

DATA CHALLENGE INTRODUCTION

Giles Strong

INTRODUCTION TO THE TASK

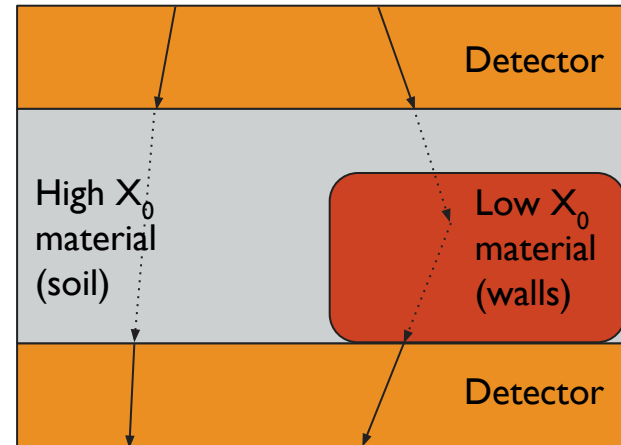
- Britain's oldest city, Colchester, was founded by the Romans around 40AD as a barracks on the site of a Celtic stronghold
- Throughout the 2nd & 3rd centuries, the city expanded, eventually becoming a *colonia* -- an extension of the city of Rome
- Roman landmarks remain to this day, with more discoveries still being made
- Whilst performing some construction, workers uncovered indications of ruined walls in a site that was previously thought to be empty
- You have been brought in to scan the site using *muon tomography* and to map out the locations of the walls to help aid the archeologists
 - They have provided you with a set of simulated data on which to develop a suitable inference algorithm



Example of Roman walls in Colchester,
credits [Maria](#), CC-BY-SA 3.0

TOMOGRAPHY VIA MULTIPLE SCATTERING

- Muon tomography allows us to infer the material composition of unknown volumes of space
- Cosmic muons are scattered by materials in the volume according to their radiation-length (X_0 [m]) of the material
 - Radiation-length = average distance between scatterings
- By using detectors, we can measure muons above and below volume
 - The changes in trajectory provide information on material composition



