Australian System Safety Conference 2023

VDML™

Validation Driven Machine Learning

A Systematic Approach to ML Model Training and Validation

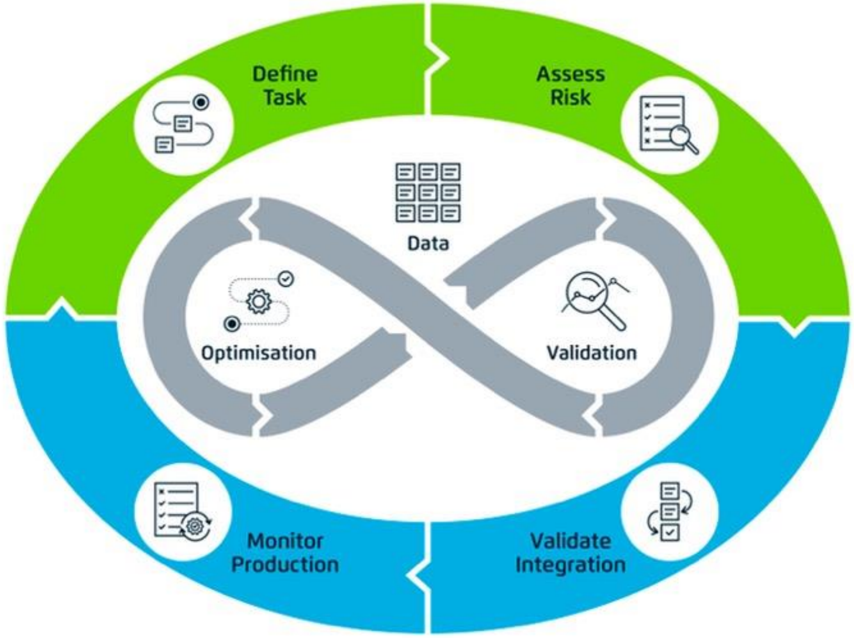
Dr Kelvin Ross
18 Oct 2023
kelvin.ross@kjr.com.au
@kelvinjross

KJR™

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

VDML™

Validation Driven Machine Learning (VDML) is a methodology developed by KJR to guide development of robust and reliable Machine Learning (ML) models. VDML emphasises understanding the risks inherent within the application context and the limitations that arise from the available data and model building processes applying iterative validation and optimisation methods to deliver an acceptable solution which can be integrated and governed within a real-world context.



DEFINE CONTEXT

Define the goals of applying machine learning to a specific problem area, being sure to include the data being used, the context of use (historical analytics vs live decision support) and expected benefits from a range of different stakeholders. Given this context, assess risks, including the impacts of potential failure, the required governance processes.

RESOLVE LIMITATIONS

Direct use of pre-built models or naïve approaches to machine learning can lead to unreliable performance. Key to validating and optimising model performance is the selection of training and testing data sets which are close to real world usage, and detailed error analysis which can uncovers underlying faults and limitations.

GOVERN BEHAVIOUR

Track the integrity of the model through build, deploy and operation, monitoring for residual risks, model drift / sabotage, identifying opportunities for further optimisation and risk reduction.

KJR™

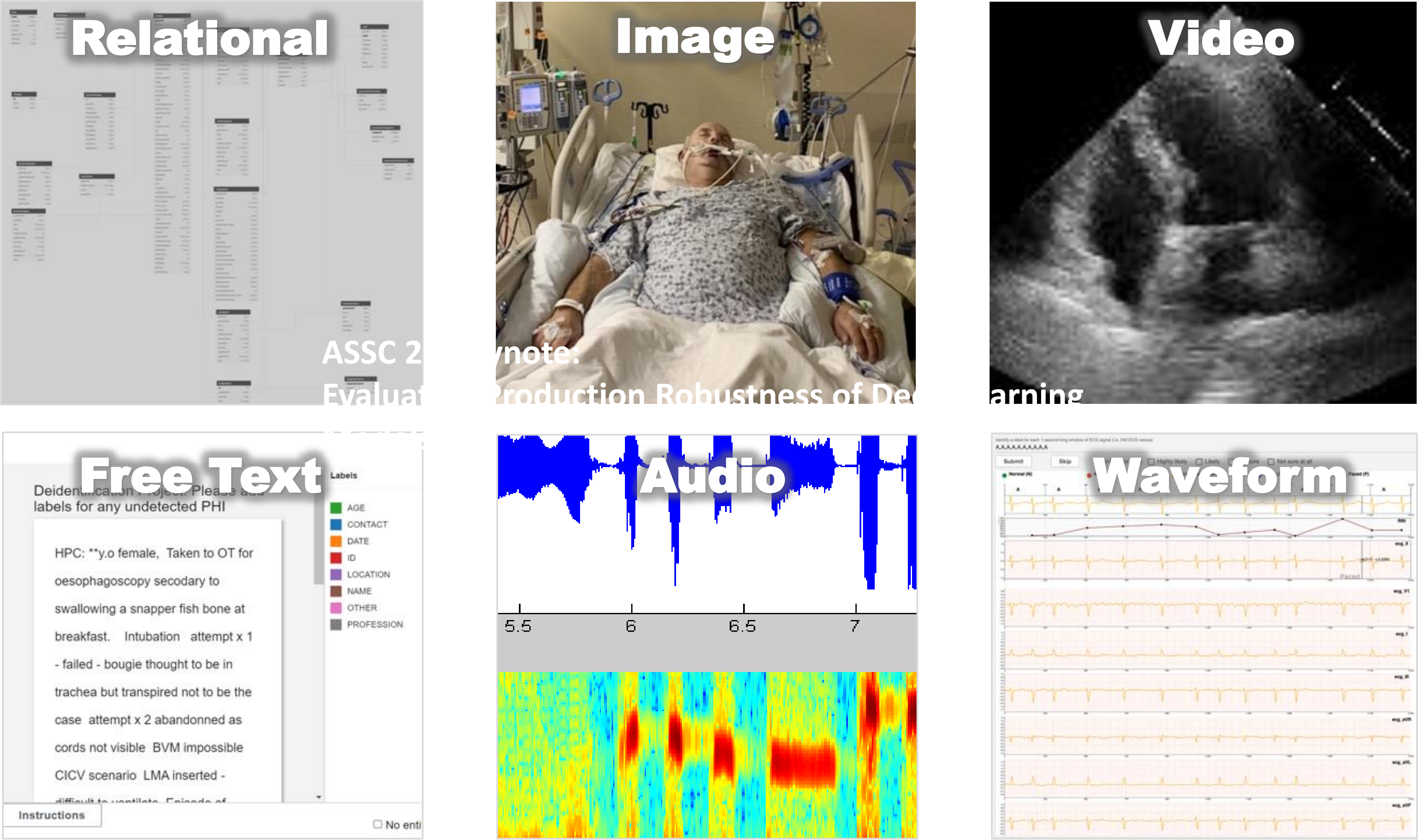
Software Quality Engineering

datarwe

Smart AI Connect

KJR™

Drone / Remote Data / ISR


Relational
Image
Video
ASSC 2
ynote:
Evaluat
Production Robustness of De
arning
Free Text
Labels
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 -
AGE
CONTACT
DATE
ID
LOCATION
NAME
OTHER
PROFESSION
Instructions
No enti
Audio
5.5
6
6.5
7
Waveform

Why is ML development different to traditional processes?

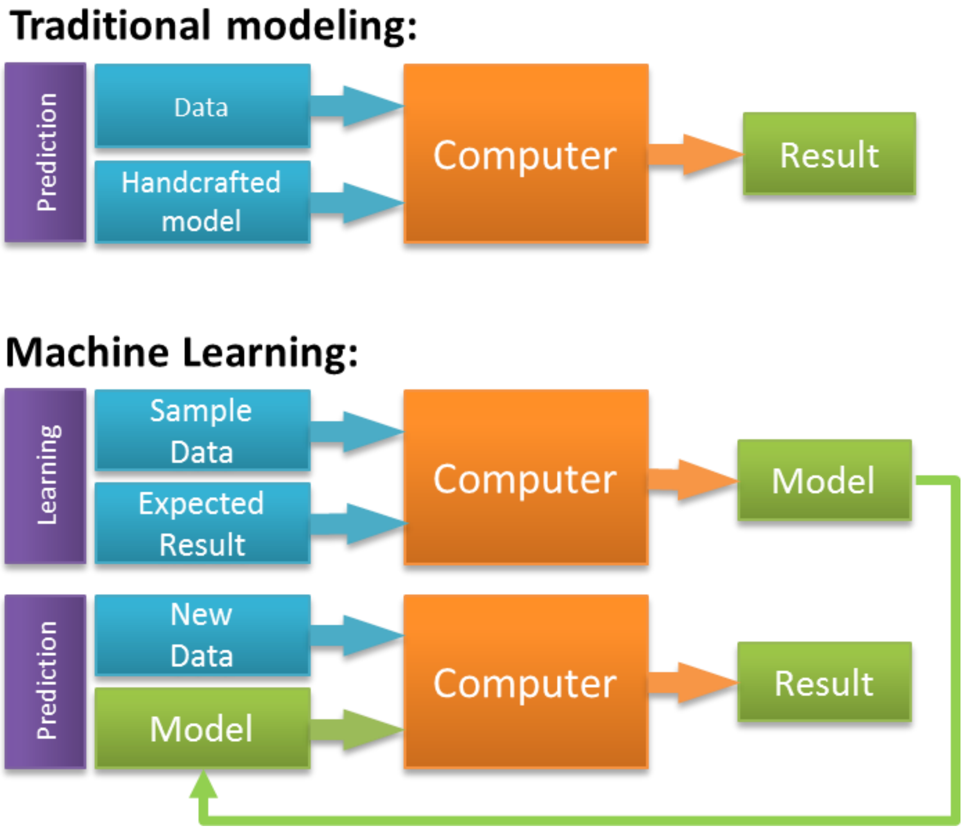
ML development is frequently exploratory and opaque

- Now, rather than starting from a blank slate, it is common for ML development to start from finding a model which is close to the target problem and transferring that solution to the target domain.
- Traditional software development often uses the same approach, but in the case of ML, it is harder to understand why a model may not perform well in the new domain, as there is no human readable code to inspect.
- Instead, a carefully designed set of data-driven experiments is required to identify the root cause issues identified above.
- VDML provides a structure approach to this stage of model development

ML development is shifting toward fine tuning existing models

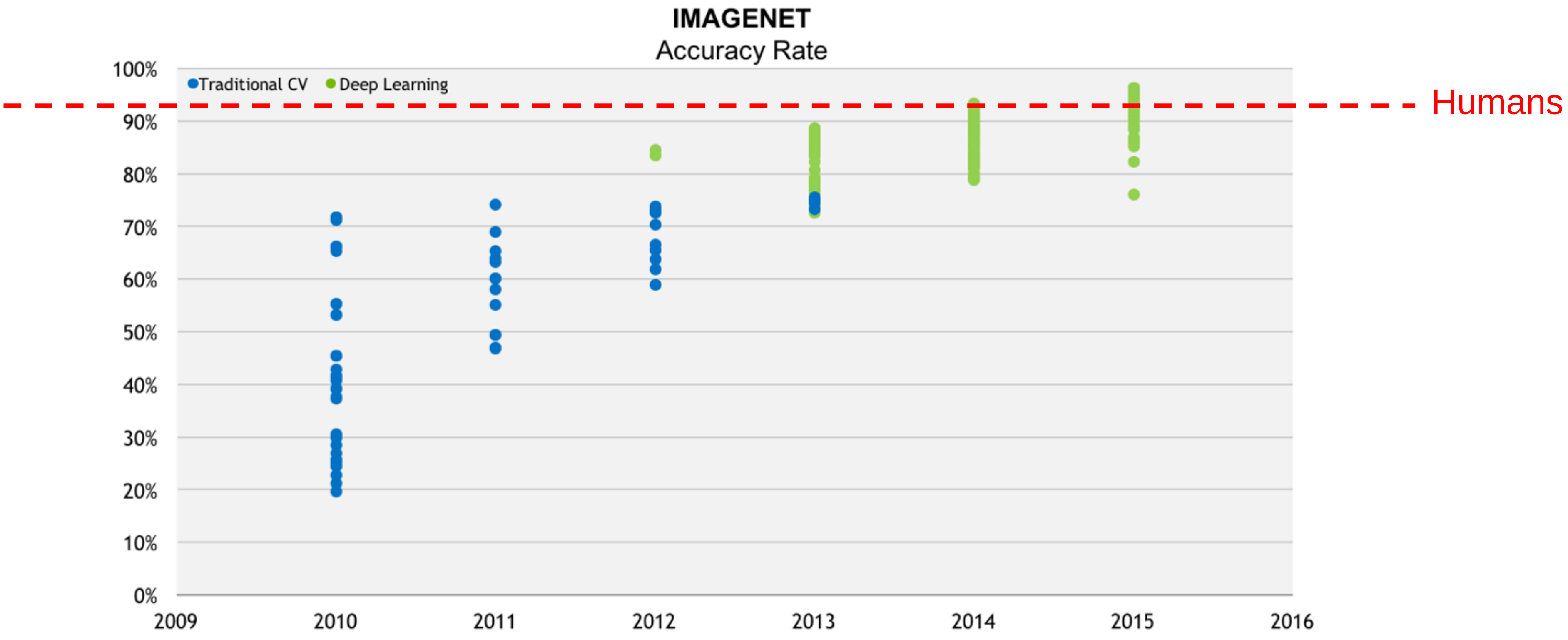
- As large models with a strong general capability, such as ChatGPT, become available as base solutions, ML development is increasingly focused on fine-tuning those based models.
- By using these existing models to tag new training data, using zero shot classification and active learning techniques, ML development becomes less about amassing a large number of examples, and more about having a very clear understanding of what is needed to fine tune model performance for a specific task.
- In this scenario, the VDML optimisation and validation process become more central to model development in general.

