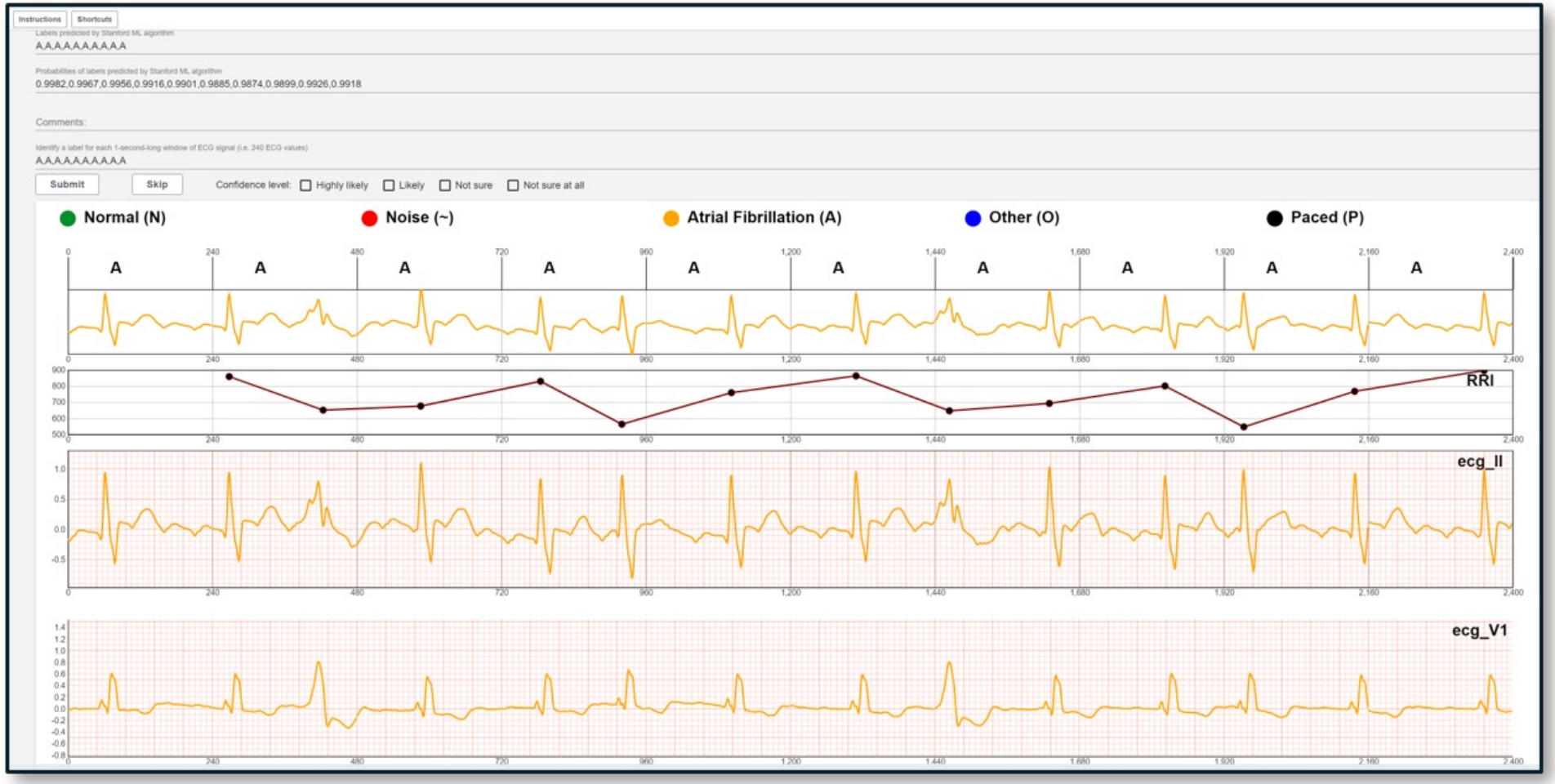
https://stanfordmlgroup.github.io/projects/ecg/

	Seq		Set	
	Model	Cardiol.	Model	Cardiol.
Class-level F1 Score				
AFIB	**0.604**	0.515	**0.667**	0.544
AFL	**0.687**	0.635	**0.679**	0.646
AVB_TYPE2	**0.689**	0.535	**0.656**	0.529
BIGEMINY	**0.897**	0.837	**0.870**	0.849
CHB	**0.843**	0.701	**0.852**	0.685
EAR	**0.519**	0.476	**0.571**	0.529
IVR	**0.761**	0.632	**0.774**	0.720
JUNCTIONAL	0.670	**0.684**	**0.783**	0.674
NOISE	**0.823**	0.768	**0.704**	0.689
SINUS	**0.879**	0.847	**0.939**	0.907
SVT	**0.477**	0.449	**0.658**	0.556
TRIGEMINY	**0.908**	0.843	**0.870**	0.816
VT	0.506	**0.566**	0.694	**0.769**
WENCKEBACH	**0.709**	0.593	**0.806**	0.736
Aggregate Results				
Precision (PPV)	**0.860**	0.723	**0.809**	0.763
Recall (Sensitivity)	**0.784**	0.724	**0.827**	0.744
F1	**0.776**	0.719	**0.809**	0.751

