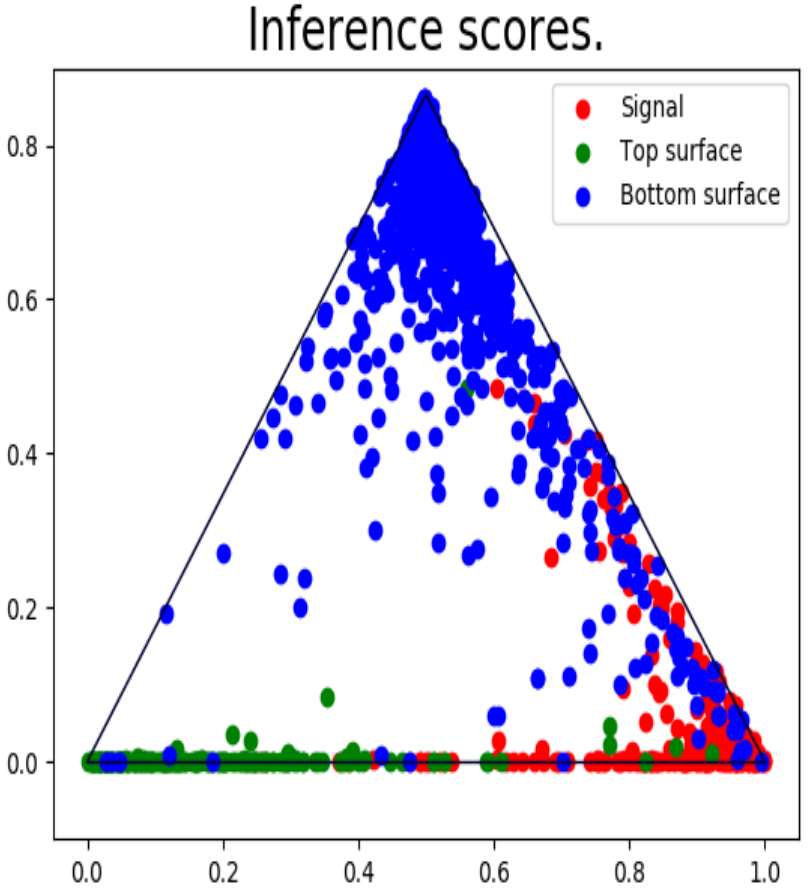

Queen Mary
University of London

# ML – 22 Mar Update

l.millward@qmul.ac.uk

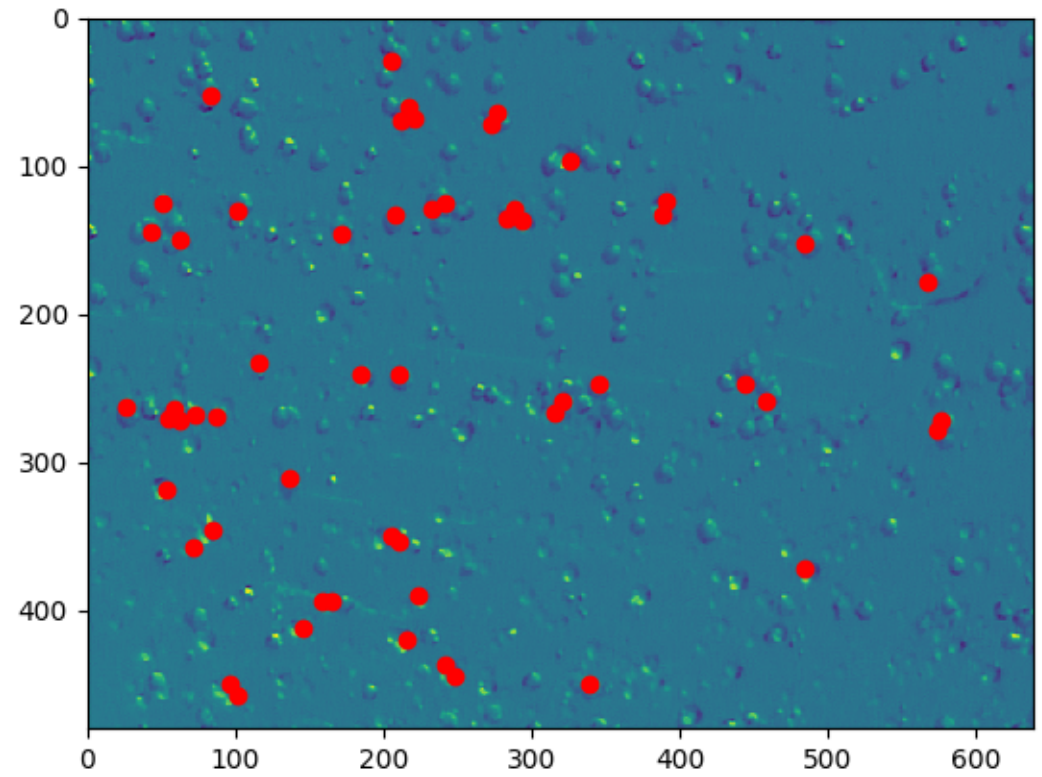
# Recap: Ensemble

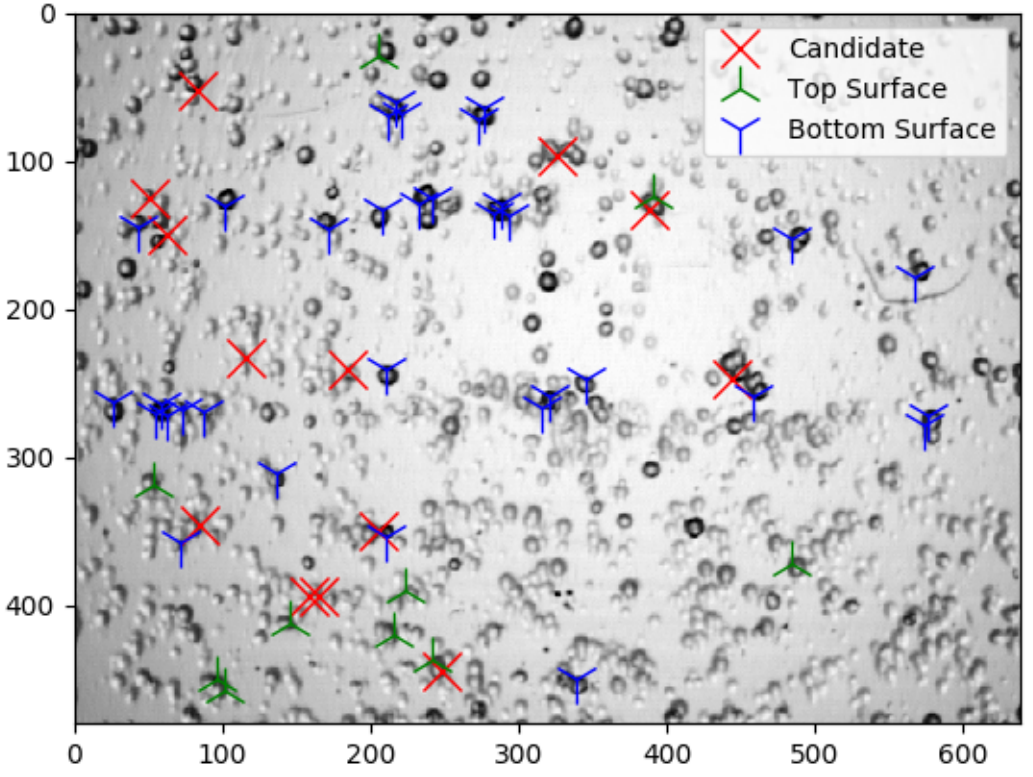
- Combining 3 Experts
  Non-linear ensemble
- Each trained and tested for given sub-classification
  eg, ‘Top vs bottom’
- Robust and accurate for each sub-task
  each expert 97%+ stable validation
- overall ensemble classification




Ternary Plot; 2d normalised representation of 3-tuple classification score.
Color = Truth label