What is Machine Learning



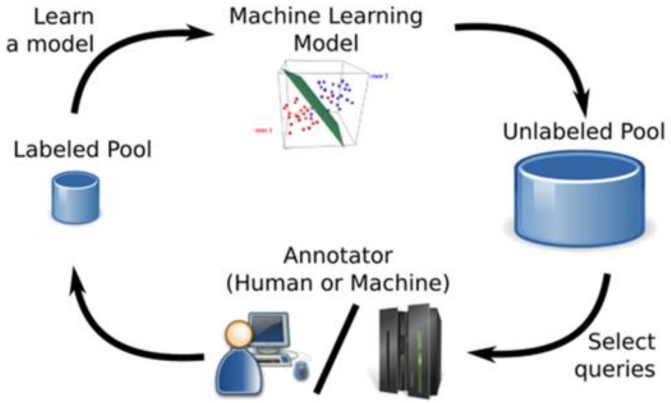
Source: zeiss.com

ML Improvement



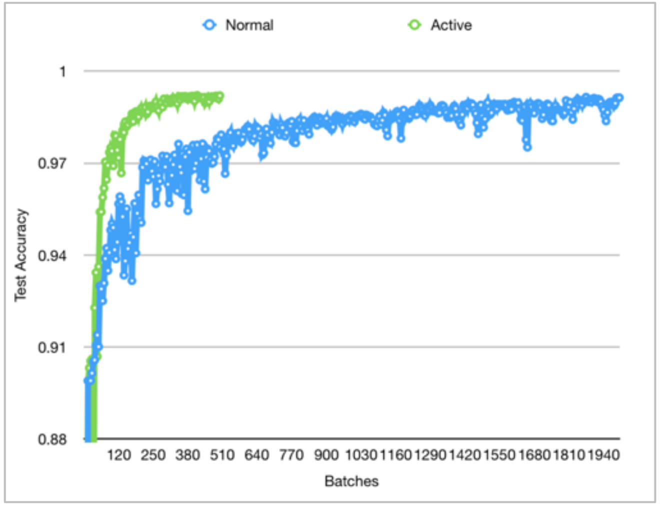
More info: https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence

Active Learning

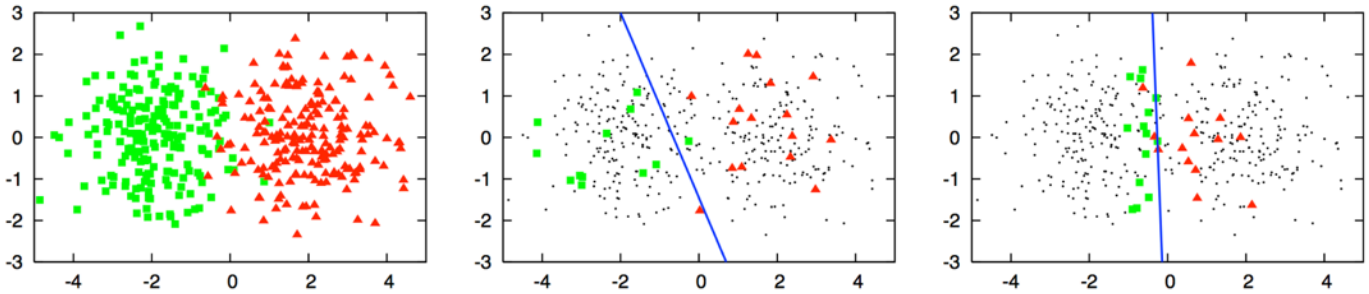


https://www.datacamp.com/community/tutorials/active-learning

- Uncertainty
- Decision Boundaries
- Similarity

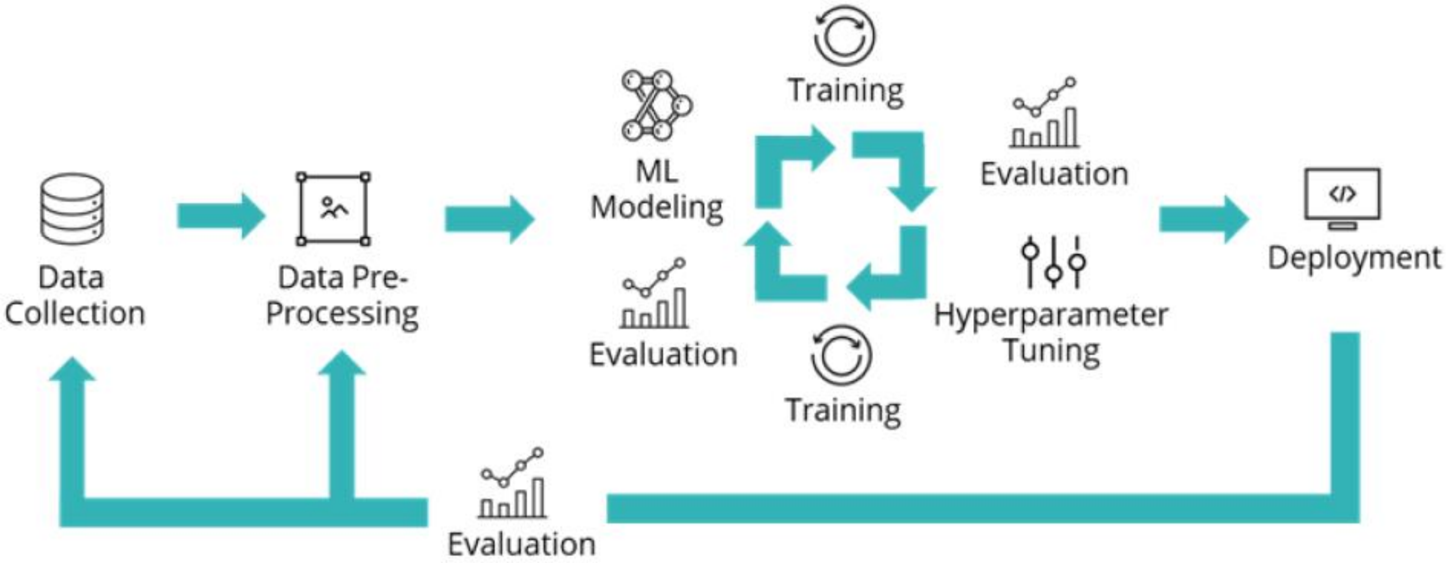


https://becominghuman.ai/accelerate-machine-learning-with-active-learning-96cea4b72fdb



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

Data-Centric AI

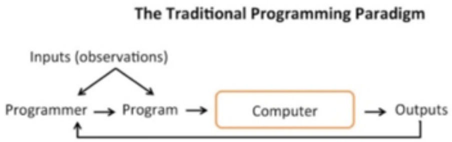


https://dida.do/blog/data-centric-machine-learning

Software 2.0 Stack

Andrej Karpathy, Tesla
https://vimeo.com/274274744

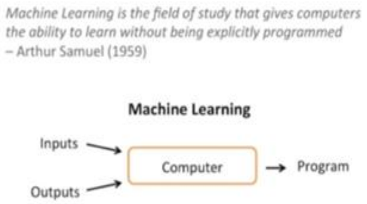
Software 1.0



Supporting developers to write rules (programs) to produce outputs from inputs

E.g. IDEs, Test Automation

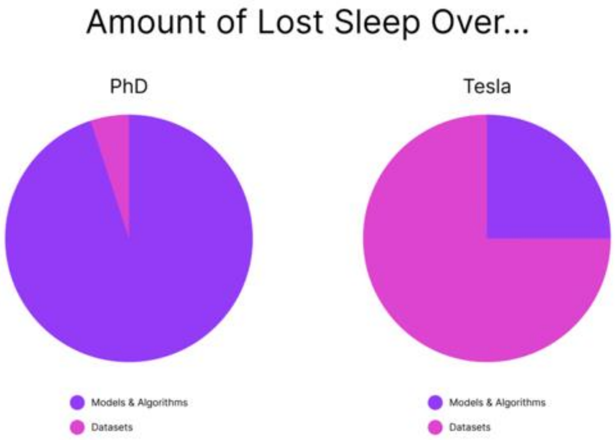
Software 2.0



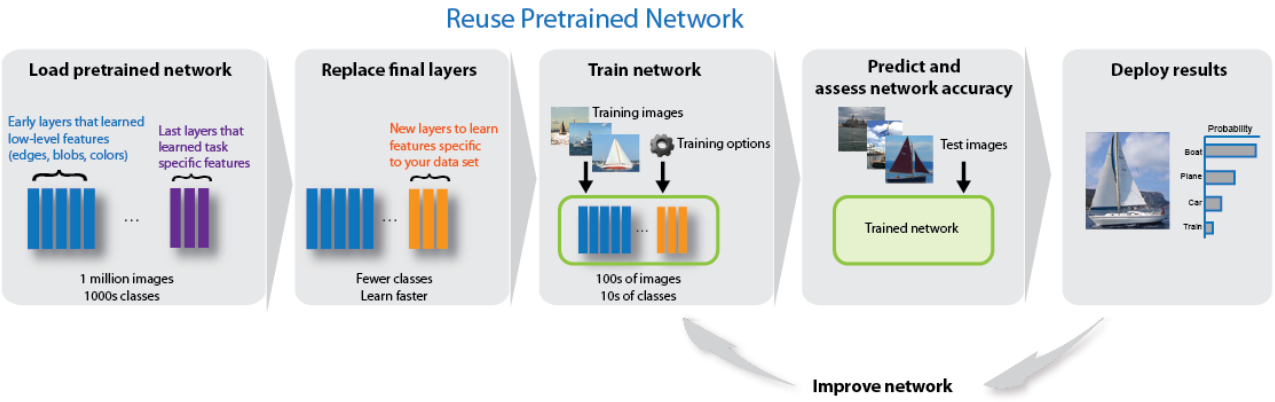
Supporting developers to learn optimal rules (ML architectures and weights) from example inputs and outputs

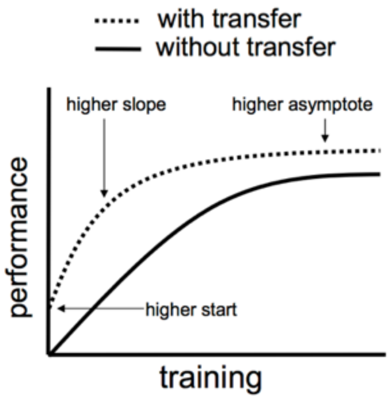
2 main areas supporting teams:

- Label
- Maintain surrounding "Data Infrastructure"
    - Visualise datasets
    - Create/edit labels
    - Bubble up likely mislabeled examples
    - Suggest data to label
    - Flag labeler disagreements
    - ...



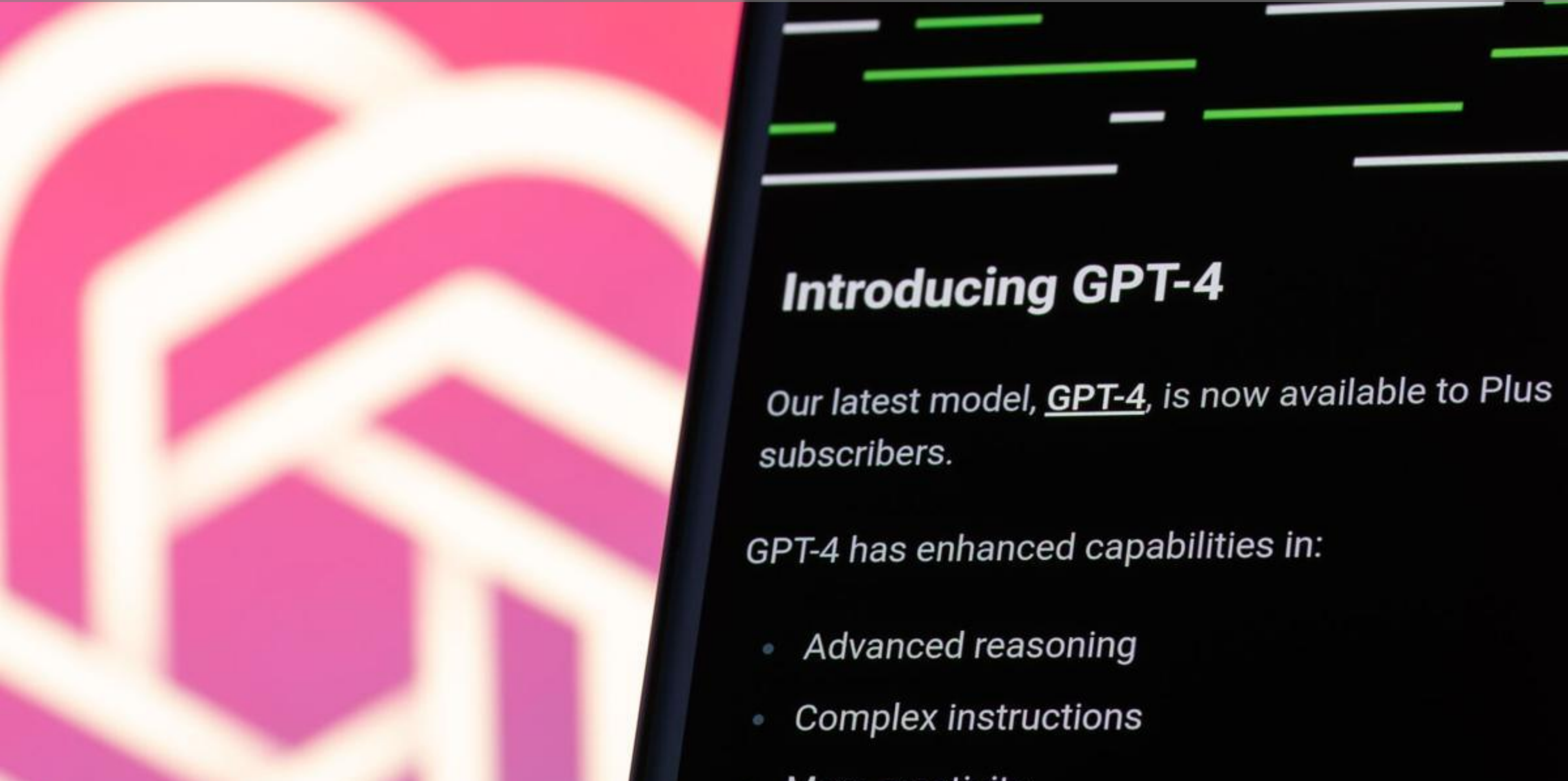
Transfer Learning



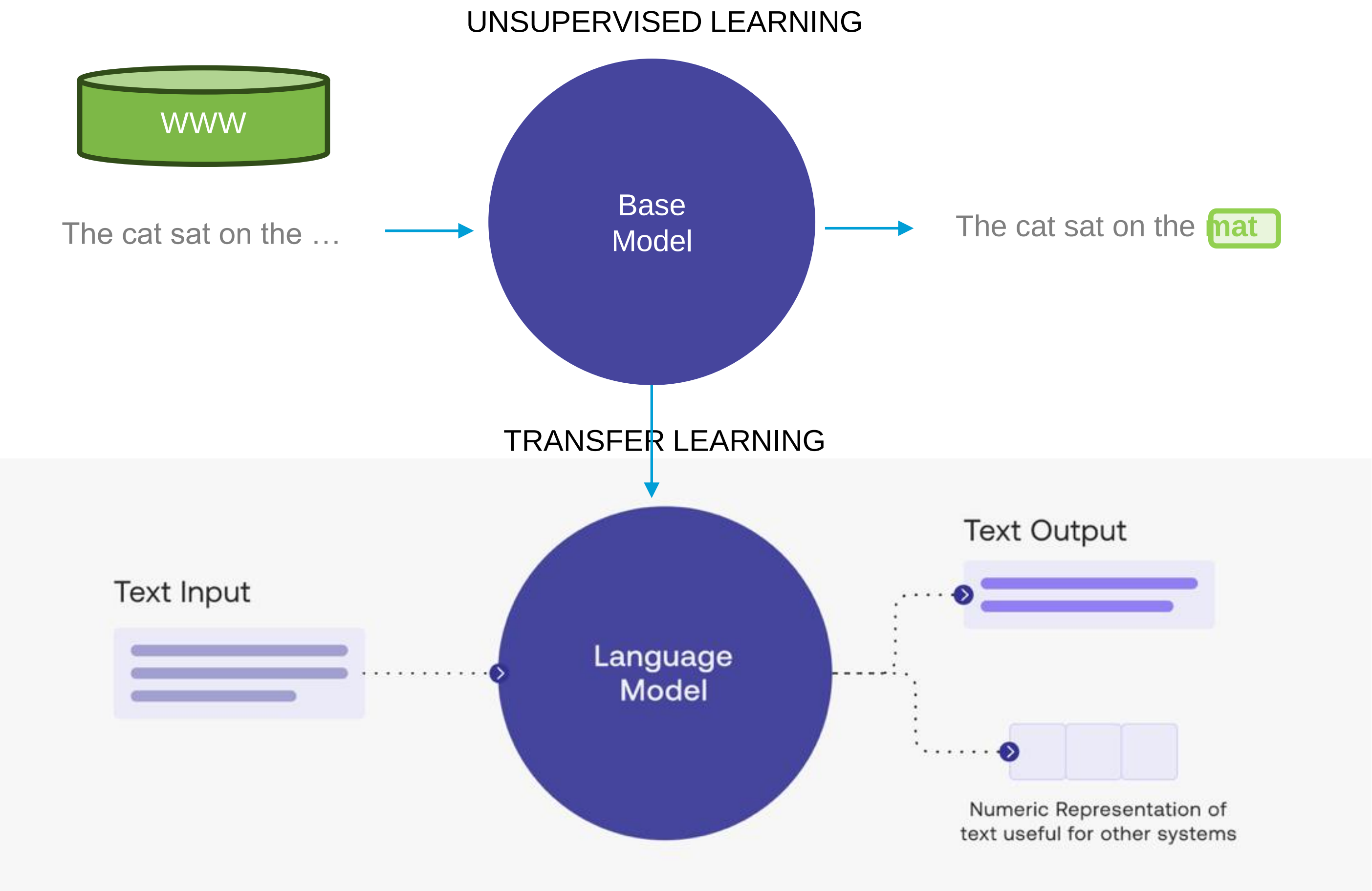


Source: https://it.mathworks.com/help/deeplearning/ug/transfer-learning-using-alexnet.html