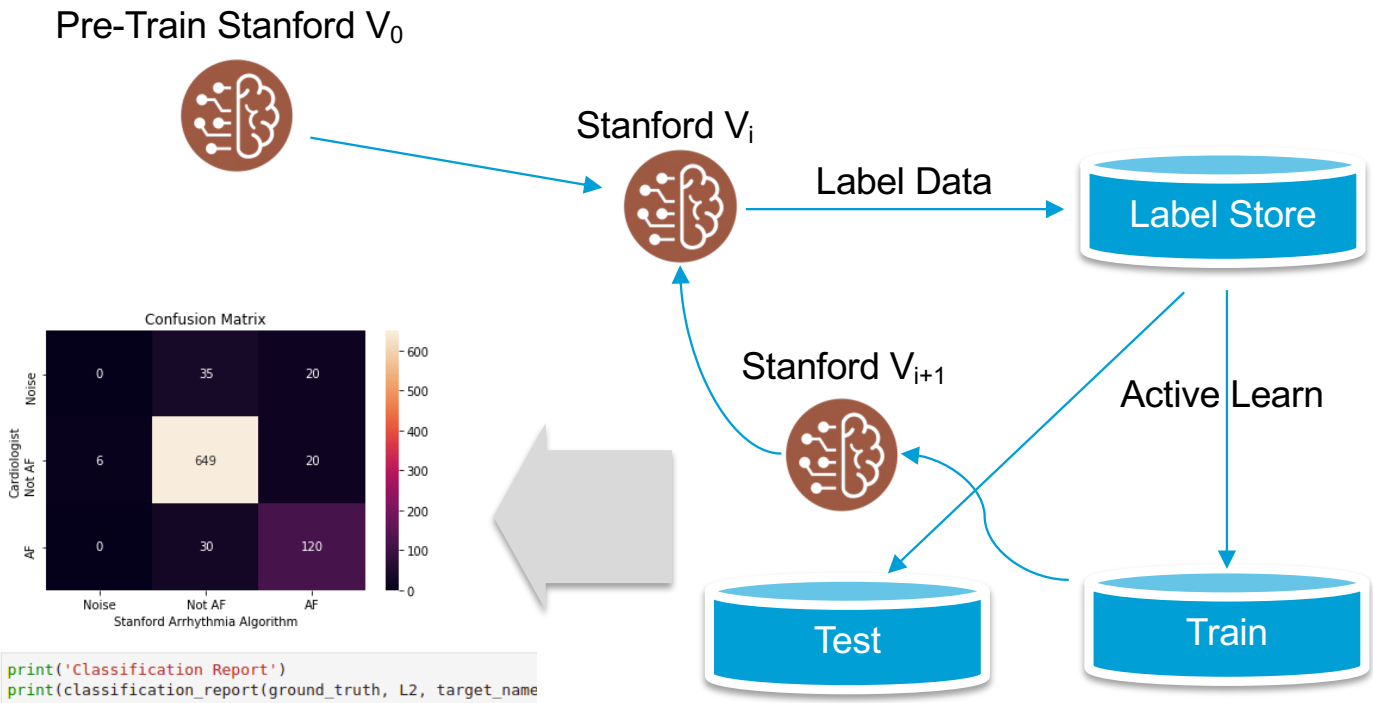
Compare human/expert performance



ECG Labelling Tool

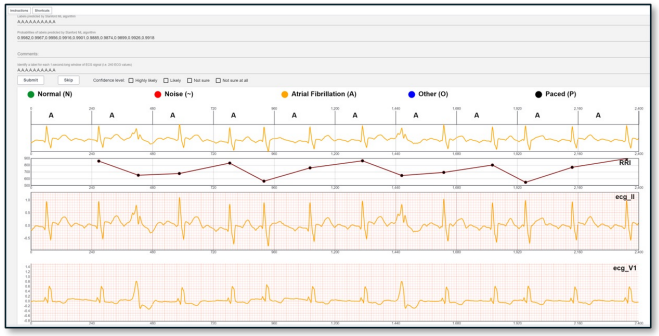


Iterative Improvement

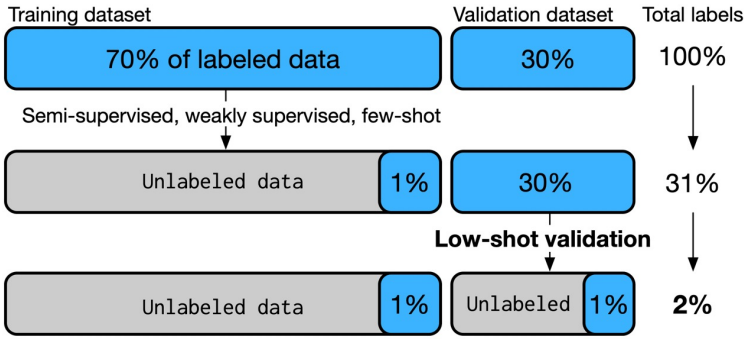


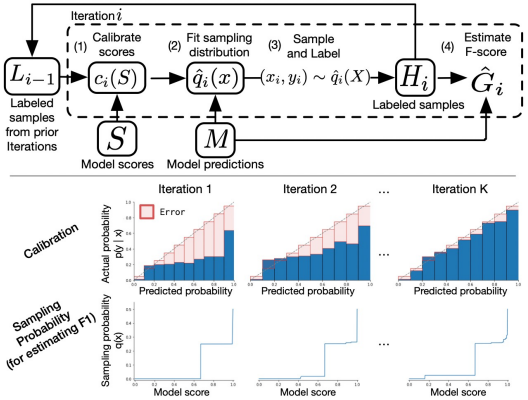

```
56]: print('Classification Report')
     print(classification_report(ground_truth, L2, target_name
```

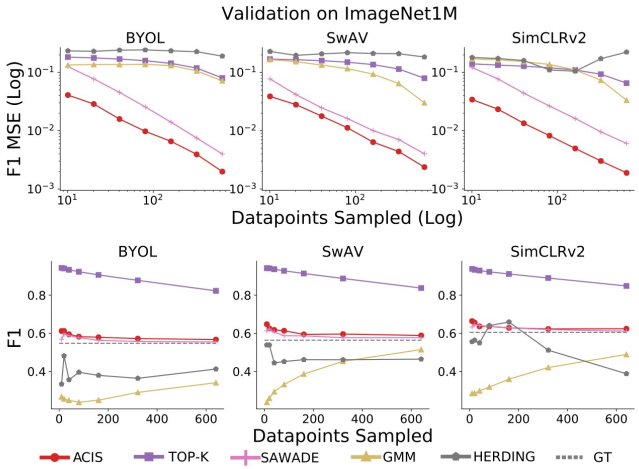
```
Classification Report
              precision    recall  f1-score   support

       Noise       0.00      0.00      0.00        55
       NotAF       0.91      0.96      0.93       675
          AF       0.75      0.80      0.77       150

    accuracy                           0.87       880
   macro avg       0.55      0.59      0.57       880
weighted avg       0.83      0.87      0.85       880
```