# Recap: Inference




- Can be applied as inference tool on unseen foil area
- Pre-selection (old) – finds pits of interest in 3d feature space (hip, entry / exit)
- Can then sub-classify via ensemble
- ( bottom=front range out , top ~ range in, sig = in-out pair)
- Visual Inspection of 3d feature space to confirm

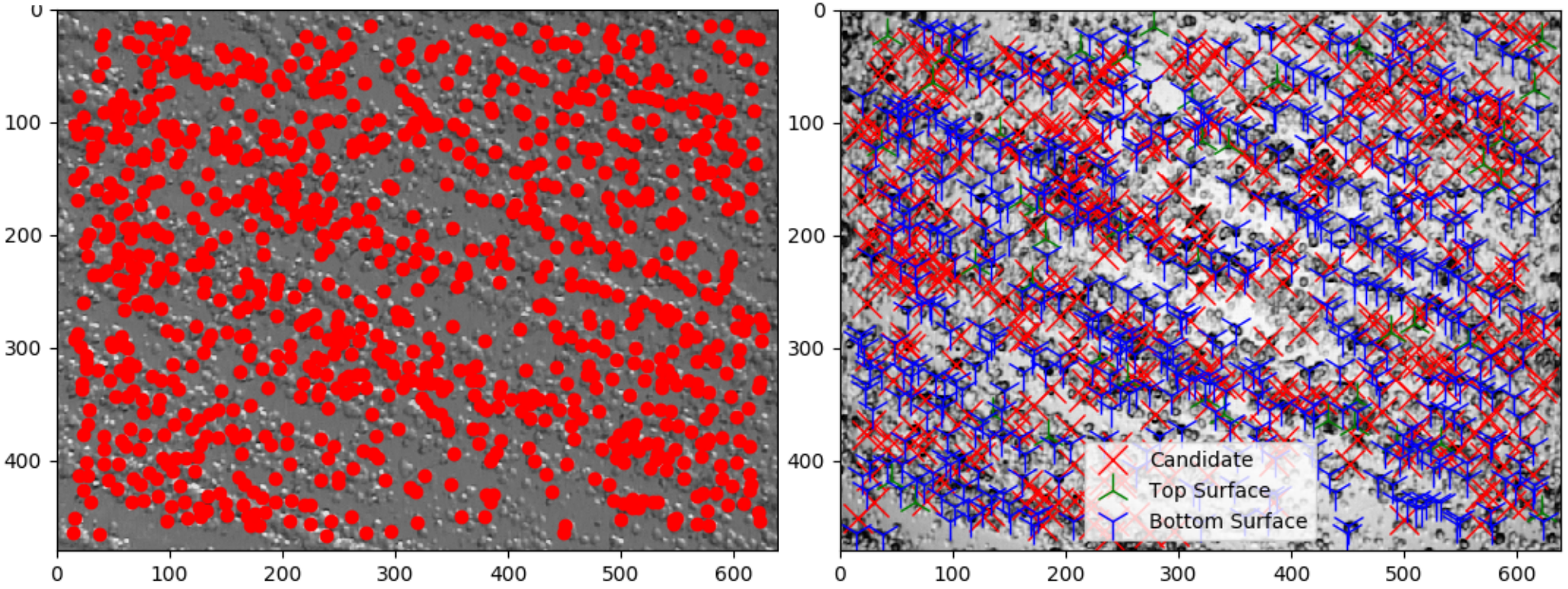
# Recap: Summary

- Ensemble classifier stronger than S / B  or multi-label classifier
- Trained with clustered background, ie can handle overlaps
- Understandable decision-making process + easy to extend to new subcatergories of background and signal
- Looked at different hyper-params / optimisers etc..