Generative AI

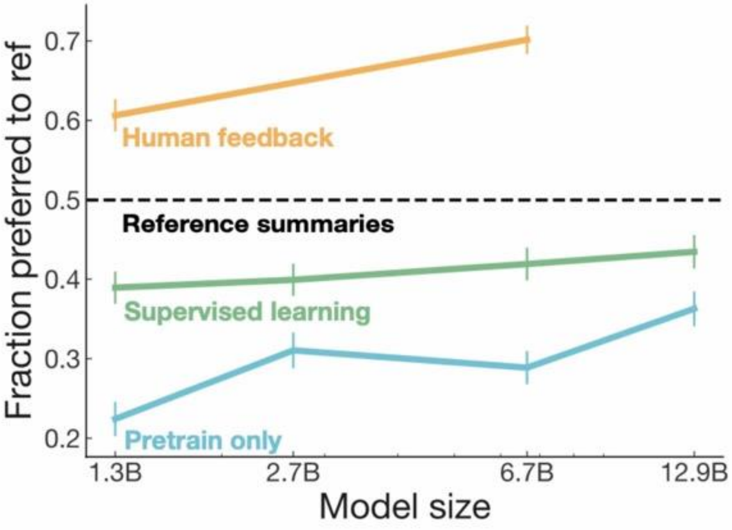


Large Language Models

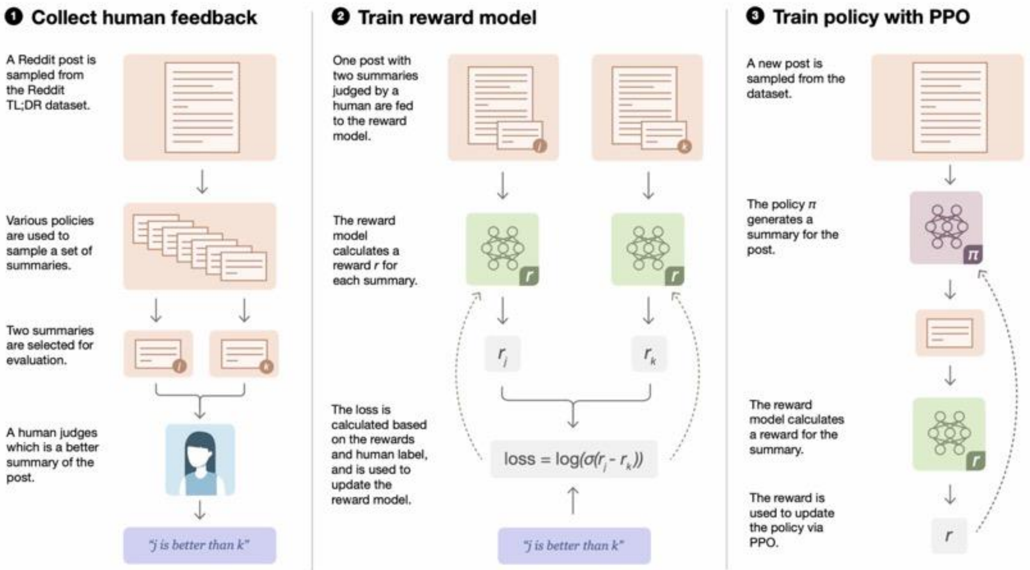


Large Language Models

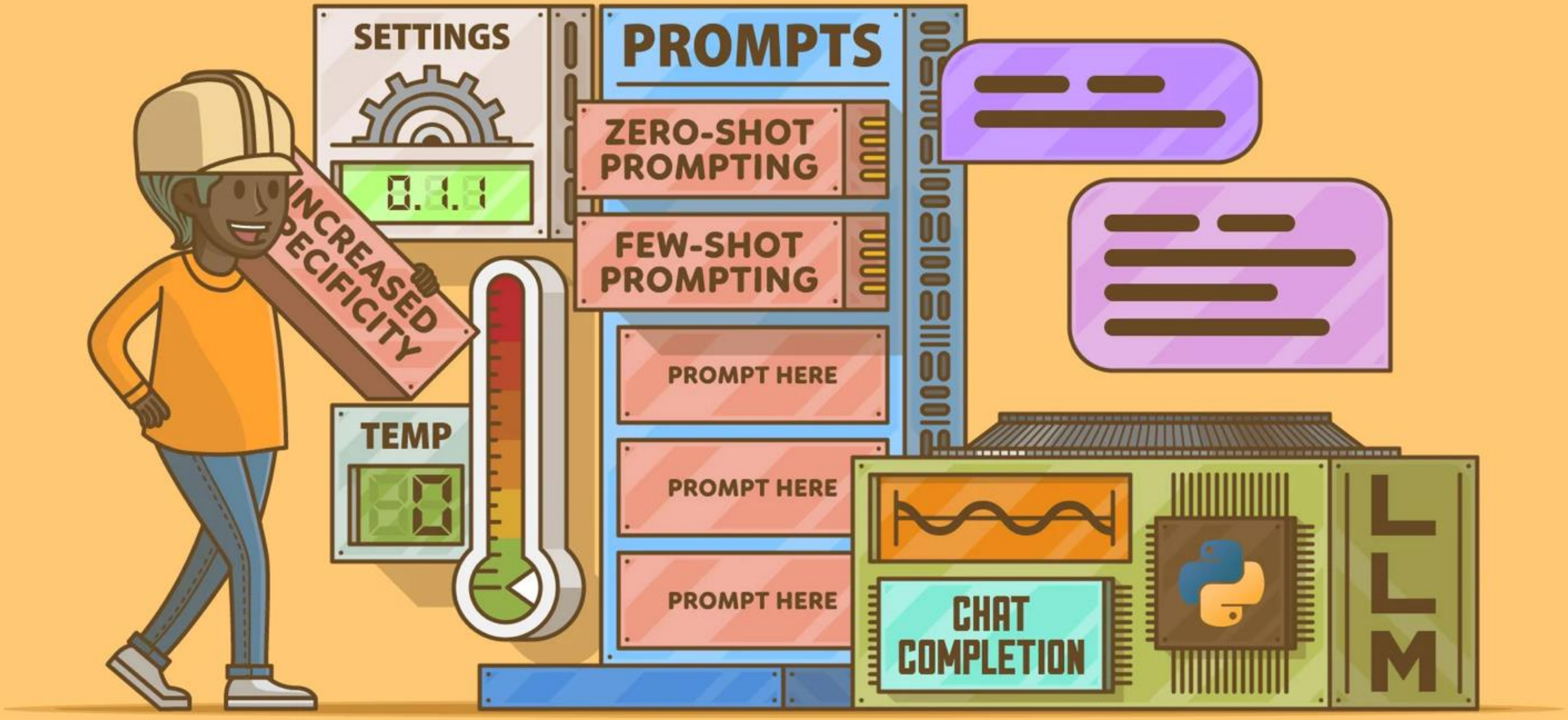
RL with PPO results in "better" LLMs
than using regular supervised learning



The original RLHF method for summarization
("Learning to Summarize from Human Feedback")



https://sebastianraschka.com/blog/2023/llm-reading-list.html



https://realpython.com/practical-prompt-engineering/

Real Python

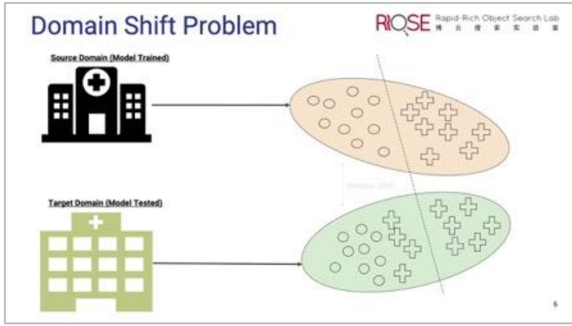
Common ML Faults

Unintended Signal Correlation

Hidden Stratification

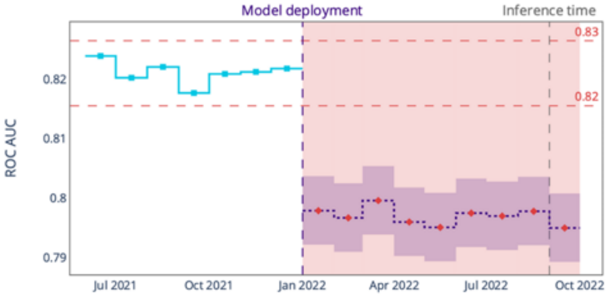
https://youtu.be/_4gn7ibByAc

Domain Shift



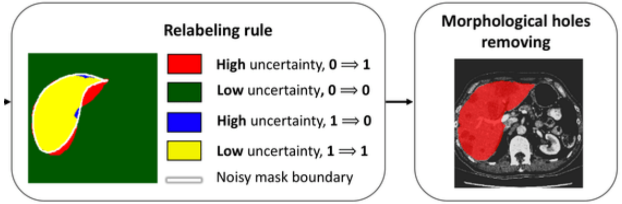
https://youtu.be/diJAM-Z6u0Y

Data Leakage



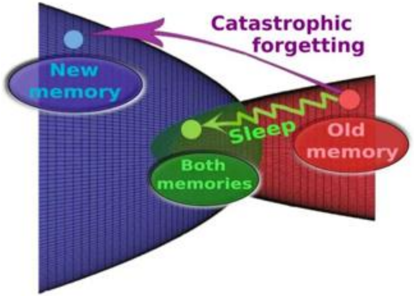
https://www.nannyml.com/blog/
3-common-causes-of-ml-model-failure-in-production

Ground Truth Inconsistency



https://www.researchgate.net/publication/
349363942_Uncertainty-based_method_for_
improving_poorly_labeled_segmentation_datasets

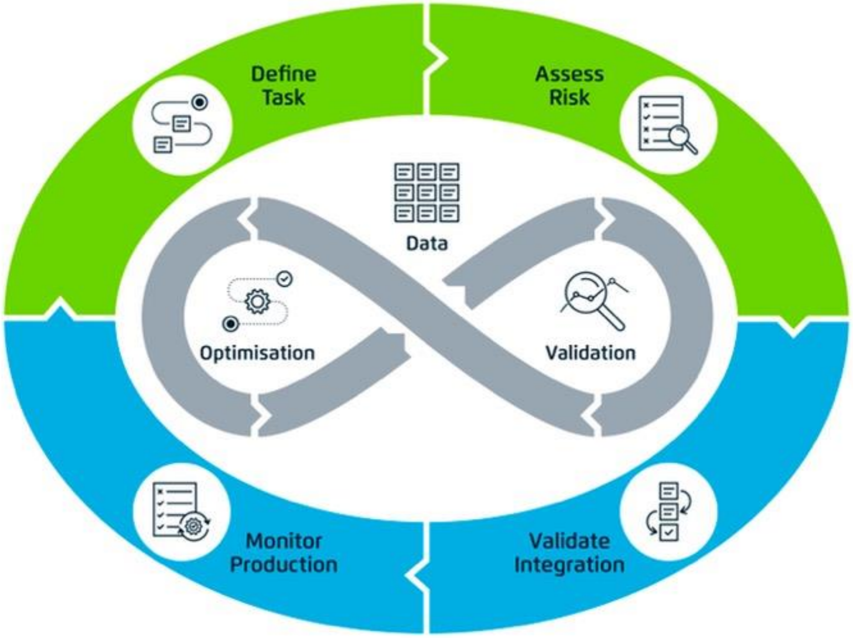
Forgetting



https://spectrum.ieee.org/
catastrophic-forgetting-deep-learning

VDML™

Validation Driven Machine Learning (VDML) is a methodology developed by KJR to guide development of robust and reliable Machine Learning (ML) models. VDML emphasises understanding the risks inherent within the application context and the limitations that arise from the available data and model building processes applying iterative validation and optimisation methods to deliver an acceptable solution which can be integrated and governed within a real-world context.



DEFINE CONTEXT

Define the goals of applying machine learning to a specific problem area, being sure to include the data being used, the context of use (historical analytics vs live decision support) and expected benefits from a range of different stakeholders. Given this context, assess risks, including the impacts of potential failure, the required governance processes.

RESOLVE LIMITATIONS

Direct use of pre-built models or naïve approaches to machine learning can lead to unreliable performance. Key to validating and optimising model performance is the selection of training and testing data sets which are close to real world usage, and detailed error analysis which can uncovers underlying faults and limitations.

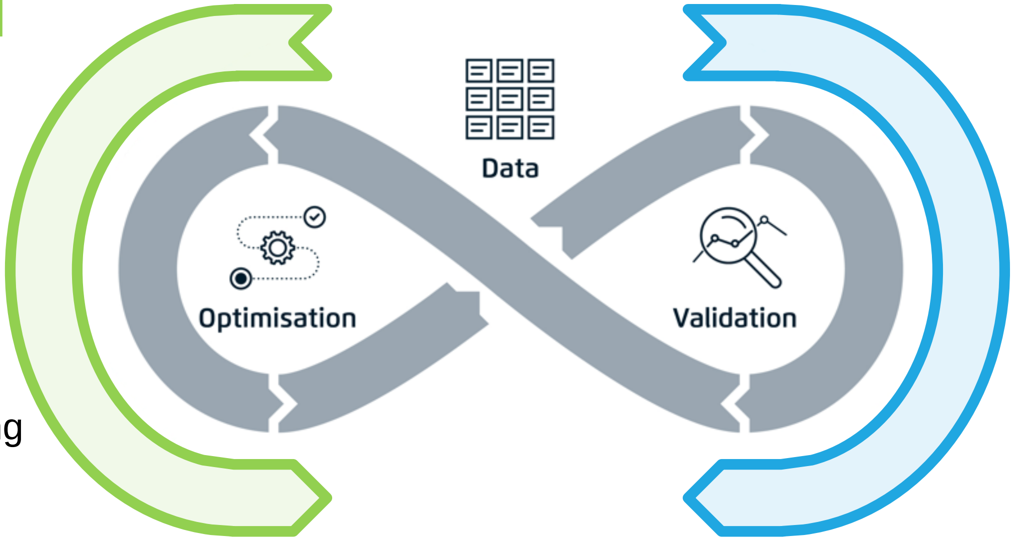
GOVERN BEHAVIOUR

Track the integrity of the model through build, deploy and operation, monitoring for residual risks, model drift / sabotage, identifying opportunities for further optimisation and risk reduction.