Low-Shot Validation: Active Importance Sampling for Estimating Classifier Performance on Rare Categories







https://arxiv.org/abs/2109.05720

Snorkel - Weak Supervision

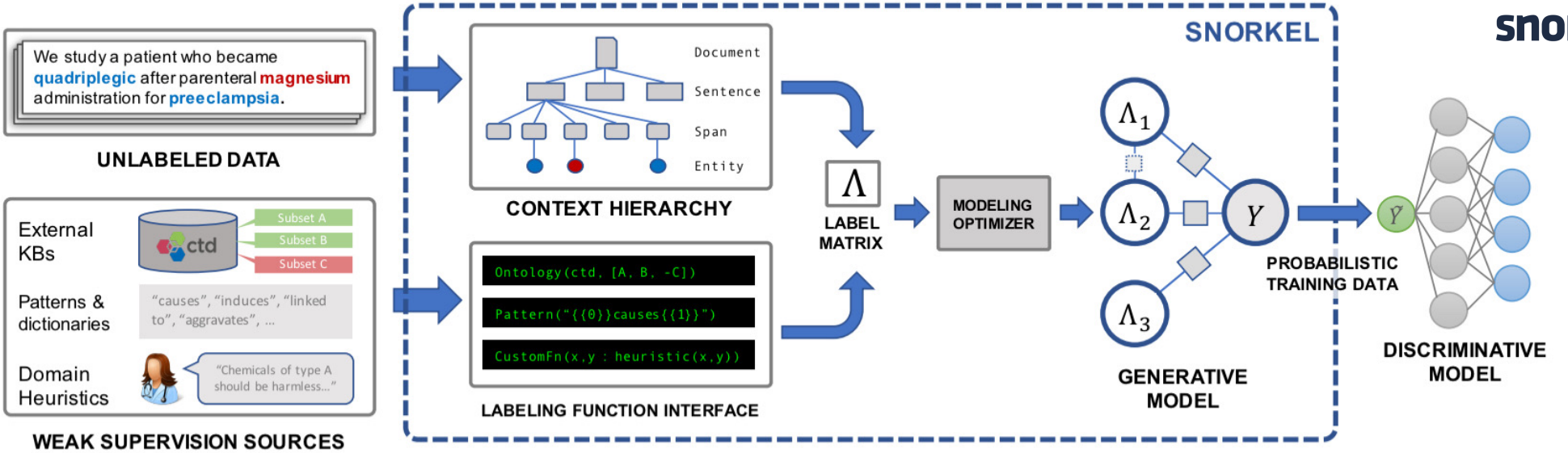
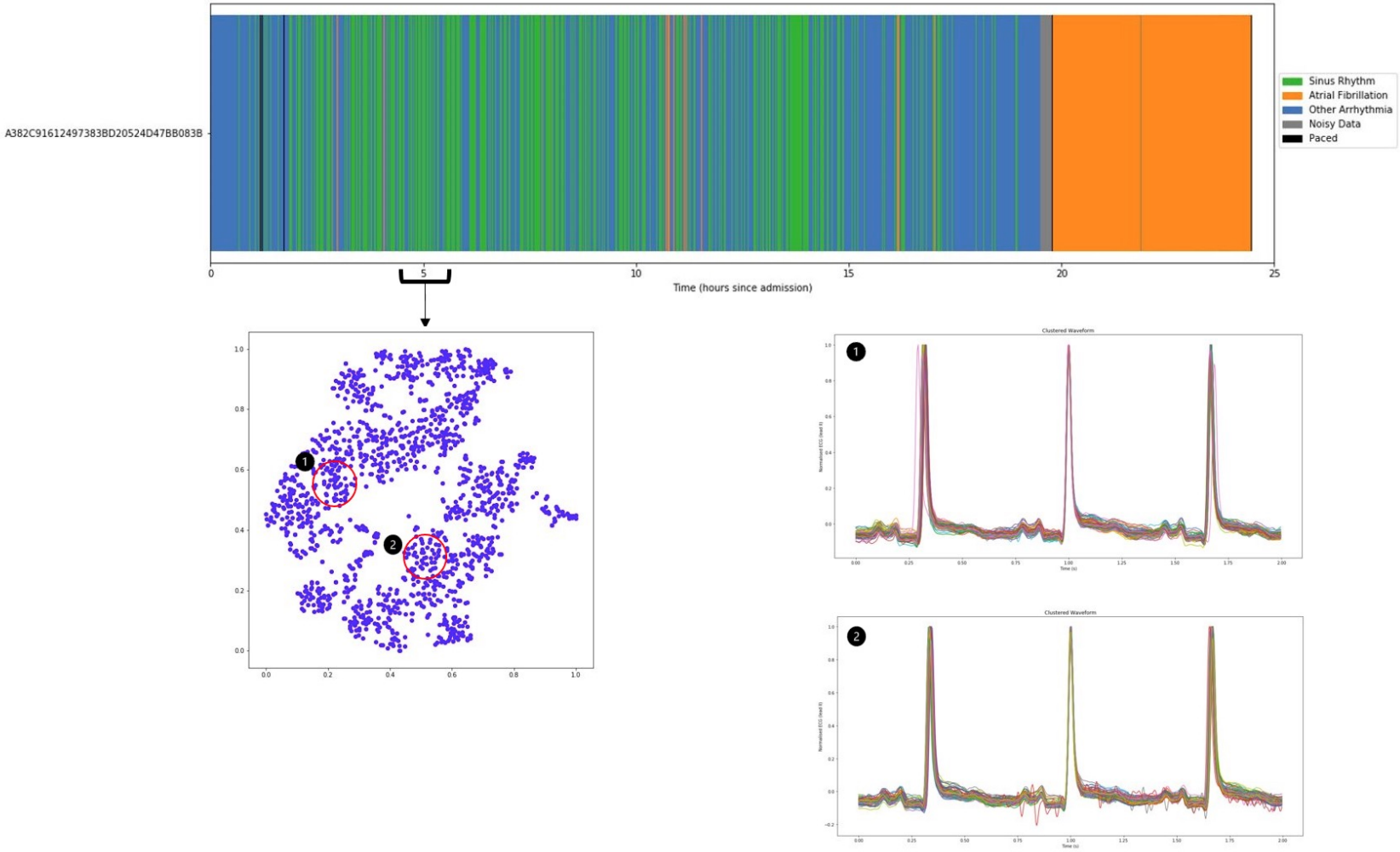