- Glorot – Xavier
  Did find bug with normalisation being lost between
  array – image – dataset
- Batch norm issues
- Weaknesses;
  Trained on data under a certain pit density limit
- Will start to break down when pushing significantly beyond this limit

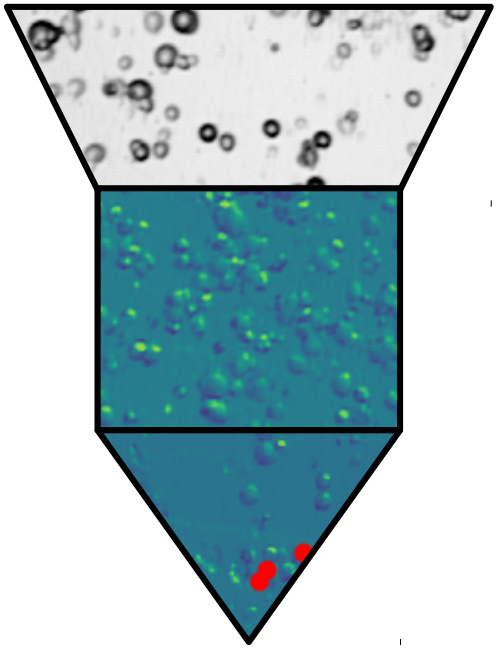
Example……..

# Recap: Background limit




- Eg, ~1000 pits per image heavy background region
- Preselection breaks down – because even uncorrelated pits look like entry exit pairs
- Classifier stage relies on pre-selection to substantially whittle down number of objects for further classification.

# Pipeline: Current Work



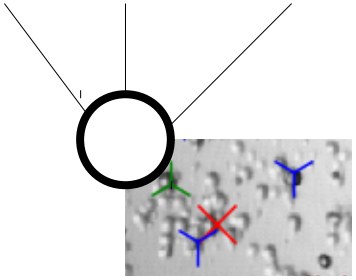
Raw Data

Image processing
Normalisation

Pre-Selection

Labeling
Manual / Inference

Trained sub-classifiers

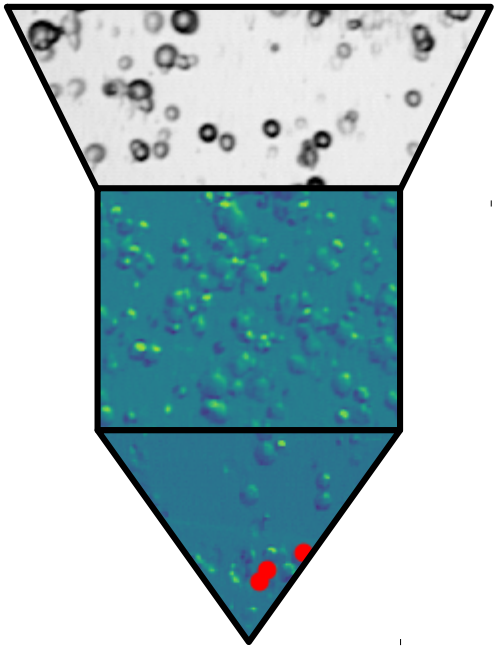


Ensemble / output

- Bug in normalisation
- Quick fix for TA inference
- Fixing properly / prep for new data
- Dictionary / meta-data track at different levels
  [foil] [image] [etch-pit]

  scope cfg, location, label, pit density
- easy to add / remove data
  # prevent data excess from scanning