Resolve Limitations

Optimisation

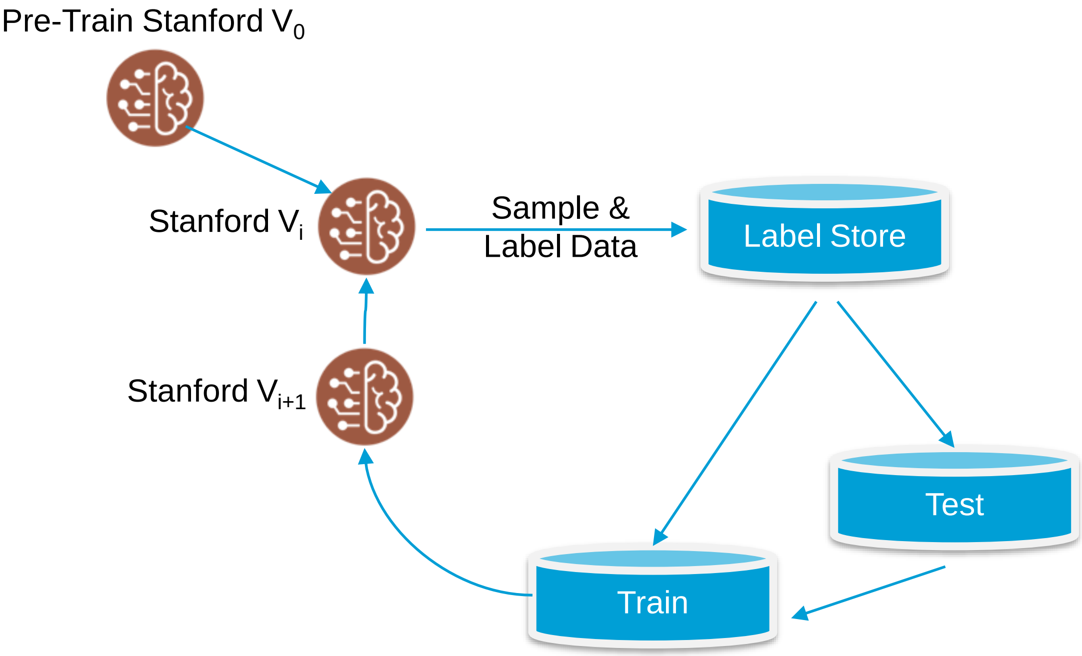
- Transfer Learning
- Fine-Tuning
- Active Learning
- Data Augmentation
- Synthetic Data
- Class Balancing
- Ensembling
- Hyperparameter Tuning

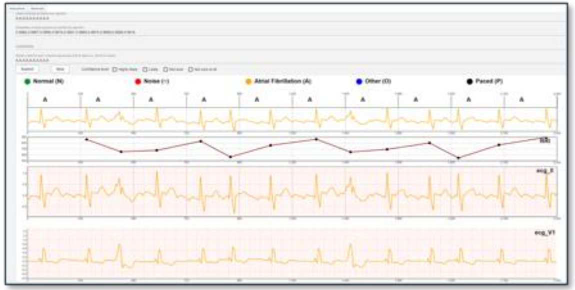


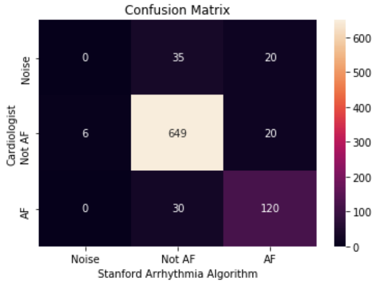
Validation

- Sanity Testing
- Ground Truth Labelling
- Sampling
- Error Analysis
- Threshold Analysis
- Uncertainty Analysis
- Model Explanation
- Stratification Analysis
- Label Reasoning
- Regression Testing

Iterative Improvement





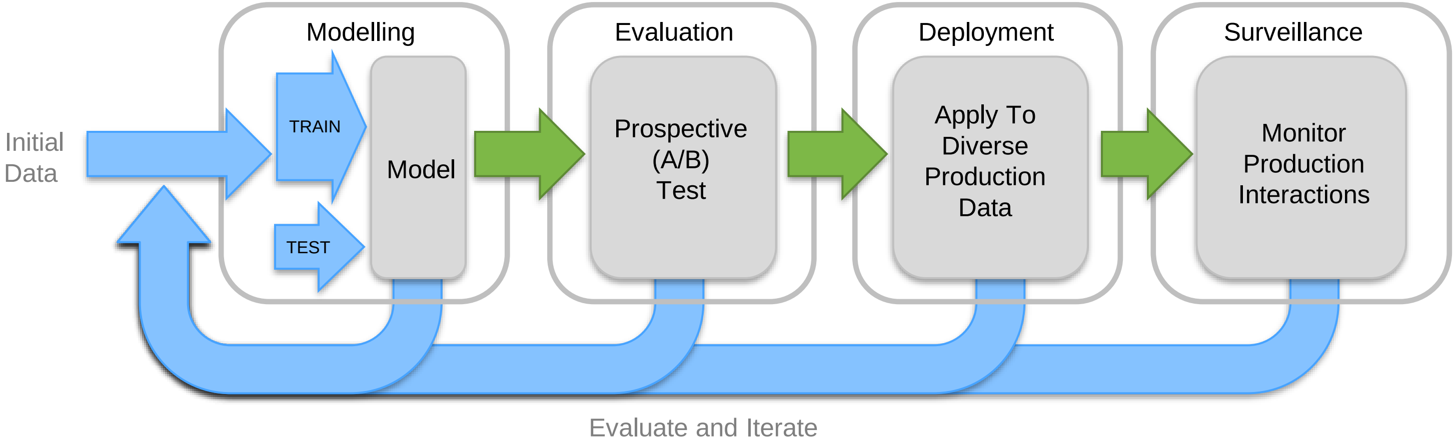



```
56]: print('Classification Report')
     print(classification_report(ground_truth, L2, target_name

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
```

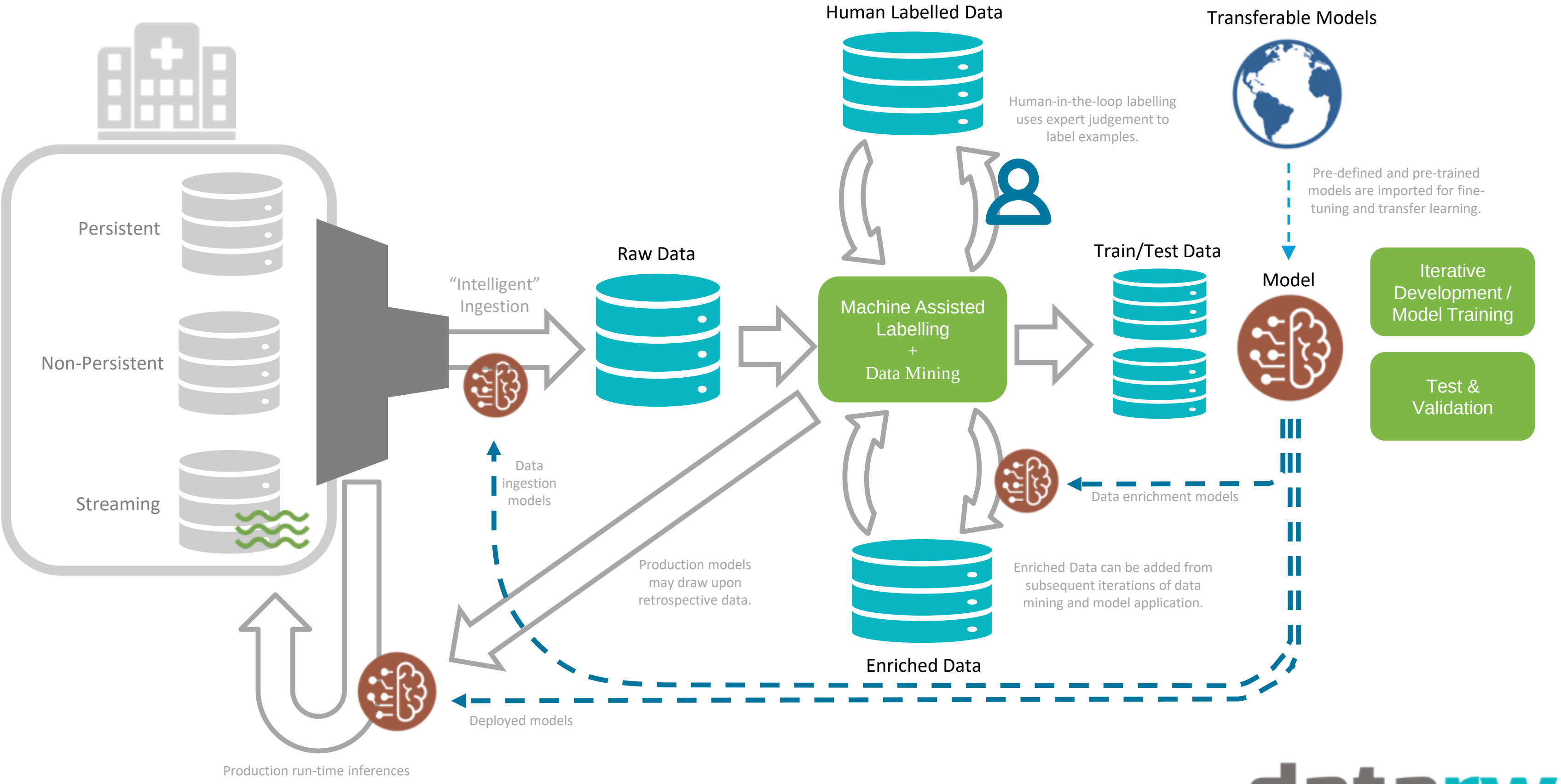
Consider Data Leakage?

Clinical Trial Approach

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

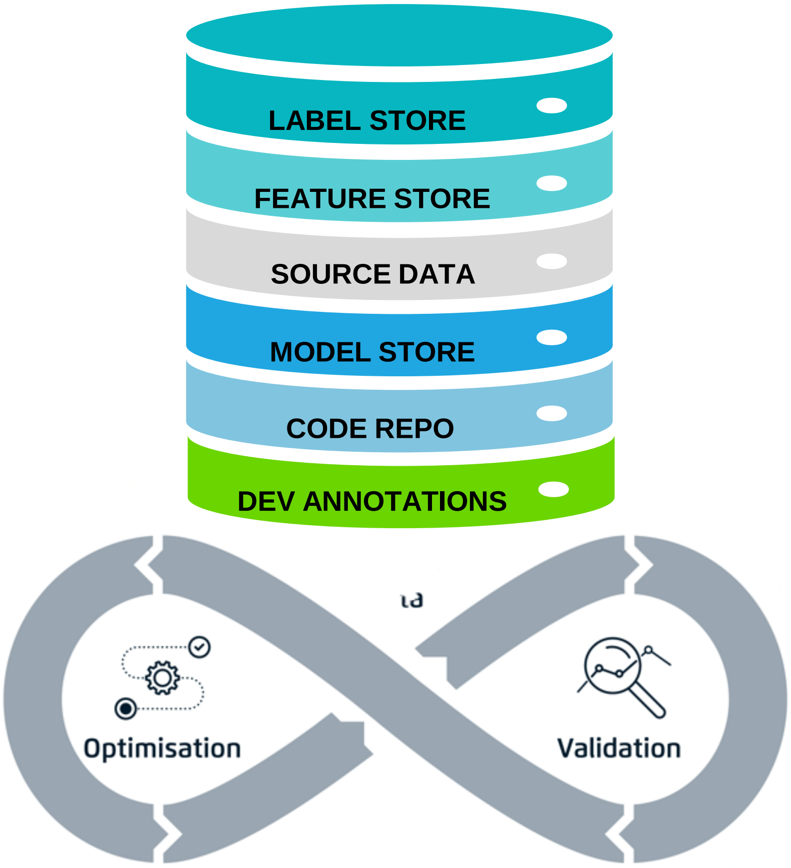


Datarwe Data Engine



datarwe

ML Data Store



Source Data

- Data sampled from real-world data
- Provenance information stored as meta-data

Feature Store

- Data shaped by data engineers as input

Label Store

- Human annotated labels
- Labels from other ML models
- Derived labels

Model Store

- Repository of ML models
- Record of how ML model was derived

Code Repository

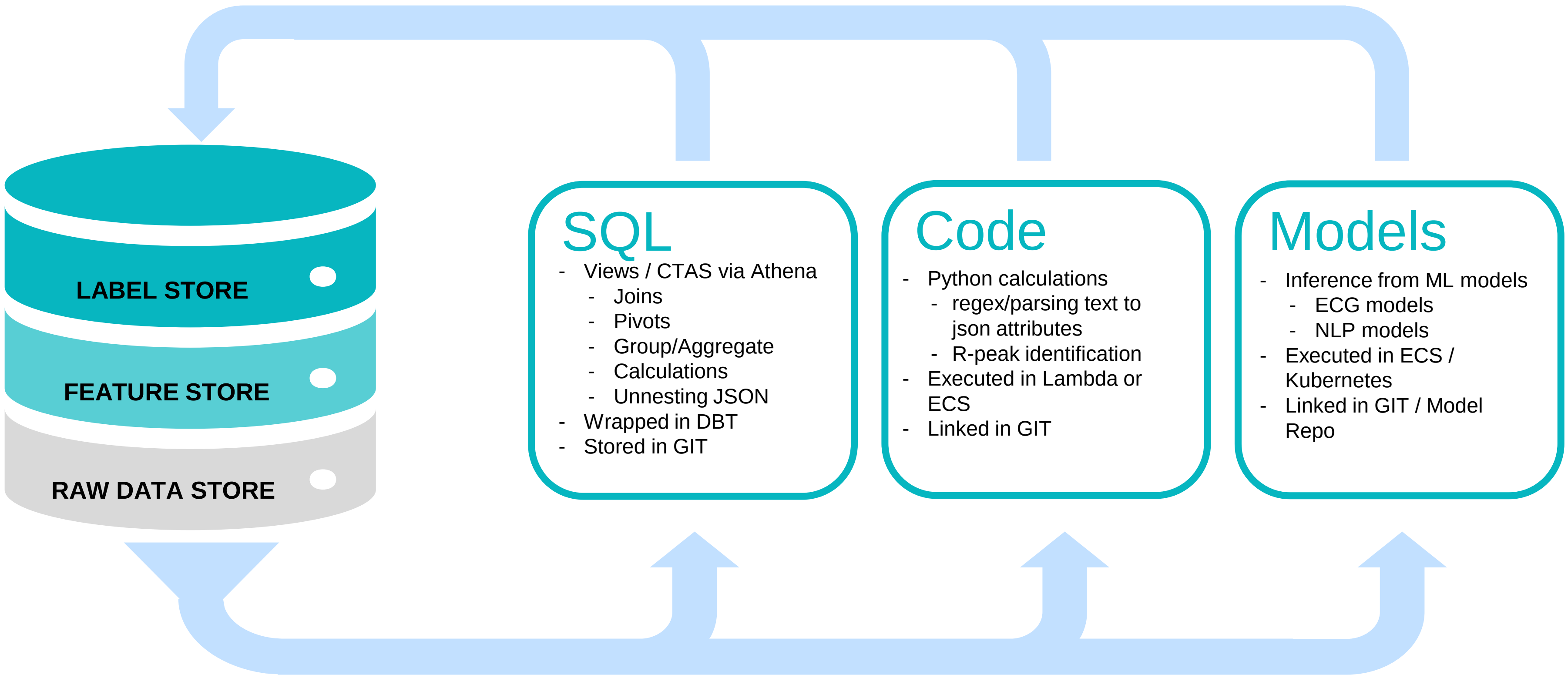
- Supporting code for ML model application, such pre- and post-processing
- Inference code for feature and label store derivation
- Mechanisms for generating data dynamically

Development Annotations

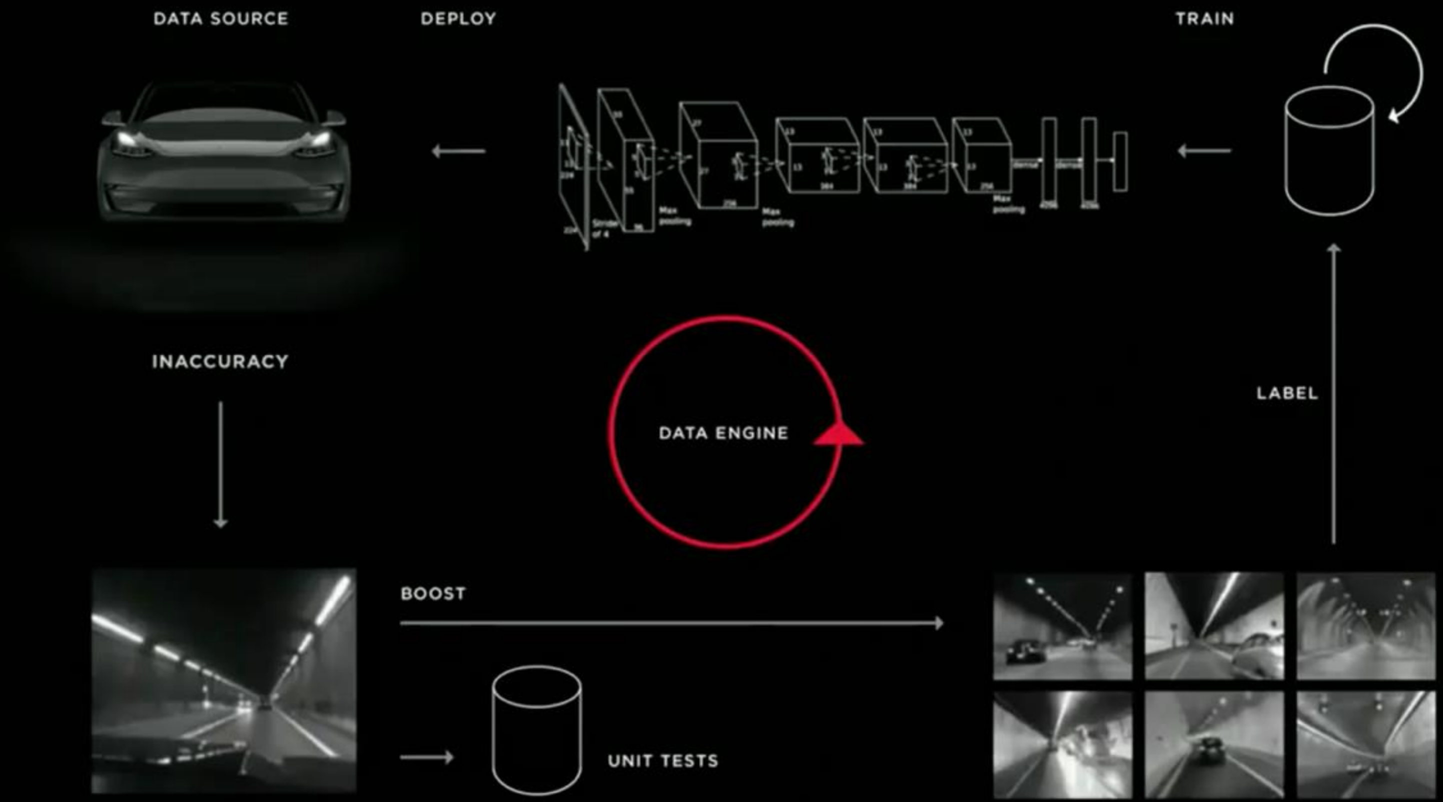
- Notes and tags
- Record of what data belongs to various training, test and validation datasets
- Error analysis findings

Data/ML Ops Pipelines

Elastic Scalability is key!