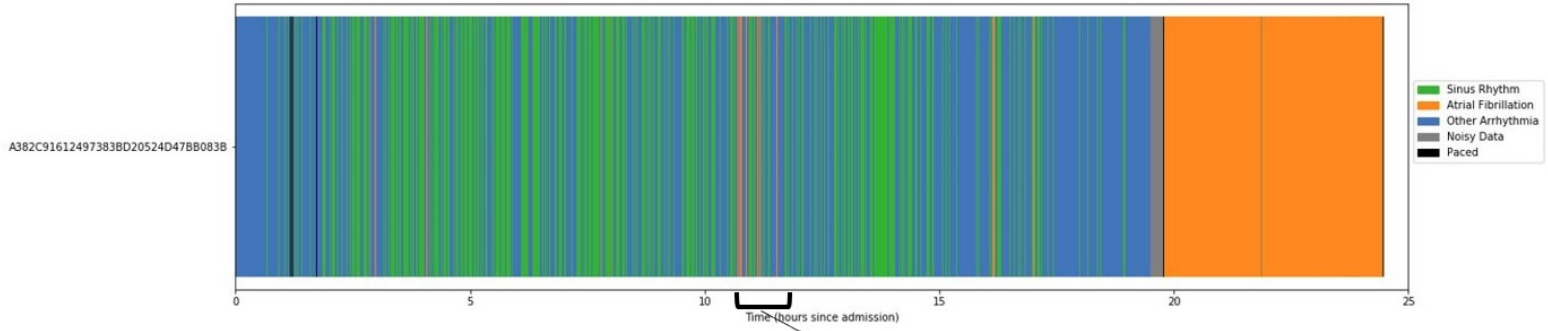
Figure 2: An overview of the Snorkel system. (1) SME users write *labeling functions (LFs)* that express weak supervision sources like distant supervision, patterns, and heuristics. (2) Snorkel applies the LFs over unlabeled data and learns a generative model to combine the LFs' outputs into probabilistic labels. (3) Snorkel uses these labels to train a discriminative classification model, such as a deep neural network.

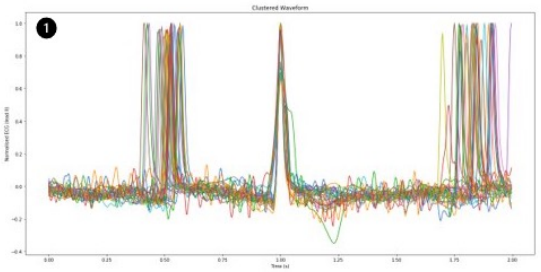
https://www.snorkel.org/blog/weak-supervision