# Pipeline: Current Work

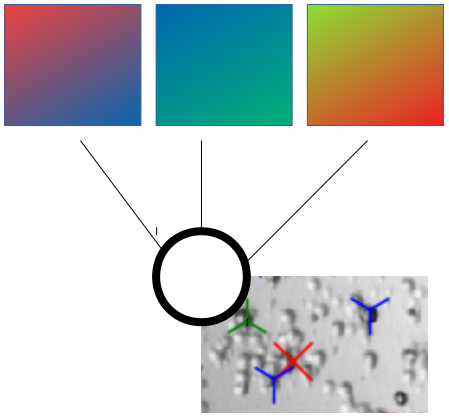


Raw Data

Image processing
Normalisation

Pre-Selection

Labeling
Manual / Inference

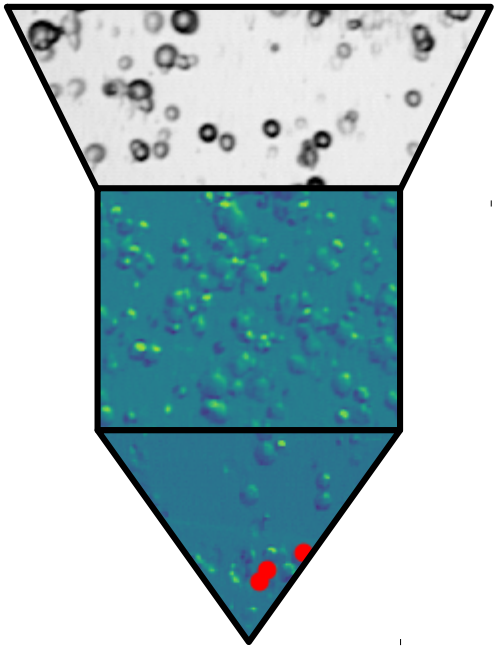


Trained sub-classifiers

Ensemble / output

- Look at optimising decision boundary within this space
- Currently going on simplistic ‘highest score’
- Decide on ROC / working point
- Can ‘stack’ NN’s ie using another NN to decide on final object classification
- Extend classes / subtypes allowed

# Pipeline: Next – higher bkg



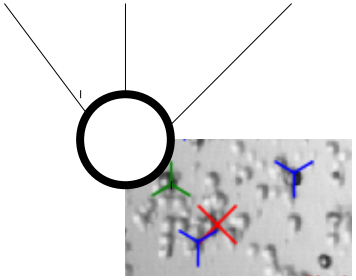
Raw Data

Image processing
Normalisation

Pre-Selection

Labeling
Manual / Inference

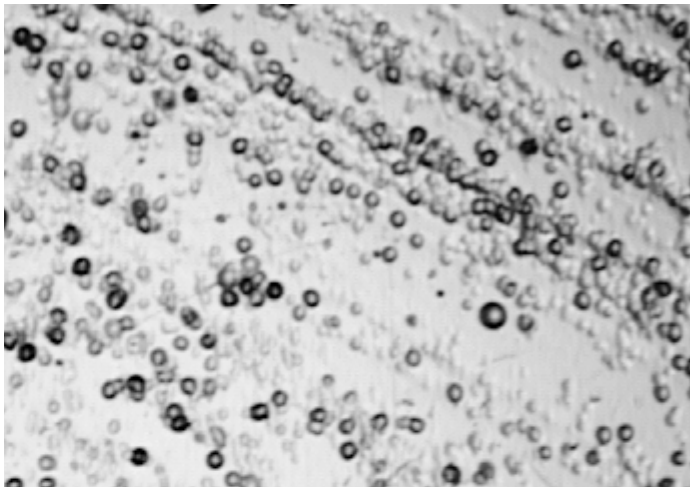
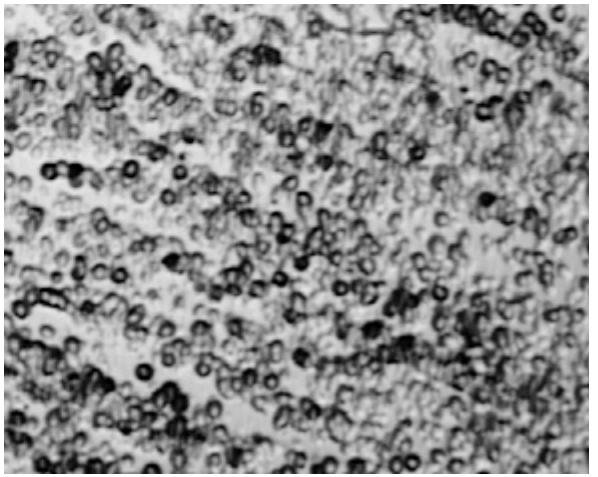
Trained sub-classifiers