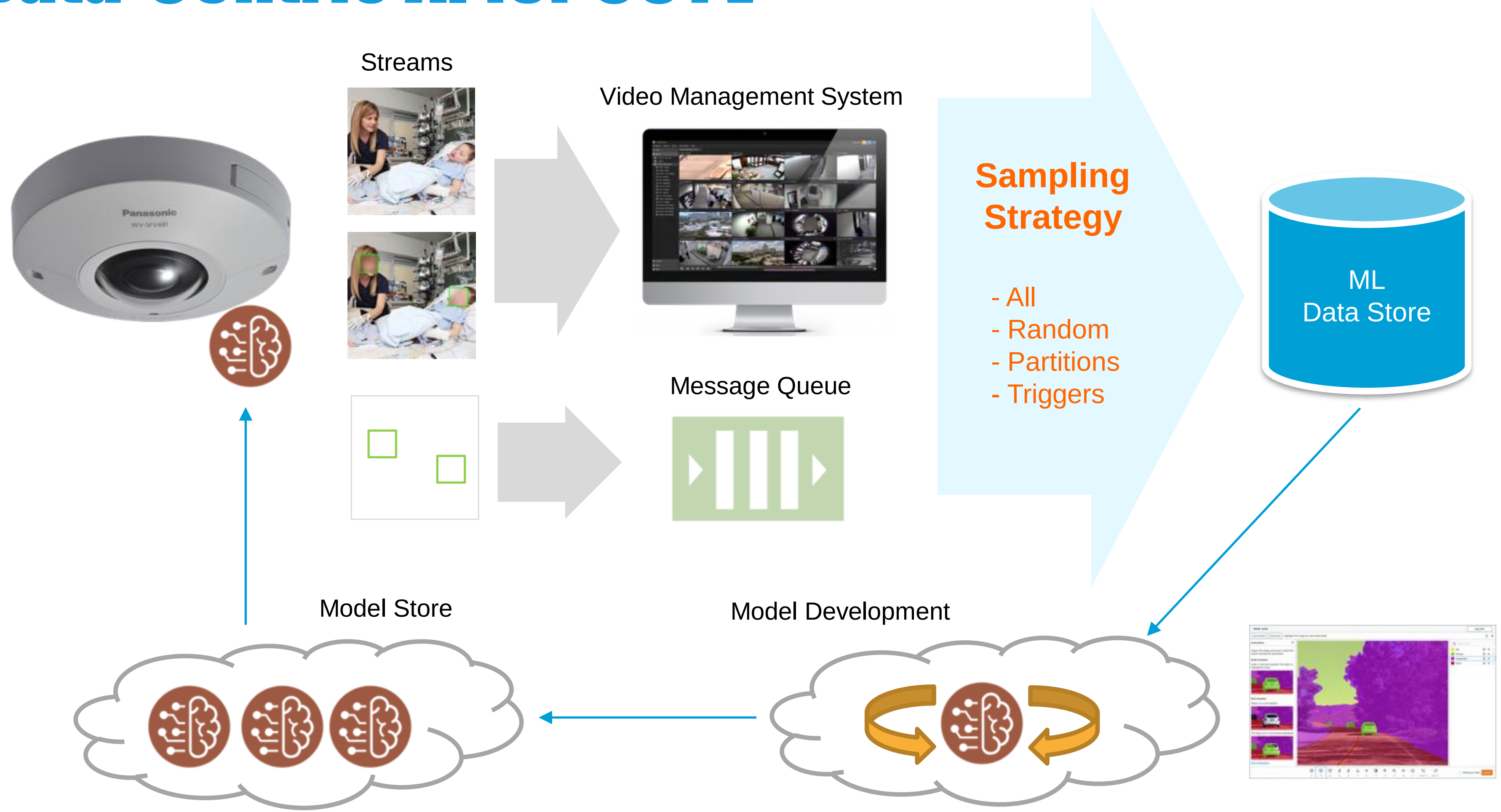


Tesla Data Engine

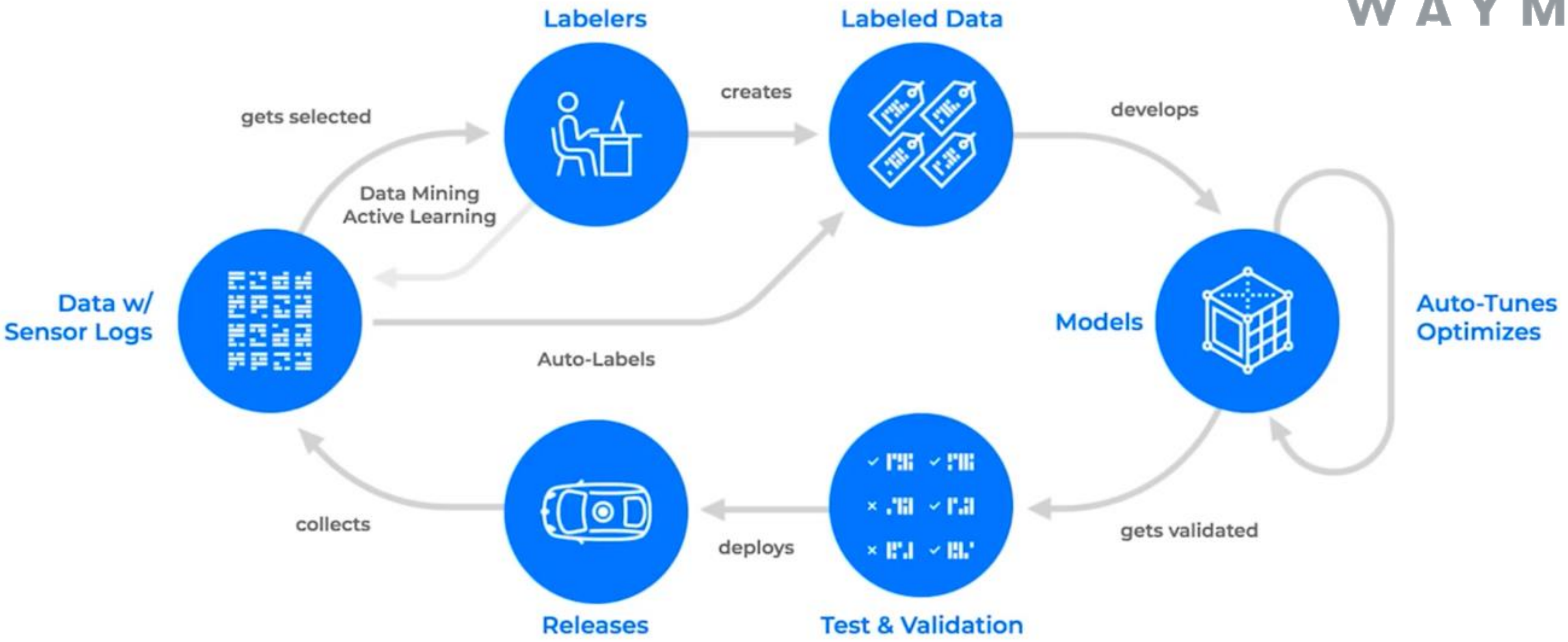


https://youtu.be/j0z4FweCy4M

Data-Centric AI for CCTV



ML Factory For Self Driving Models



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

Deidentification: Named Entity Recognition (NER)

- 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

Source: Aiimi Labs

Groundtruth Definition

PII
- Date
- Name
- ID
- Age
- Contact
- Location
- Profession

When is an entity PII?
- Date
  - Time, Day of Week
  - Do we include “of” in ”12th of March”
- Name
  - Do we include “Dr” in “Dr Brent Richards”
- …

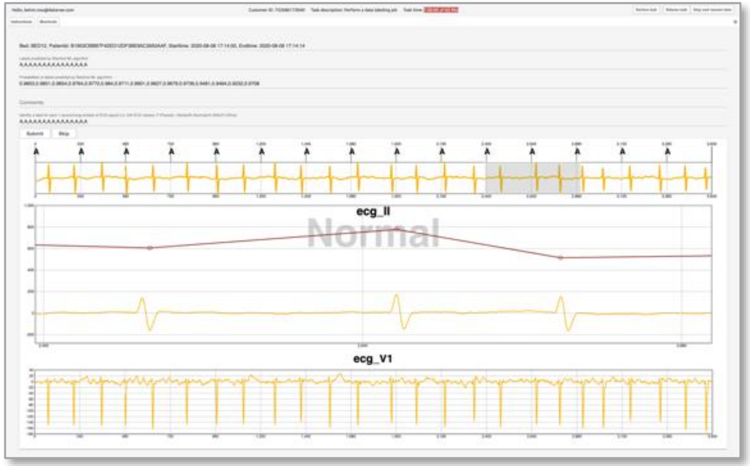
Ground Truth Labeling

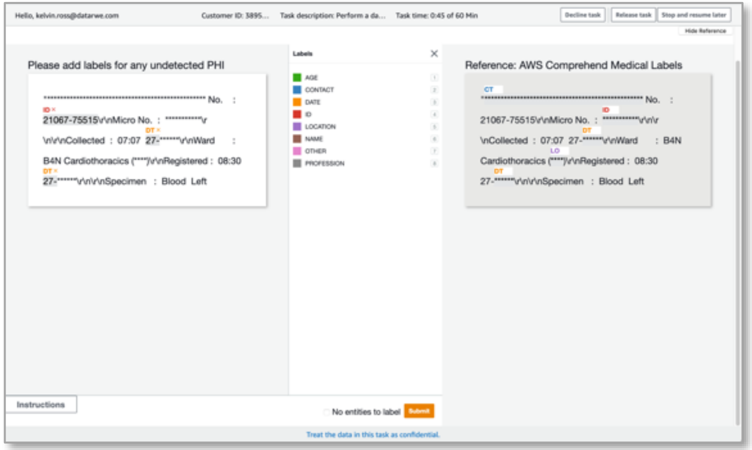
Labeling tasks undertaken with different levels of clinical expertise:

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]

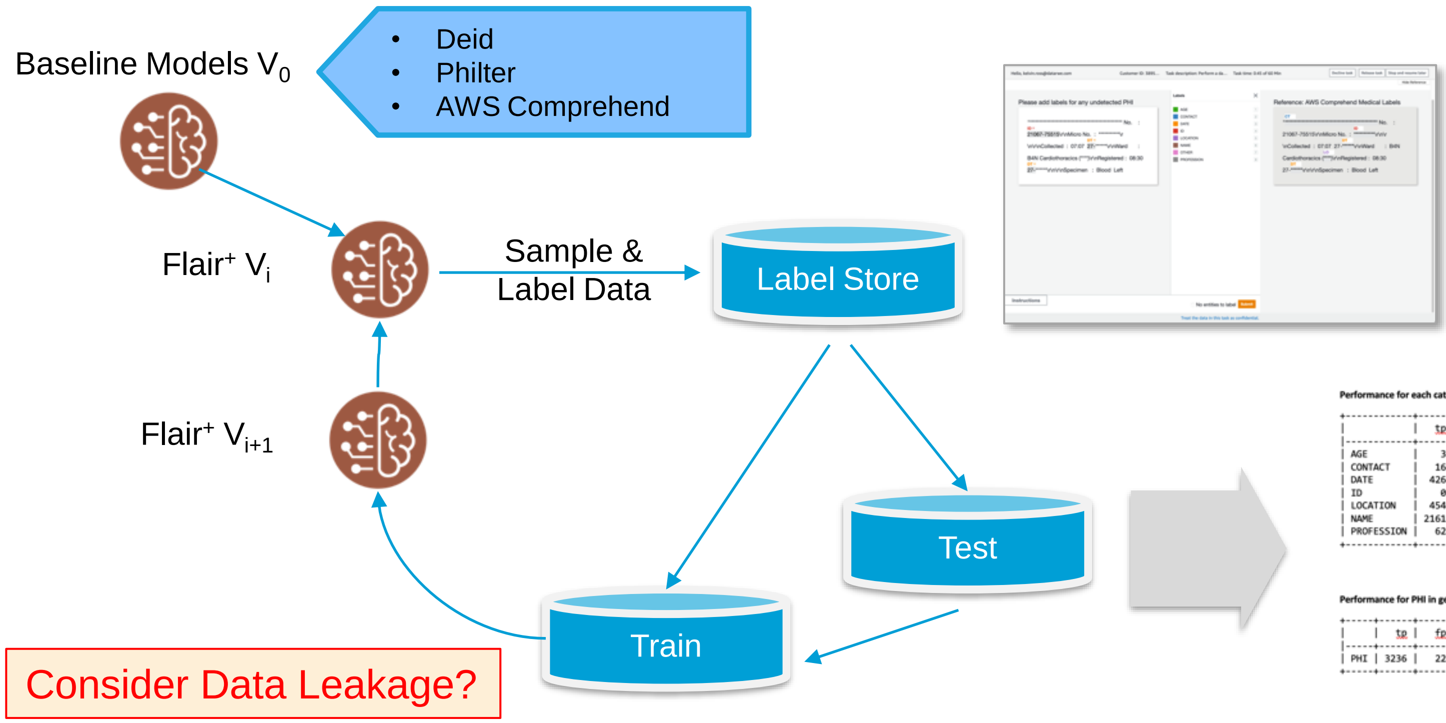
Cognito





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

Iterative Improvement



Performance for each category:

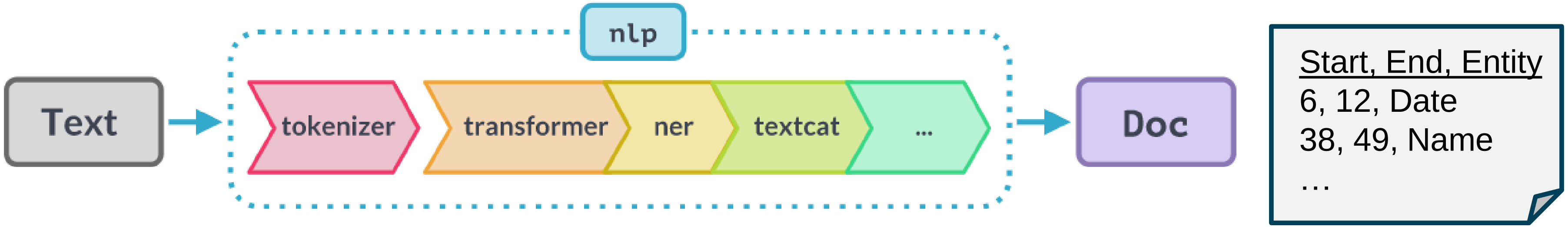
	tp	fp	fn	precision	recall	F1
AGE	3	0	0	1	1	1
CONTACT	16	5	60	0.761905	0.210526	0.329897
DATE	426	0	25	1	0.944568	0.971494
ID	0	60	5	0	0	0
LOCATION	454	38	10	0.922764	0.978448	0.949791
NAME	2161	33	161	0.984959	0.930663	0.957042
PROFESSION	62	0	29	1	0.681319	0.810458

Performance for PHI in general:

	tp	fp	fn	precision	recall	F1
PHI	3236	22	176	0.993247	0.948417	0.970315

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

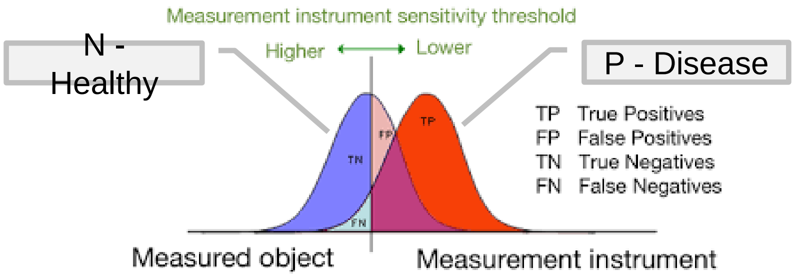


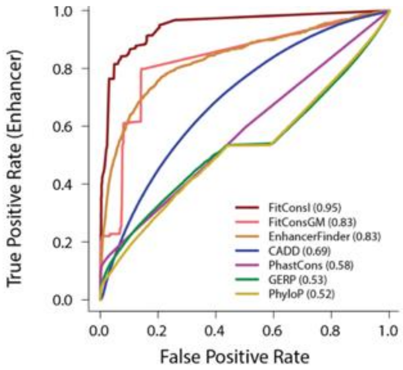
Source: spacy.io

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

Evaluation Metrics

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	





https://en.wikipedia.org/wiki/Receiver_operating_characteristic

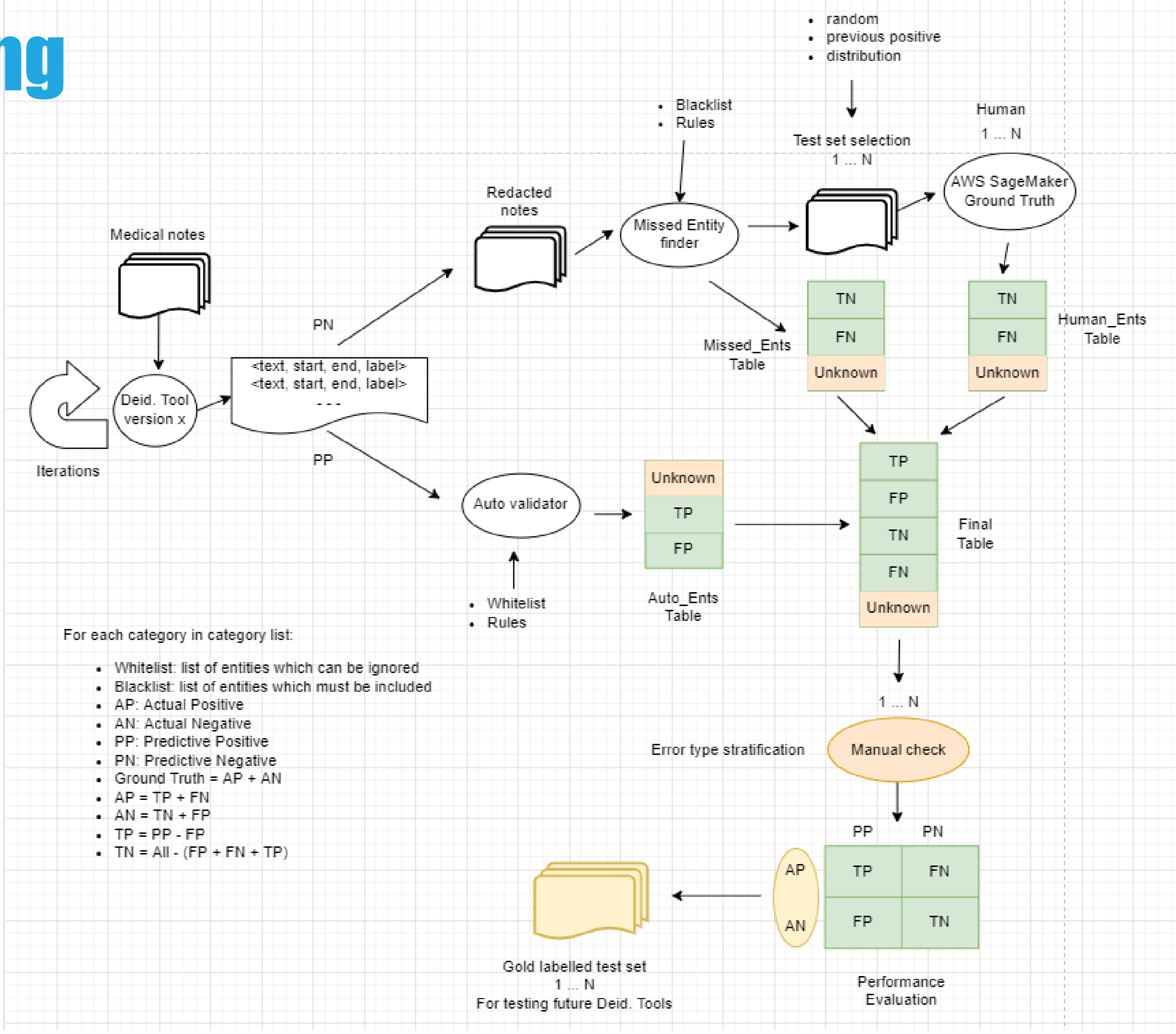
Evaluation Metric

Admitted to ICU for observation and management of W1F4 SAH (*******), complicated by vasospasm of ACA & MCA, requiring 6 sessions of IA verapamil.
********* coiled on *******.

Issues with polyuria, requiring sodium replacement

Entity	Token	Character
Kelvin	Kel-vin	K-e-l-v-i-n

Auto Assist Labelling



Error Analysis

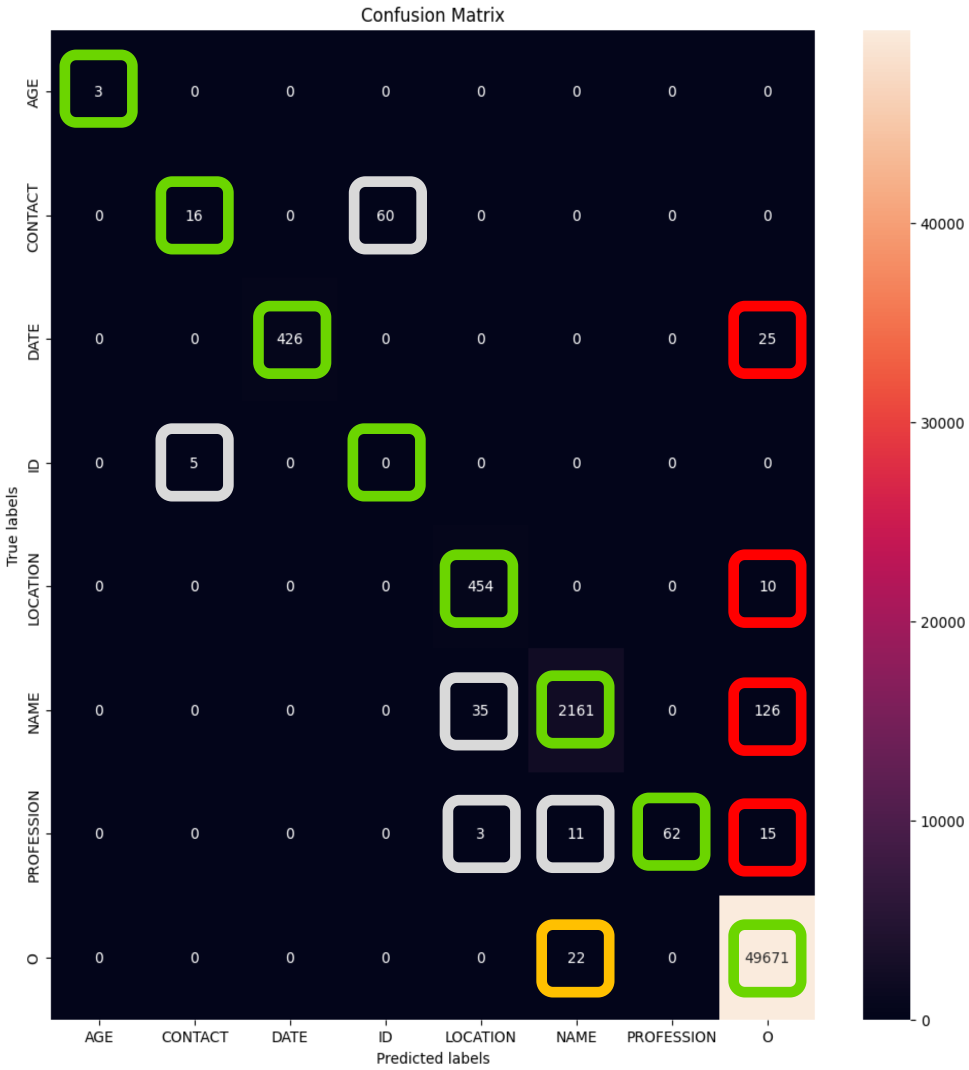
Deidentification NER on Family Notes

Performance for each category:

	tp	fp	fn	precision	recall	F1
AGE	3	0	0	1	1	1
CONTACT	16	5	60	0.761905	0.210526	0.329897
DATE	426	0	25	1	0.944568	0.971494
ID	0	60	5	0	0	0
LOCATION	454	38	10	0.922764	0.978448	0.949791
NAME	2161	33	161	0.984959	0.930663	0.957042
PROFESSION	62	0	29	1	0.681319	0.810458

Performance for PHI in general:

	tp	fp	fn	precision	recall	F1
PHI	3236	22	176	0.993247	0.948417	0.970315



Error Analysis

Entity	Entity Type	Severity
Age > 89	AGE	High
Age =< 89	AGE	Low
Email id	CONTACT	High
Mobile number	CONTACT	High
Full date with Date, Month & Year	DATE	High
Partial date – Date & Month / Month & Year	DATE	High
Partial Date – Only date or Month or Year	N/A	N/A
Day	N/A	N/A
Time	N/A	N/A
Driving Licence Number	ID	High
Medicare Number	ID	High
UR Number	ID	Medium
Ethnicity	LOCATION	High
Nationality	LOCATION	High
Street address	LOCATION	High
City / Suburb of residence / care	LOCATION	Medium
Post code of residence / care	LOCATION	Medium
Hospital abbreviation / name	LOCATION	Medium
State of residence / care	LOCATION	Low
Country of residence / care	LOCATION	Low
Full Name	NAME	High
Last Name	NAME	High
First/Middle Name – not common	NAME	High
First/Middle Name – common	NAME	Medium
Profession – patient	PROFESSION	High
Profession / Designation - staff	PROFESSION	Low

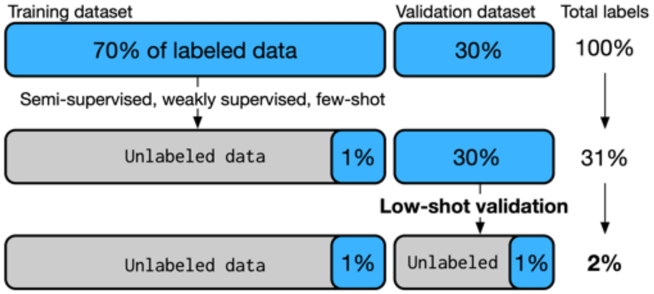
Kelvin J. Ross

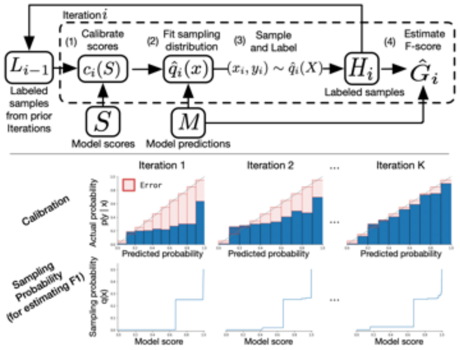
****** J. ****

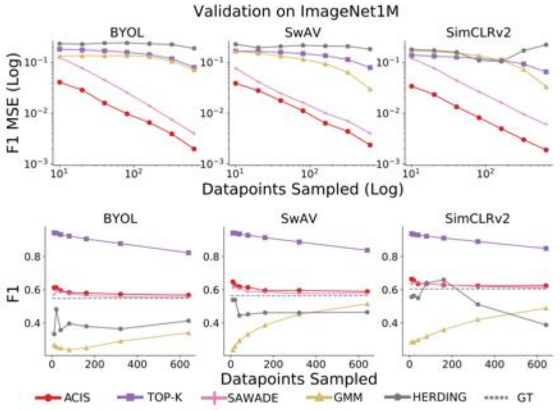
PHI Type	# True characters	# FNs (% of True)	# FNs of High severity (% of True)	# FNs of Medium severity (% of True)	# FNs of Low severity (% of True)	# FNs tagged incorrect PHI Type (% of FNs)
NAME						
CONTACT						
ID						
LOCATION						
DATE						
AGE						
PROFESSION						