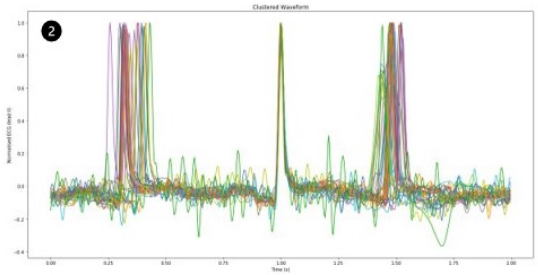
Clustering Using Model Embeddings

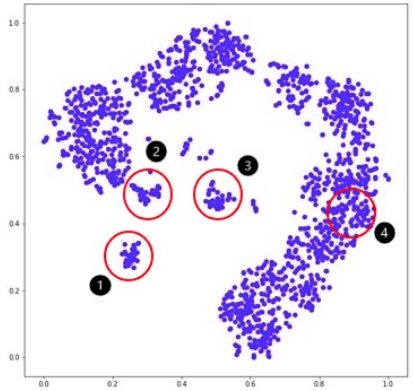


Clustering Using Model Embeddings

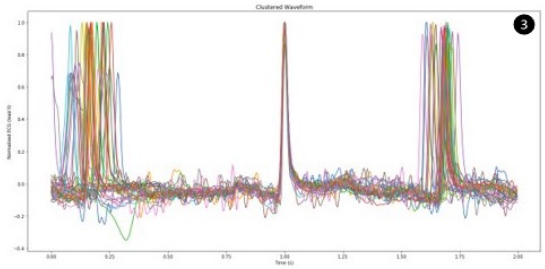


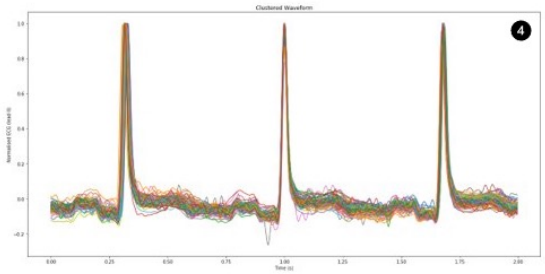




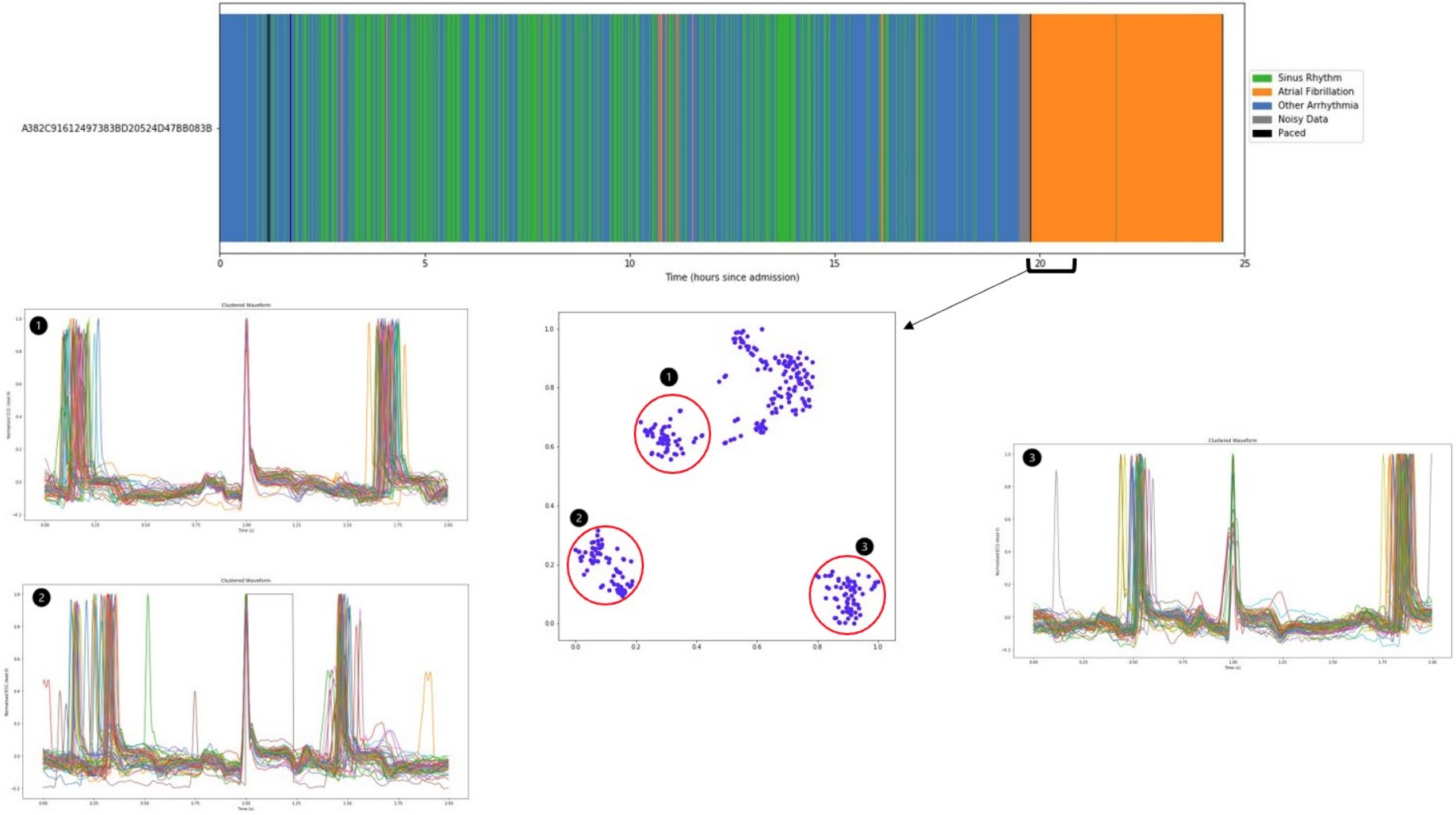


Here we can see some clusters beginning to form





Clustering Using Model Embeddings



NER: A subtask of INFORMATION extraction

- De-identification as NER (Named Entity Recognition): locate and classify named entities
- Build robust detectors of PII (Personally Identifiable Information) or PHI (Protected Health Information)
- Date, Name, ID, Age, Contact, Location and Profession
- Detected PHI entities are redacted before data being shared

NER: to tag and redact PHI before sharing data

HIPAA PHI: List of 18 Identifiers

1. Names
2. All geographical subdivisions smaller than a State
3. All elements of dates (except year) related to individual
4. Phone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger prints
17. Full face photographic images
18. Any other unique identifying number

HIPAA Safe Harbor: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard

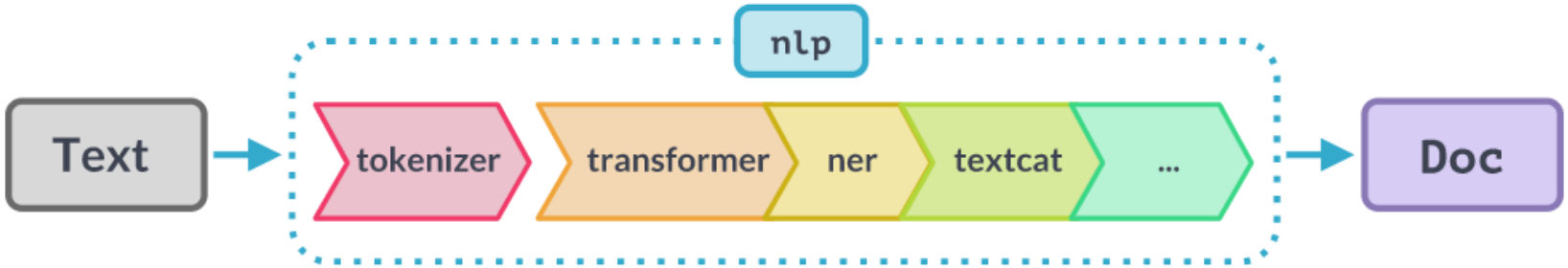
Why do we build our NER tools?

- Existing tools:
  - Amazon Comprehend Medical works better for US data
  - Ruled-based tools like: Philter, deid are slow, not generalised well
- Difficulties:
  - Ambiguity and abbreviations, typos, sentences randomly broken by new line characters
  - complicated terms (medicine names)
  - PHI entities being part of a word
  - Medical texts have tables
  - big variance in text length (1 word – 1000s words)

https://aws.amazon.com/comprehend/medical/
https://github.com/BCHSI/philter-ucsf
https://www.physionet.org/content/deid/1.1/

How do we build our NER tools?

- Data-centric approach: focus on improving data quality
- Finetune state-of-the-art pretrained transformer-based model on medical data
- Well-known NLP Library: Flair, spaCy
- Transformer-based NLP Pipeline with Customised Tokenizer



Source: spacy.io

https://github.com/flairNLP/flair
https://spacy.io/

Sota performance on n2c2 medical notes

	total	tp	fp	fn	precision	recall	F1
AGE	1612.0	1574.0	24.0	38.0	0.984981	0.976427	0.980685
CONTACT	1876.0	1867.0	93.0	9.0	0.952551	0.995203	0.973410
DATE	38159.0	37843.0	50.0	316.0	0.998680	0.991719	0.995188
ID	5447.0	5307.0	118.0	140.0	0.978249	0.974298	0.976269
LOCATION	18077.0	16804.0	289.0	1273.0	0.983092	0.929579	0.955587
NAME	29507.0	29210.0	563.0	297.0	0.981090	0.989935	0.985493
PROFESSION	2774.0	2387.0	29.0	387.0	0.987997	0.860490	0.919846

	total	tp	fp	fn	precision	recall	F1
PHI	97452.0	95426.0	732.0	2026.0	0.992388	0.97921	0.985755