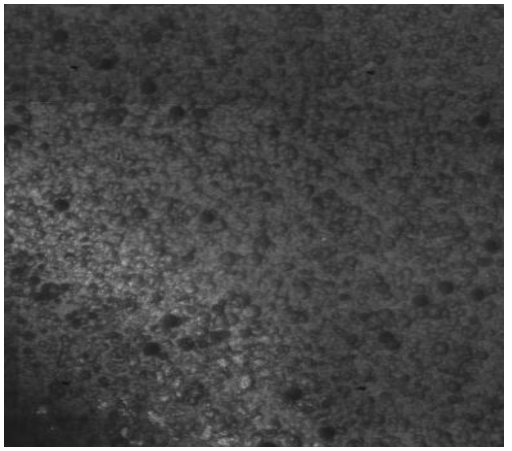
Ensemble / output

- Return to classic image processing – new tricks
- Post processing / clean up on very ‘dirty’ images
- Can still spot anomalies in very dense pile-up by looking at processed gifs
- Think about different spallation backgrounds
- Map densities on foil(s)
- ML will/(should) behave differently in different areas of foil - optimise

# Background mapping

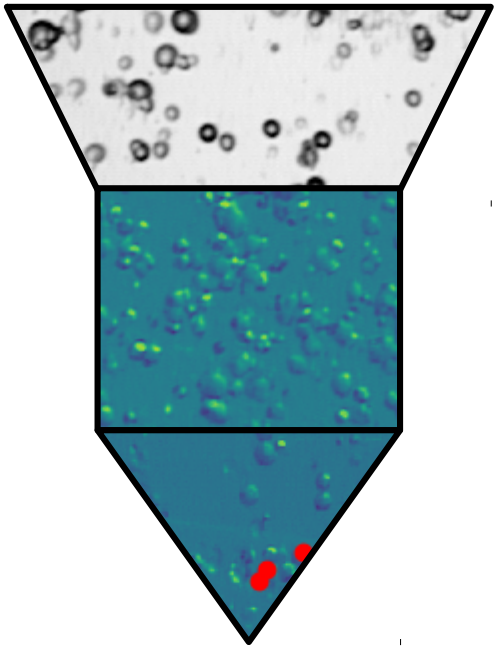




Different pit density in different regions of NTD foil
Above = 8 months exposure makrofol
Side = image from ‘seite’ exposure – ps: could do with more access to presious NTD analysis data



Represent different domains, needing different techniques,
Different challenges;
Low Bkg – want to compress whitespace, easy pattern matching
Med Bkg – want ML to tackle complex clustering / find ROIs
High Bkg – Identify where ML will / wont work
Image processing / clean up, due to high object density want to look at more ‘continous’ methods

# Pipeline: Next – higher bkg



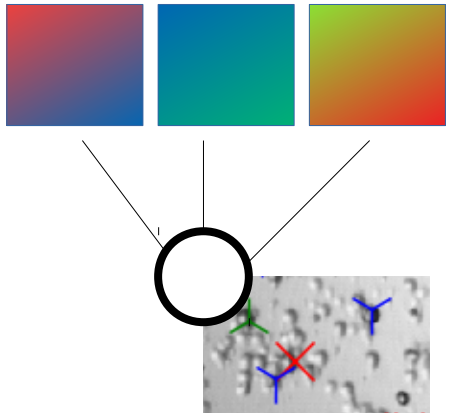
Raw Data

Image processing
Normalisation

Pre-Selection

Labeling
Manual / Inference

Trained sub-classifiers

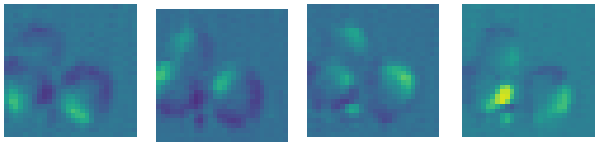


Ensemble / output

- Retrain / transfer learning adding in new data from higher pit density areas of foil. Can look at using inference to boost
- Data Augmentation
- Can look at 3d convoloution for the sub-classifiers
  (was holding off because in short term adds more data per pit to an over-constrained problem, no longer a big problem with robust sub-sclassifiers)