Importance Sampling

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







https://arxiv.org/abs/2109.05720

Label Reasoning

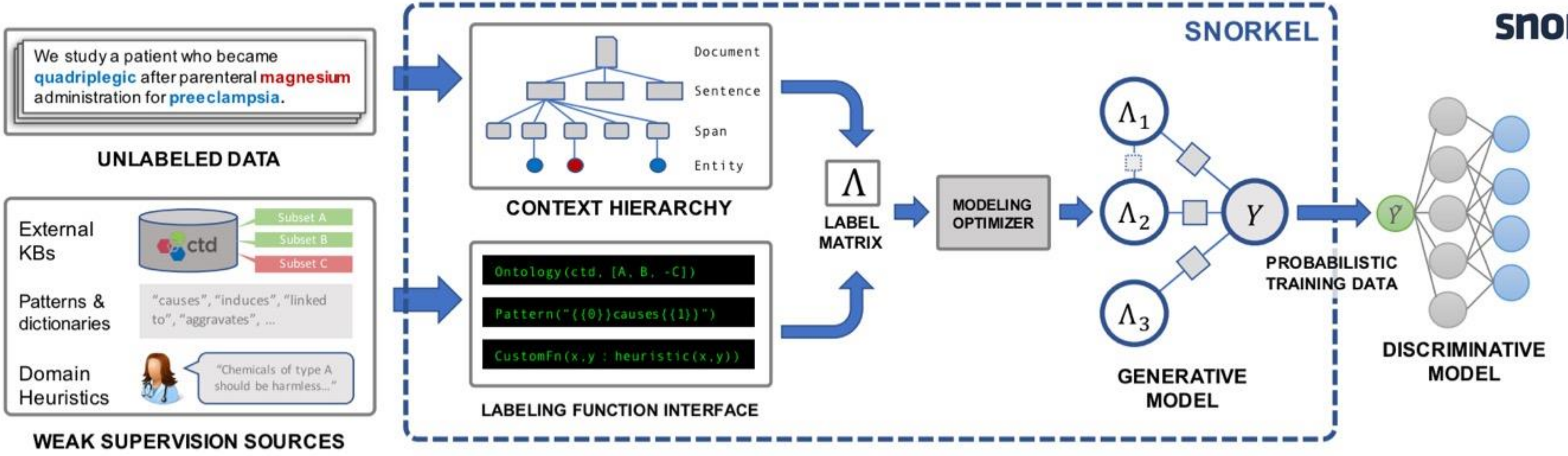
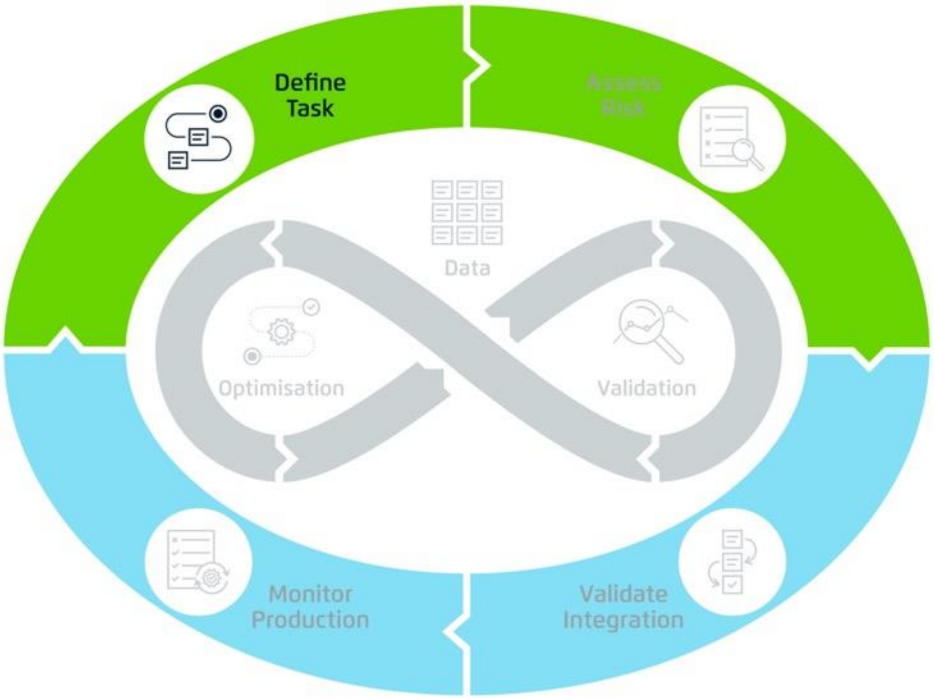


Figure 2: An overview of the Snorkel system. (1) SME users write *labeling functions (LFs)* that express weak supervision sources like distant supervision, patterns, and heuristics. (2) Snorkel applies the LFs over unlabeled data and learns a generative model to combine the LFs' outputs into probabilistic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

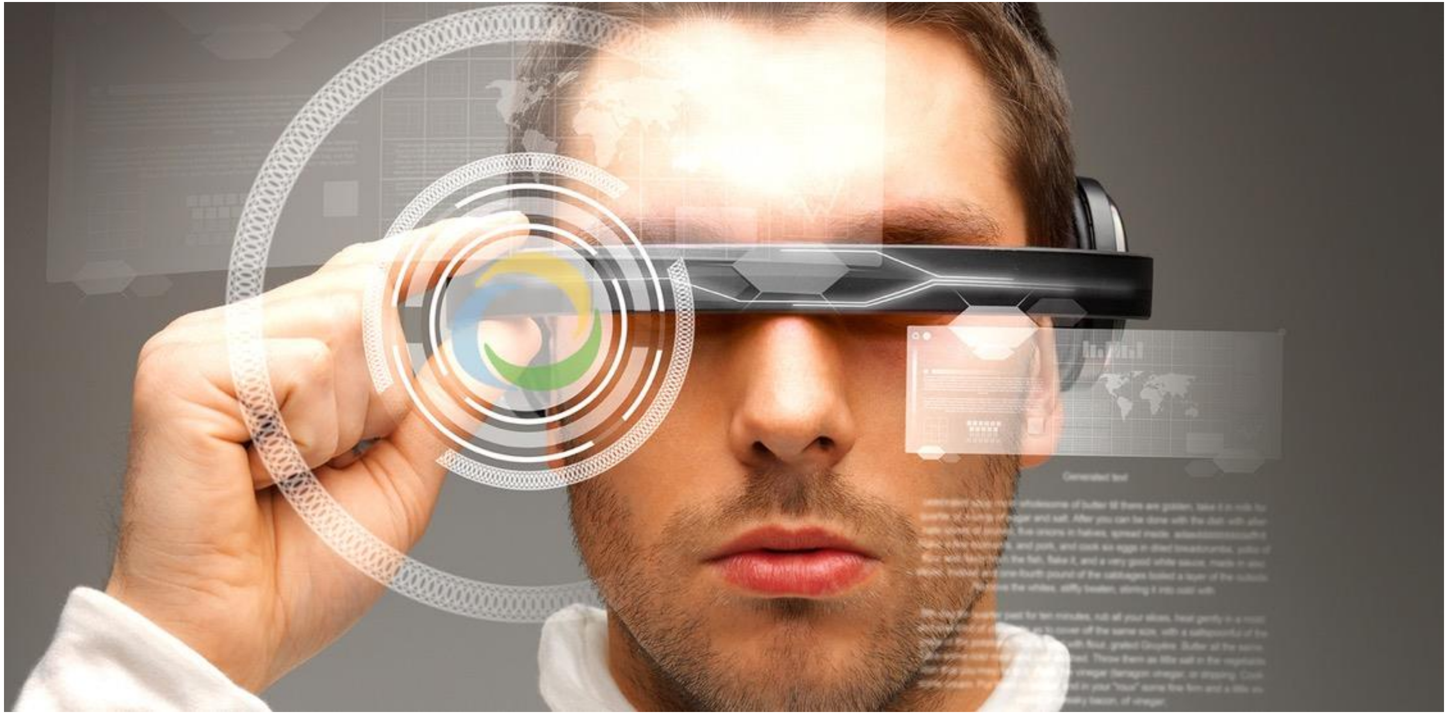
https://www.snorkel.org/blog/weak-supervision

Define Context: Define Task



- What is the task to be modelled.
- How is the task specified.
- What guidance is there for ground truth labelling.
- How might ground truth between labellers vary and confuse training and validation.

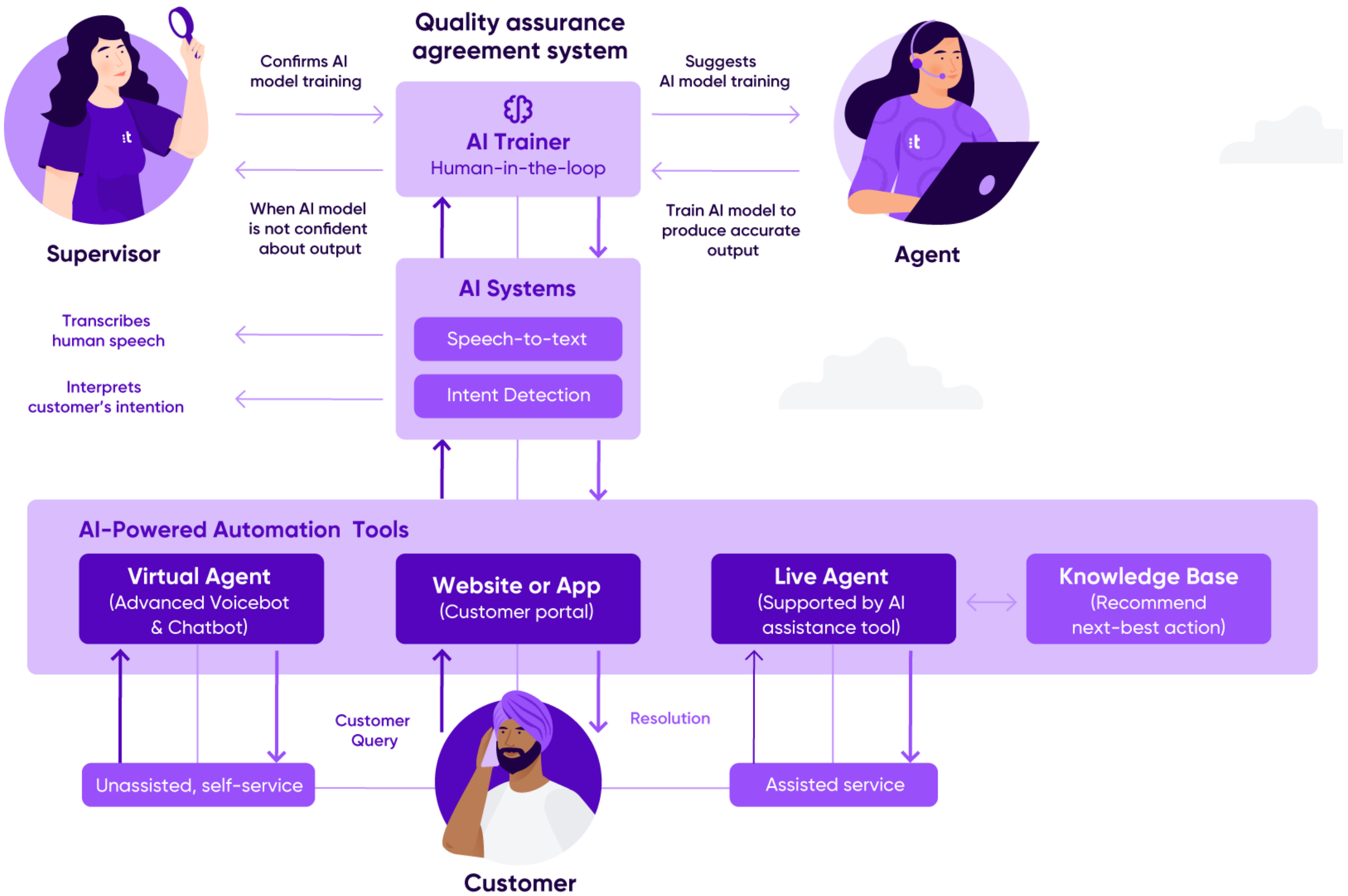
Human Augmentation



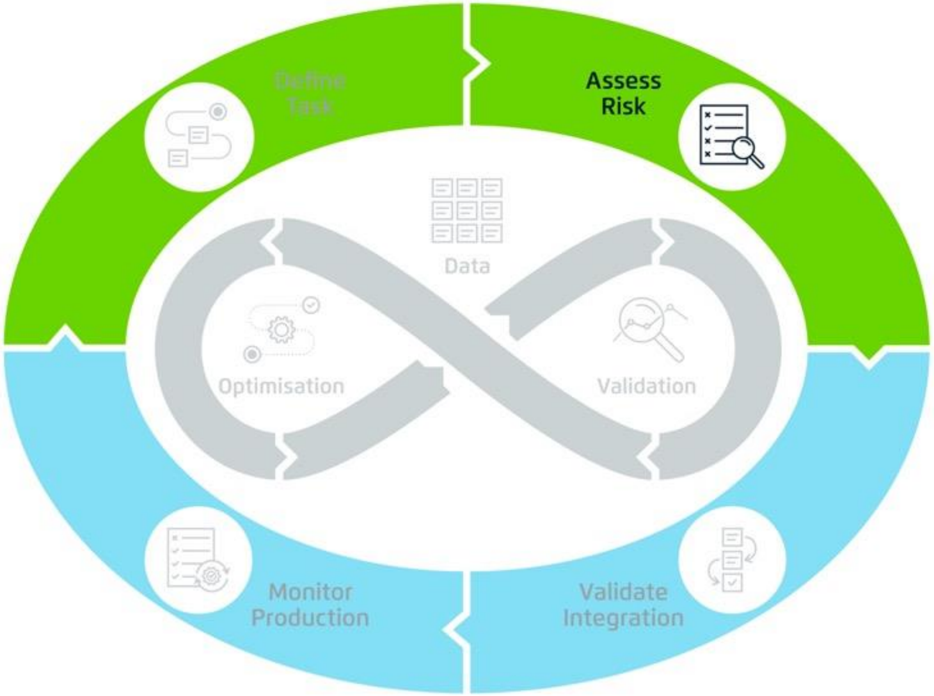
Error Rate in detection of cancer in lymph node cells	Human Pathologist	AI	Human Pathologist + AI
	3.5%	**7.5%**	**0.5%**

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



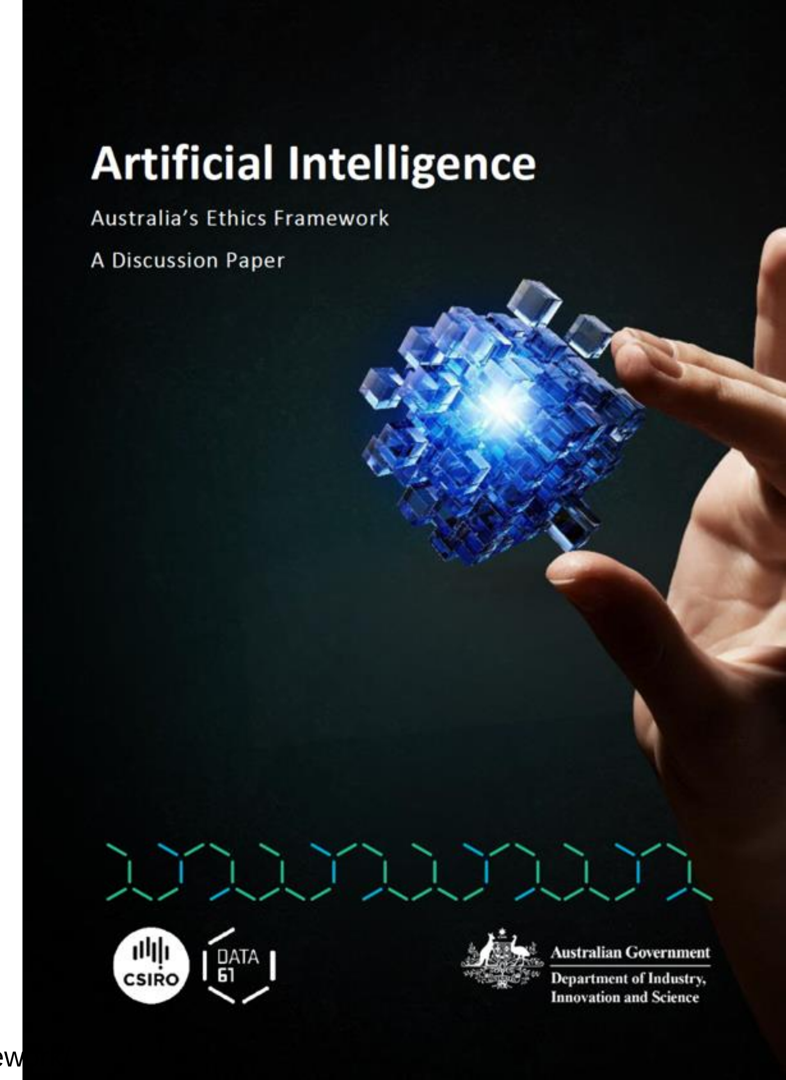
Define Context: Assess Risk



- What is the risk profile?
- Are False Negatives and False Positives equally bad?
- Are some input and output classes equally risky, or some riskier than others. E.g. False negative on a melanoma much more serious than a false negative on a benign skin lesion.
- What is the context of how the model is being integrated?
- Is human-in-the-loop checking involved, or is it real-time autonomous decision making?