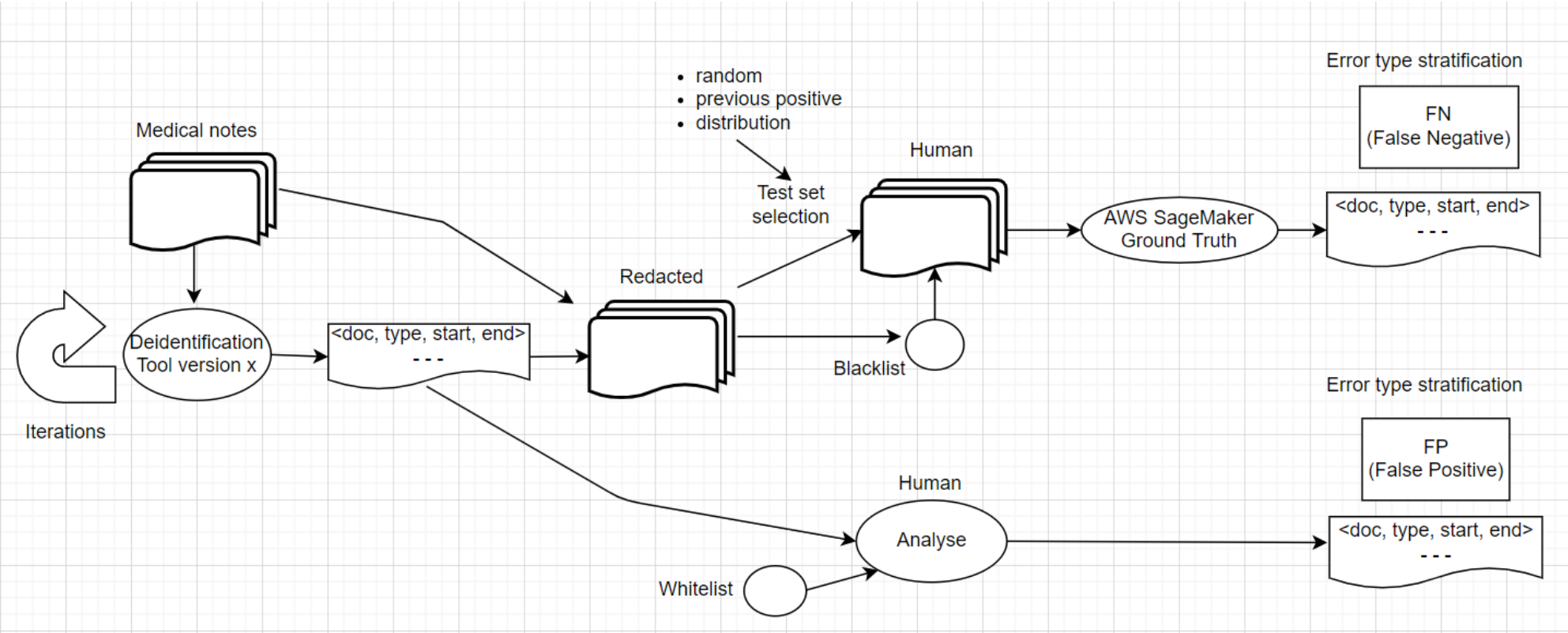
- We achieve State-Of-The-Art performance on the published n2c2 medical notes
- Character-level performance: F1 0.9858 for PHI detection

https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

Error analysis for Deidentification Tools

- False negative identification:
  - AWS SageMaker Labelling Jobs to detect PHI entities missed by NER tools
  - PHI entities detected by Deidentification Tool are redacted to avoid disclosing PHI
- False positive identification:
  - Detected PHI entities are taken out from texts and put in a list for a Datarwe QA Analyst to verify

Error analysis

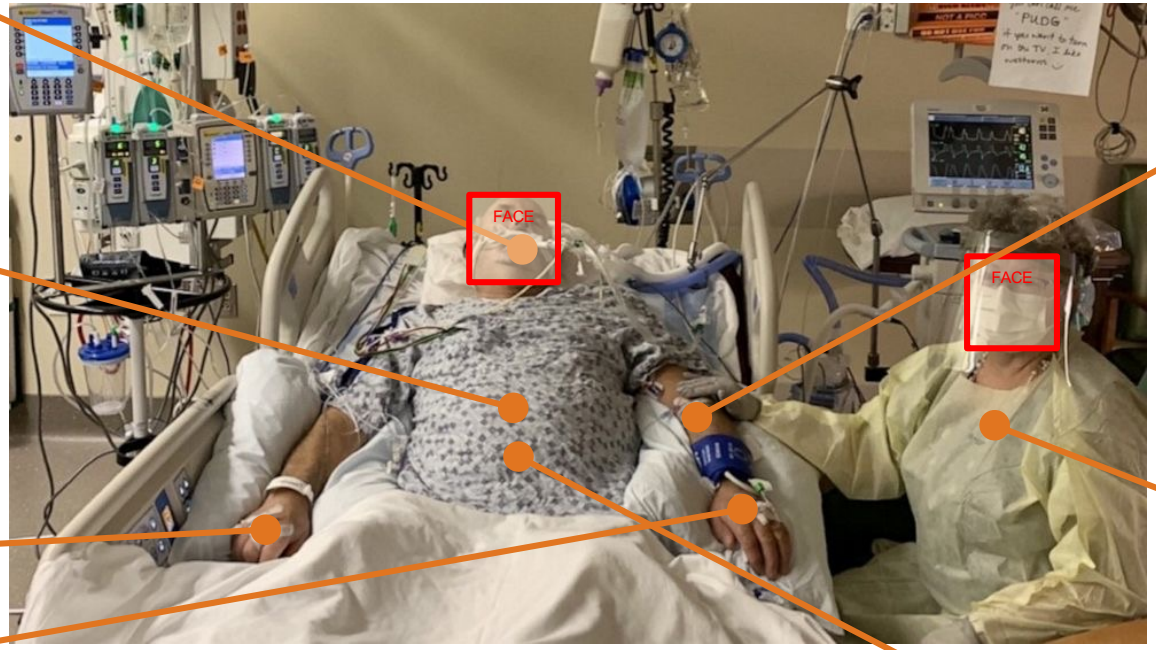


Labelling job demo: Detect missing phi

Watch My Patient

- Bed occupancy
- Sleep
- Agitation
- Bed position
- Body position
- Accidental catheter removal
- Falls
- Breathing rate (via chest movement)
- Breathing apparatus
- Staff movements
- Interruptions
- Procedure determination
  - Intubation
  - Blood Pressure Checks
  - Bolus administration