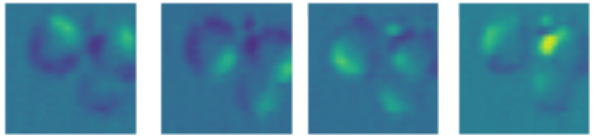
# Data Augmentation

Typical to apply data augmentation, using symmetries to expand labeled data
eg, Rotation
Wont work with the broken symmetry of the fresnell illumination



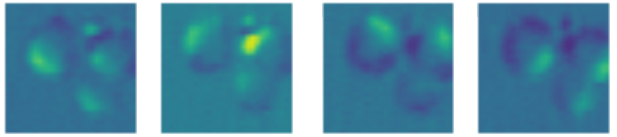
Eg, Cluster of three pits, on top surface of foil
(displayed in 4 of the illumination channels)

---



Simple 90° rotation changes physical interpretation, renders label invalid
Now appears like three bottom surface pits

---

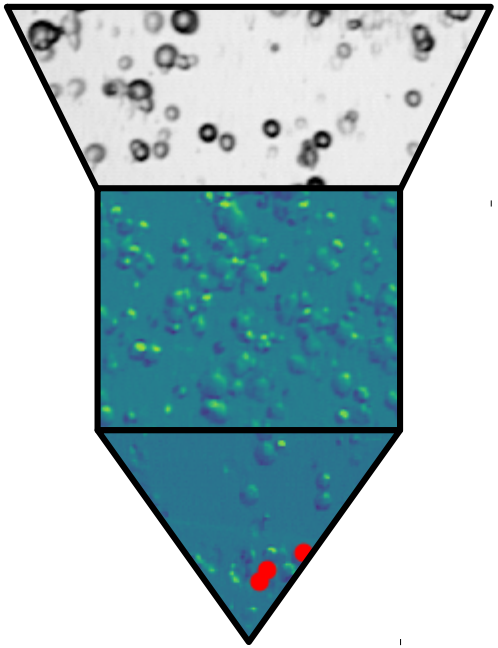


Rotation + Channel Permutation
restores original physical interpretation
(three etch pits, on top surface)

Trivial to implement for x4 augmentation of existing labels and examples (90°)

More fiddling can do x8 augmentation (45°)

# Pipeline: Next – higher bkg



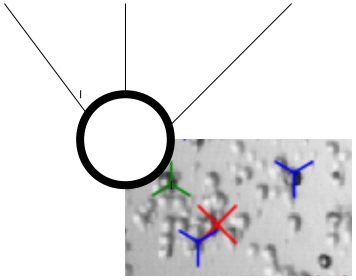
Raw Data

Image processing
Normalisation

Pre-Selection

Labeling
Manual / Inference

Trained sub-classifiers



Ensemble / output

- Collect higher Res data on test sample
- ‘Dry run’ full analysis on test sample foil
- Build representative set of through going HIPs, and common bkgs, detailed imagine both sides, different focal planes, high zoom. Send to Laura+INF

# End

A.O.B?
Suggestions / requests

IOP
High Energy Physics + Astro-Particle Physics
Conference early april
(Any requests / comments)

# Current work;

- Refining pipeline
- This to fix norm bug tf-ds, images, npy
- Also allows to future prep for new data
- Data dict, meta-data track dataflow at different levels foil – image – pit easy to add / remove data # prevent data explosion when doing analysis

- Backgrounds
- Think about different backgrounds
- ML will behave differently in different areas of foil
- Low density want to compress information
- Med density – want to use previous slide techniques
- High density Human inspection New techniques
- Het data ( eg, pit list )

# Next steps

- Push to higher bkg;
- Gradual retraining / transfer learning / boosting w inference assistance
- Return to image processing – tricks
- Post processing / clean up on very dirty images
- Weak learning / new techniques
- Tweak ensemble
- - looks ‘god enough’ atm
- Data augmentation

  (brief slides)
- More 3d ML, (was holding off due to more initial params)
- IOP – talk any comments / suggestions / requests