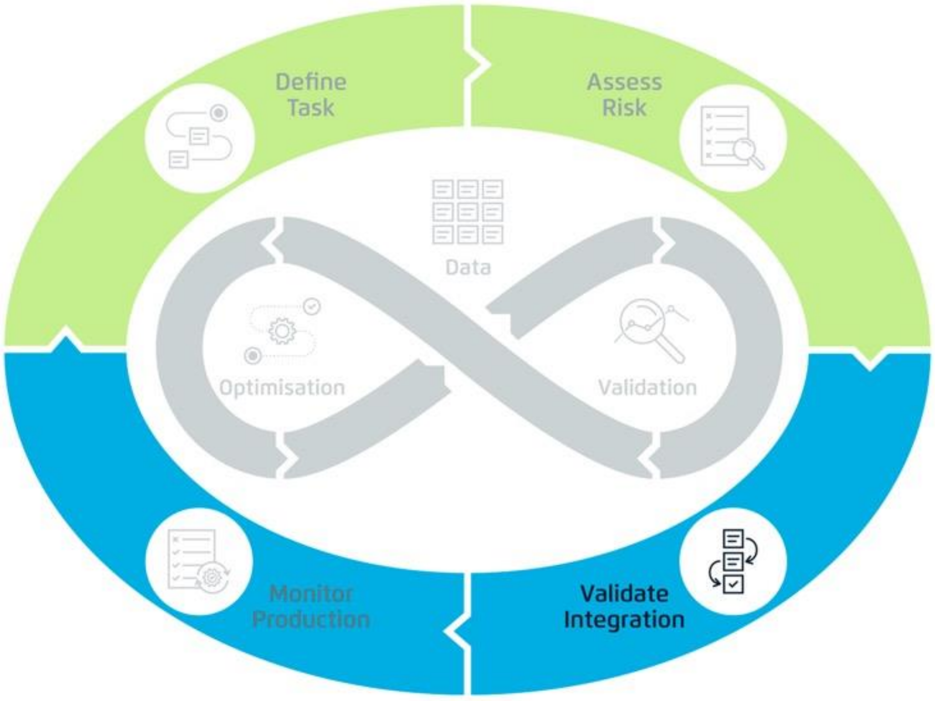
Responsible AI

Core principles for AI	
1. Generates net-benefits. The AI system must generate benefits for people that are greater than the costs.	**2. Do no harm.** Civilian AI systems must not be designed to harm or deceive people and should be implemented in ways that minimise any negative outcomes.
3. Regulatory and legal compliance. The AI system must comply with all relevant international, Australian Local, State/Territory and Federal government obligations, regulations and laws.	**4. Privacy protection.** Any system, including AI systems, must ensure people's private data is protected and kept confidential plus prevent data breaches which could cause reputational, psychological, financial, professional or other types of harm.
5. Fairness. The development or use of the AI system must not result in unfair discrimination against individuals, communities or groups. This requires particular attention to ensure the "training data" is free from bias or characteristics which may cause the algorithm to behave unfairly.	**6. Transparency & Explainability.** People must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions.
7. Contestability. When an algorithm impacts a person there must be an efficient process to allow that person to challenge the use or output of the algorithm.	**8. Accountability.** People and organisations responsible for the creation and implementation of AI algorithms should be identifiable and accountable for the impacts of that algorithm, even if the impacts are unintended.



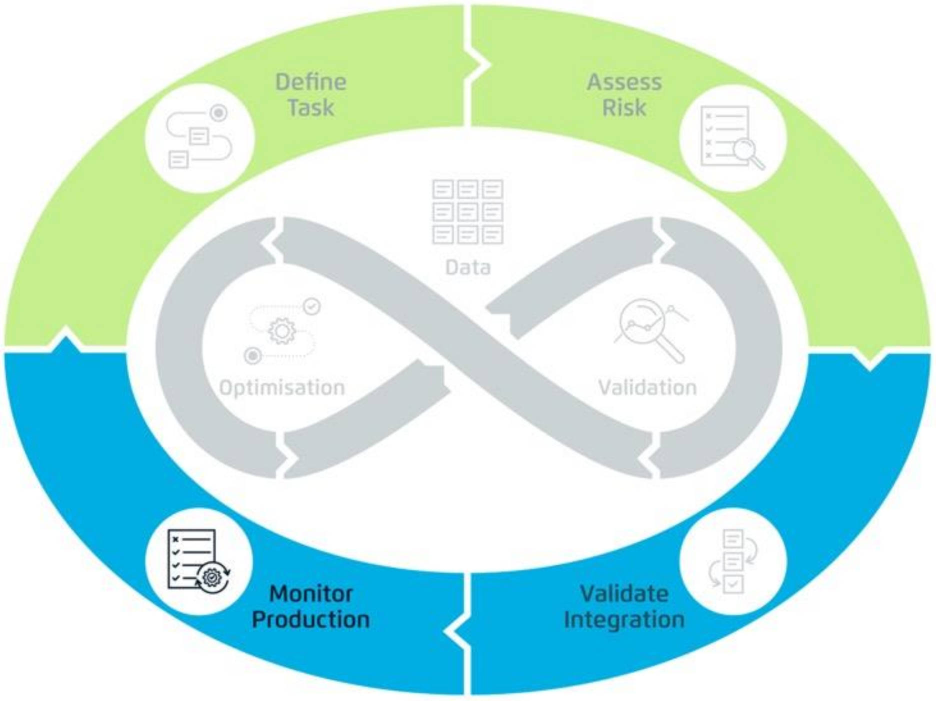
https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framew

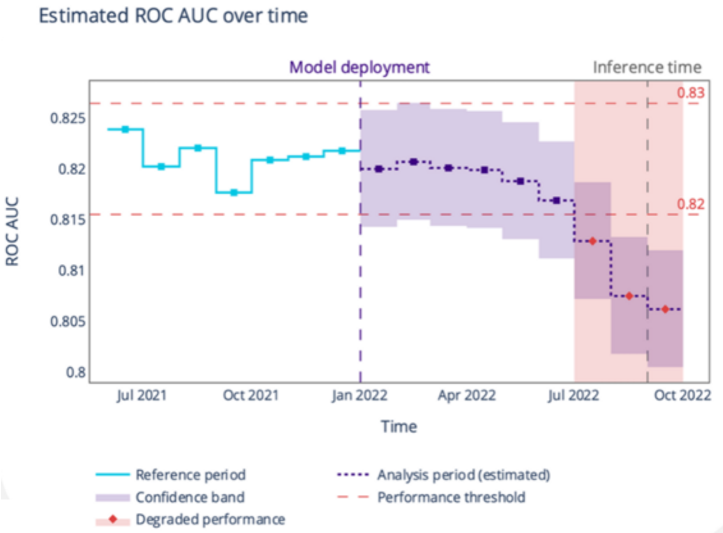
Govern Behaviour: Validate Integration



- ML models sit within a broader system.
- Integration within the system must be validated.
- Are model inferences further extended within the system to interpret or translate the results?
- E.g. are the model outputs are then incorporated in imperative code to map values to specific classes, or the way that results are written to the system/ API / database to record results.
- ML model validation is essentially Unit Testing, after which we need to run layers of integration, system and acceptance tests.
- Even if the development process is based on an iterative DevOps or MLOps driven process, this stage of the lifecycle typically requires an element of V-model testing in which system integration testing, performance testing, acceptance testing all draw upon key requirements from the broader context.

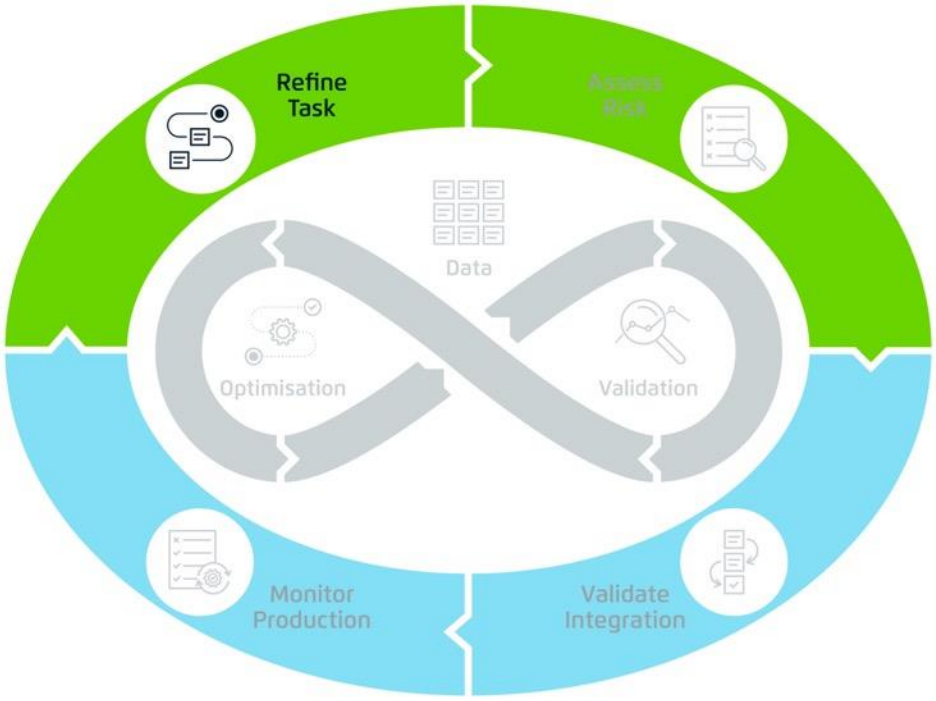
Govern Behaviour: Monitor Production





https://www.nannyml.com/blog/3-common-causes-of-ml-model-failure-in-production

Understand Context: Refine Task



Once a model is in production there is a need to:

- Continually review performance results.
- Consider new risks emerging from other activities.
- Update task definition and risk assessment.
- Propose activity reiteration for redevelopment.

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

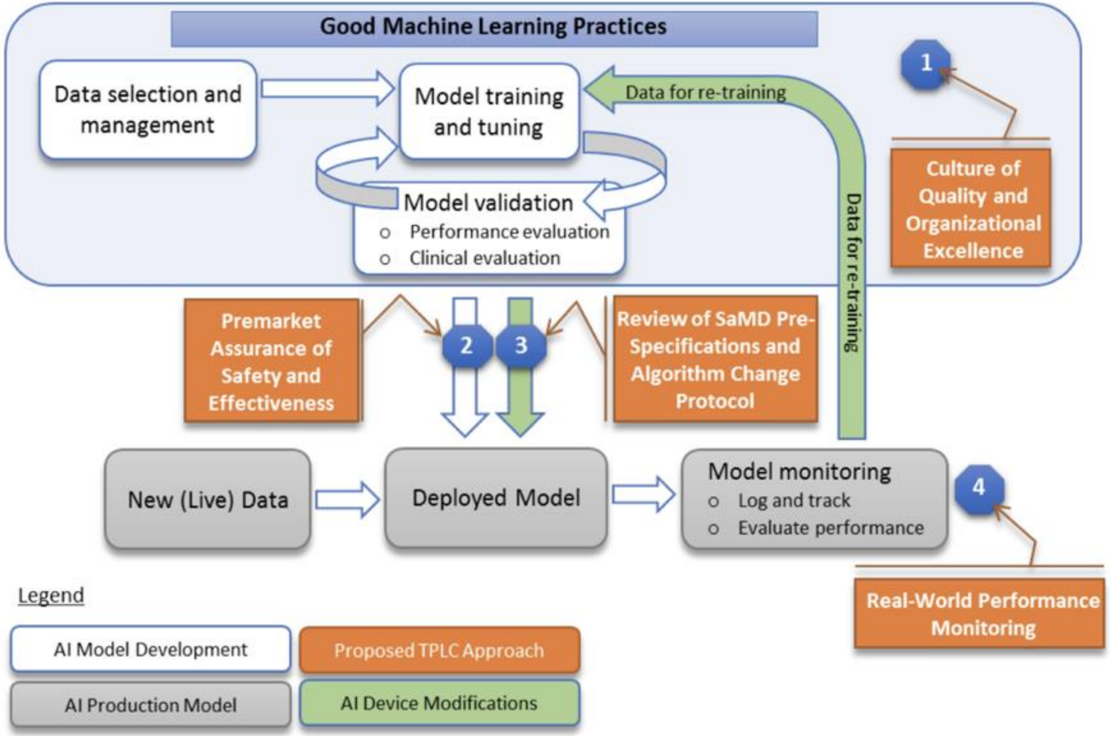
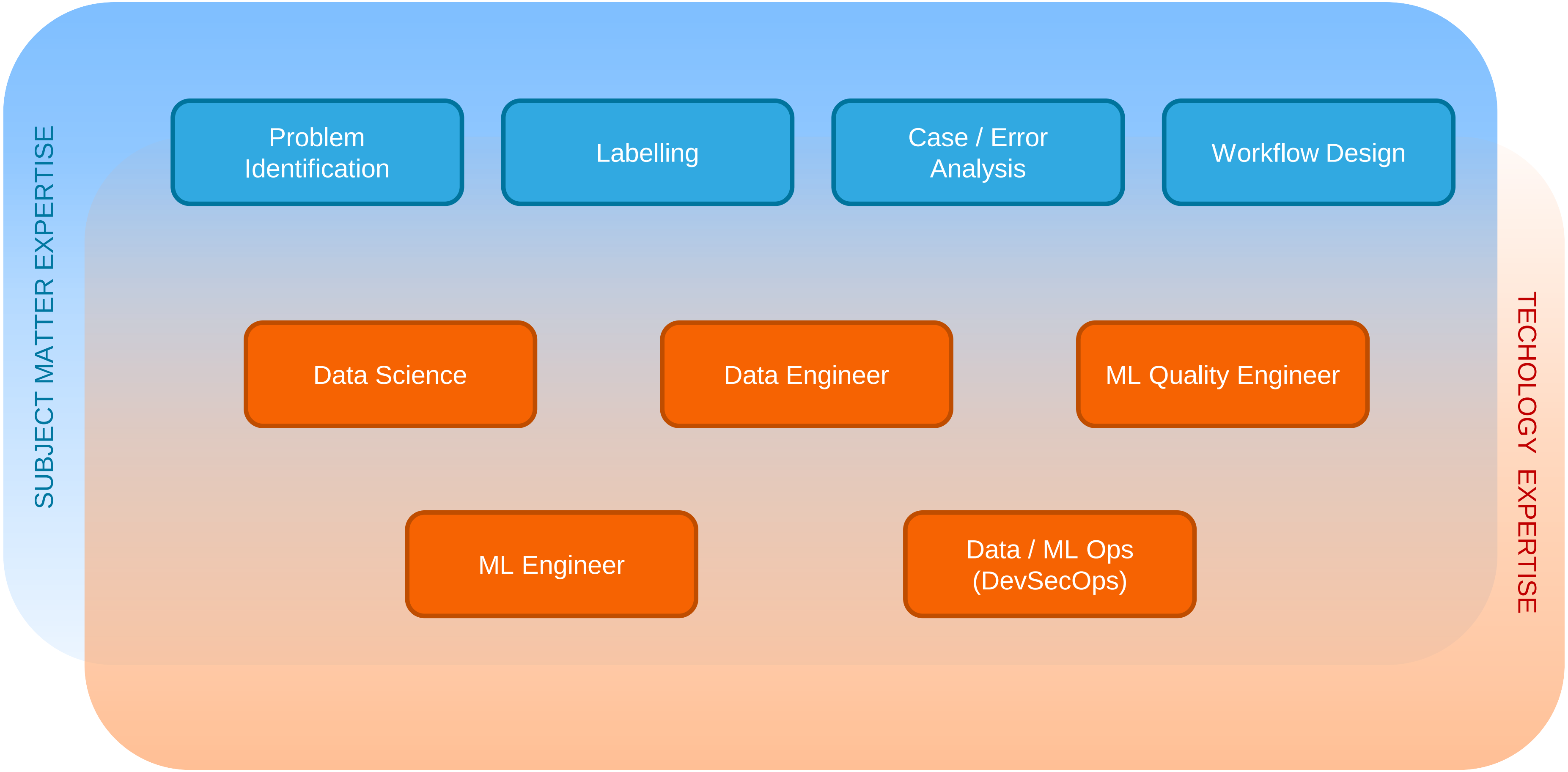


Figure 2: Overlay of FDA's TPLC approach on AI/ML workflow

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

ML: Team Sport



Key Takeaways

1. ML has improved capability dramatically
2. Shift from Software 1.0 to Software 2.0
3. Various validation considerations to understand limitations of ML model
4. Key to optimizing towards robust models

Questions?

kelvin.ross@kjr.com.au
0414 505 910
@kelvinjross