Interpreting Image



Intubated

Patient

Canula (R)

Canula (L)

Arterial Line (L)

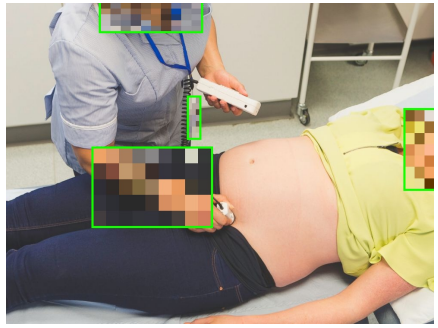
Visitor

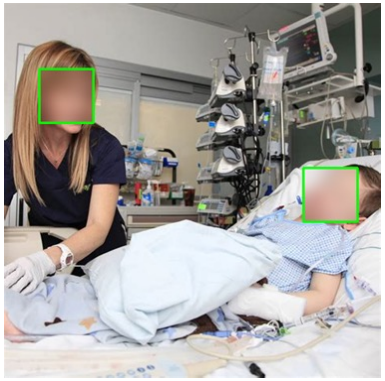
Position: Supine 30^0

Protecting the identity/modesty of patients

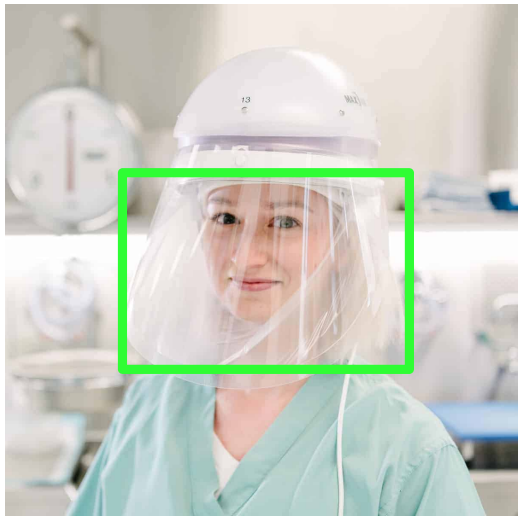
Identifying features

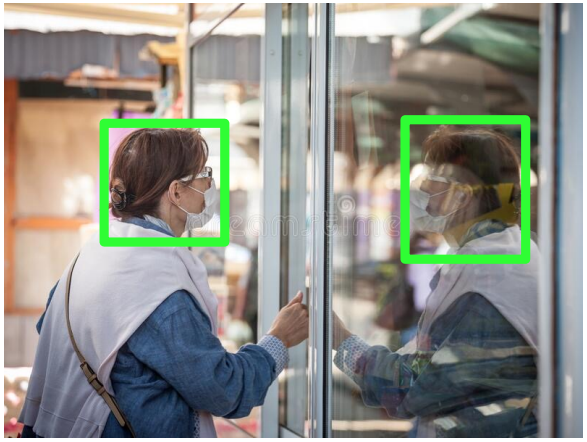
Deidentified results

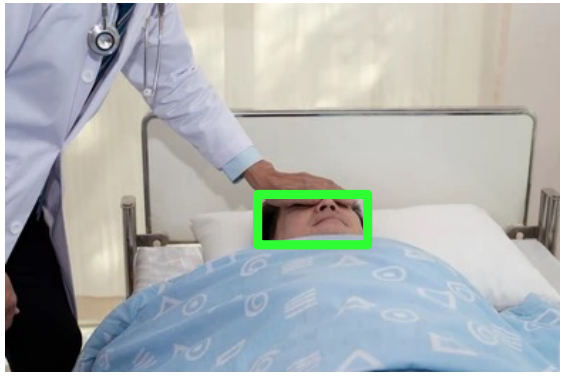




Challenges

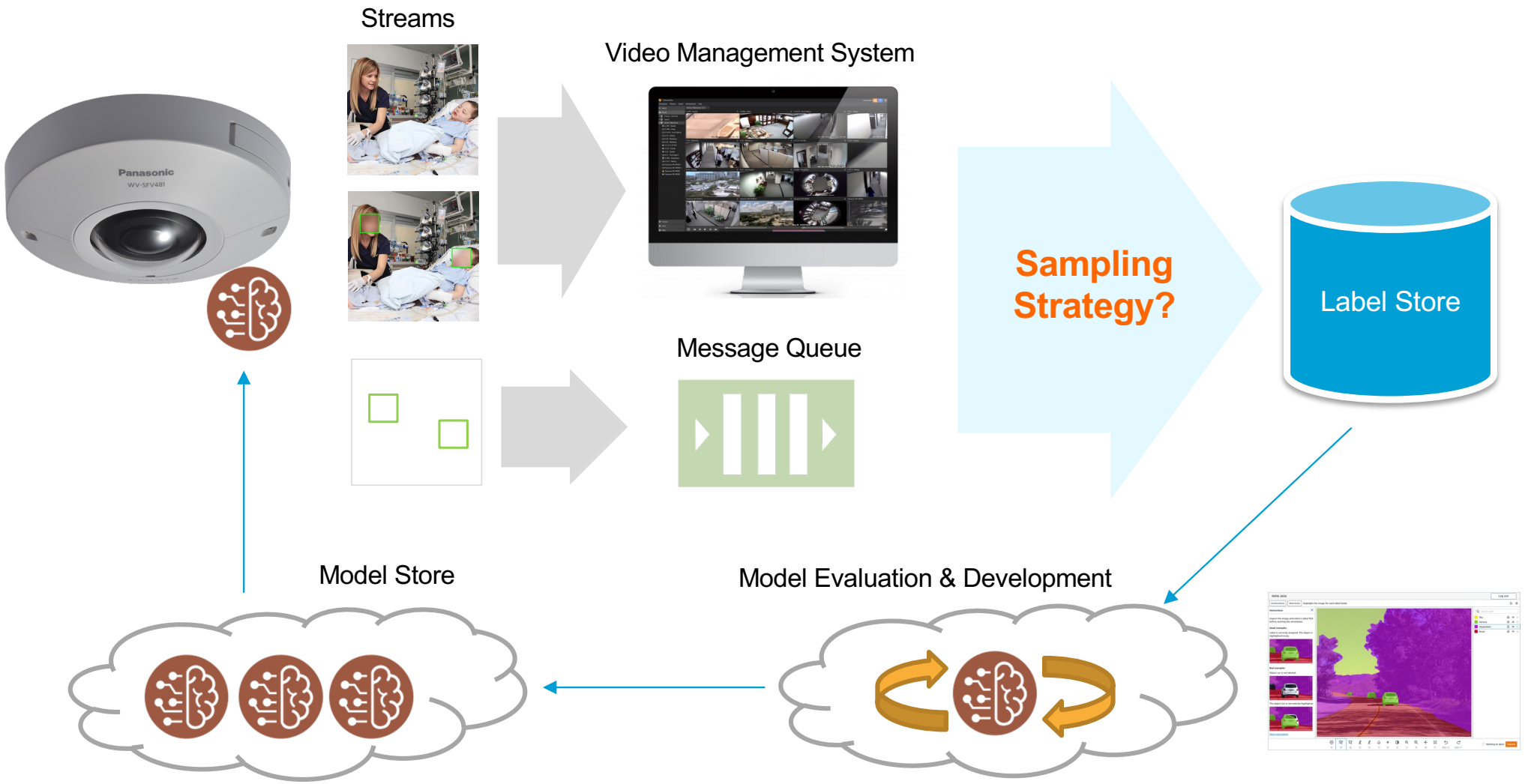






Video demonstration

Data-Centric AI for CCTV

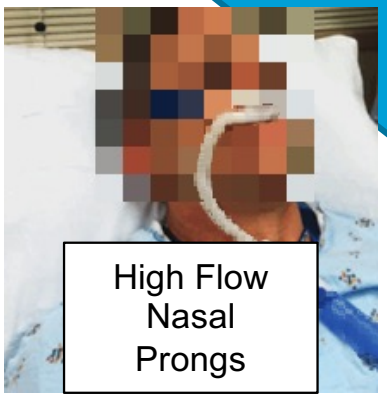


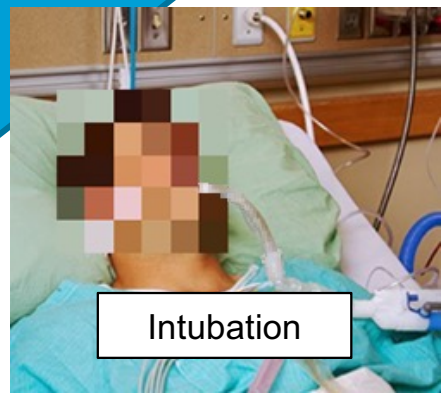
Retaining information

Nasal Prongs

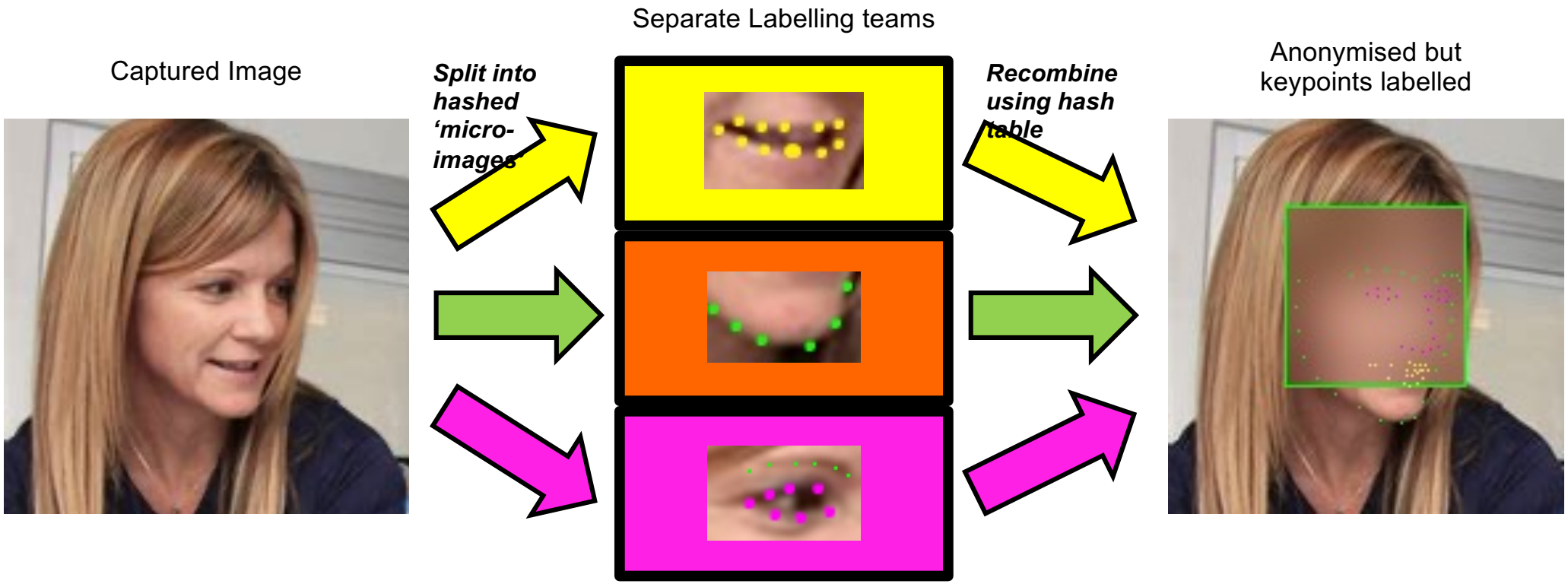
???

Tracheostomy





Preserving privacy during labelling



Key Takeaways

1. DL Can Achieve Super-Human Performance
2. Domain Shift Challenges
3. Iterative Data-Centric Approach to Evaluation and Model Improvement
4. Continuous Lifecycle

Questions?

kelvin.ross@datarwe.com
@kelvinjross

datarwe
Real world data for AI enabled medical innovation