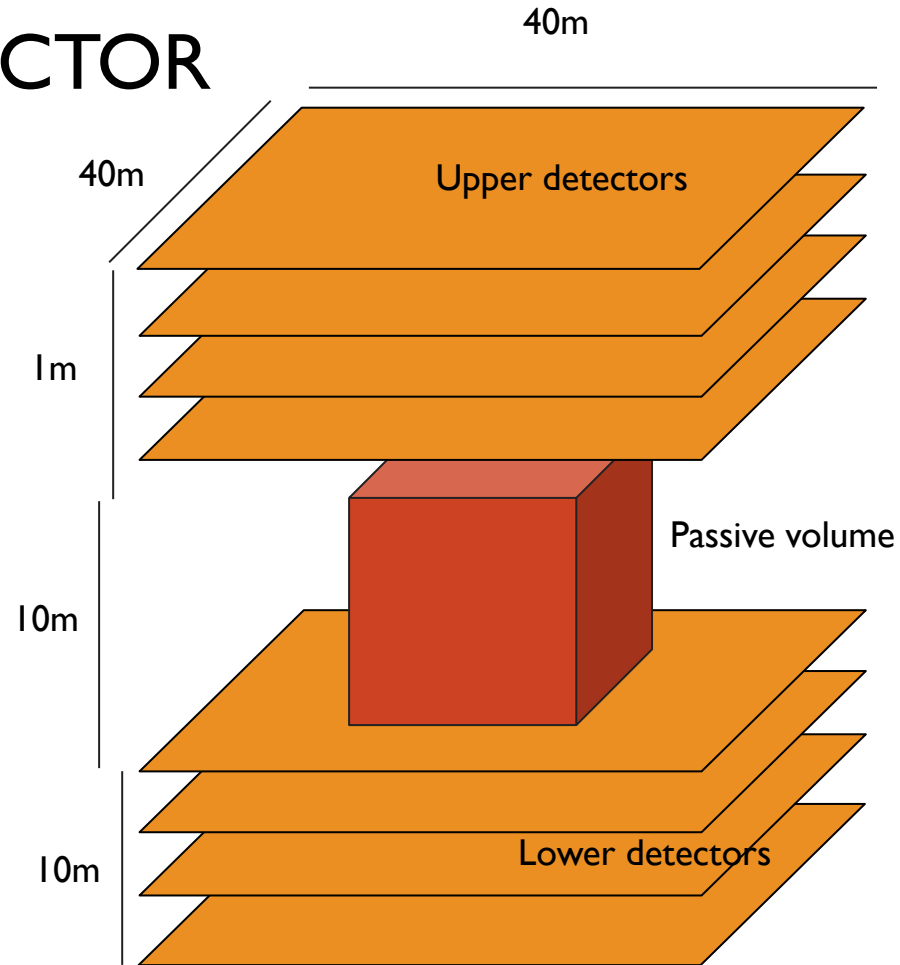
High X_0 = low scattering

Low X_0 = high scattering

X_0 = average distance between scatterings

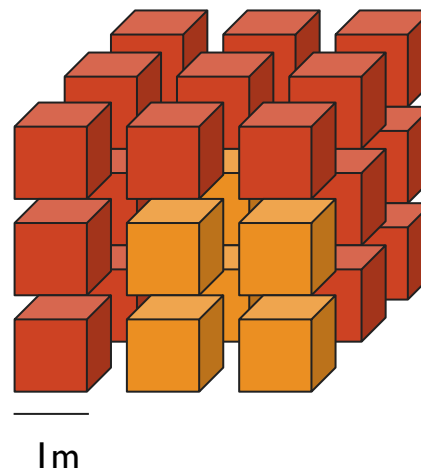
DETECTOR

- The detector setup used here is quite unrealistic
 - It assumes that a detector panel may be placed underground, directly under the passive volume
 - The simulated detector is very large (40x40m), but this is for simulation convenience; a smaller detector could be used and placed in several spots to create a combined scan
- The detector consists of two layers (1m height), placed above and below the volume
 - Each layer contains 4 equally spaced panels
 - Each panel records muon positions with an xy resolution of 0.1mm

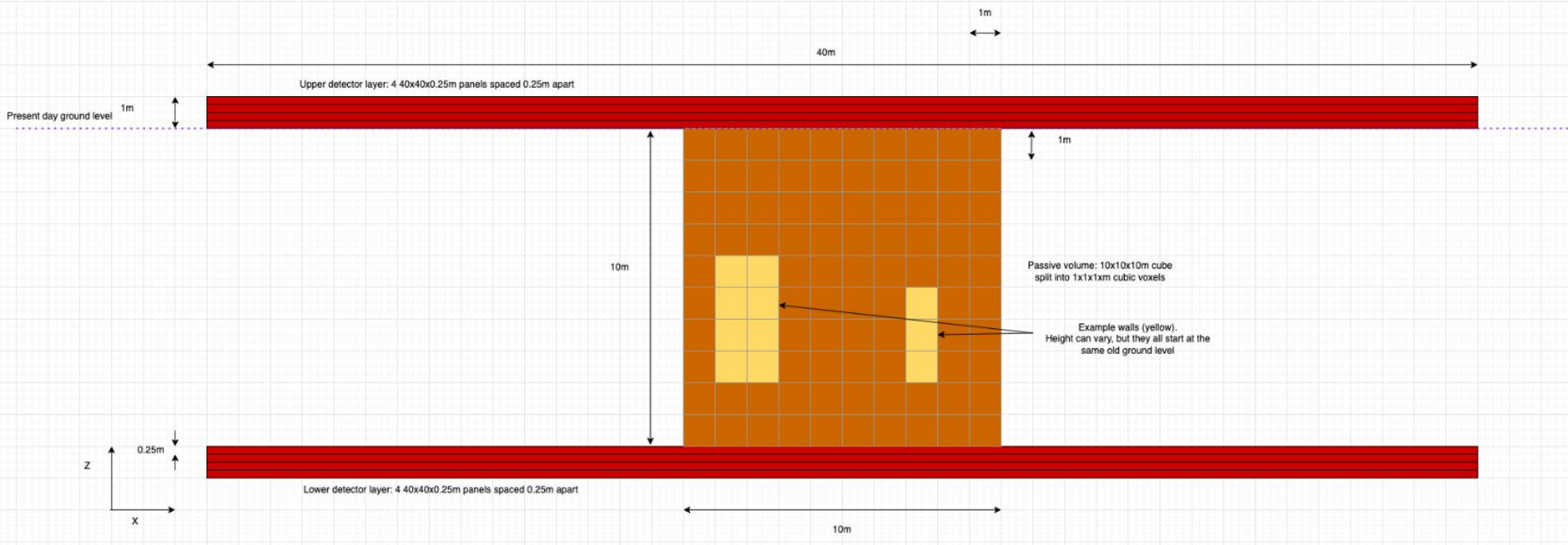


PASSIVE VOLUME

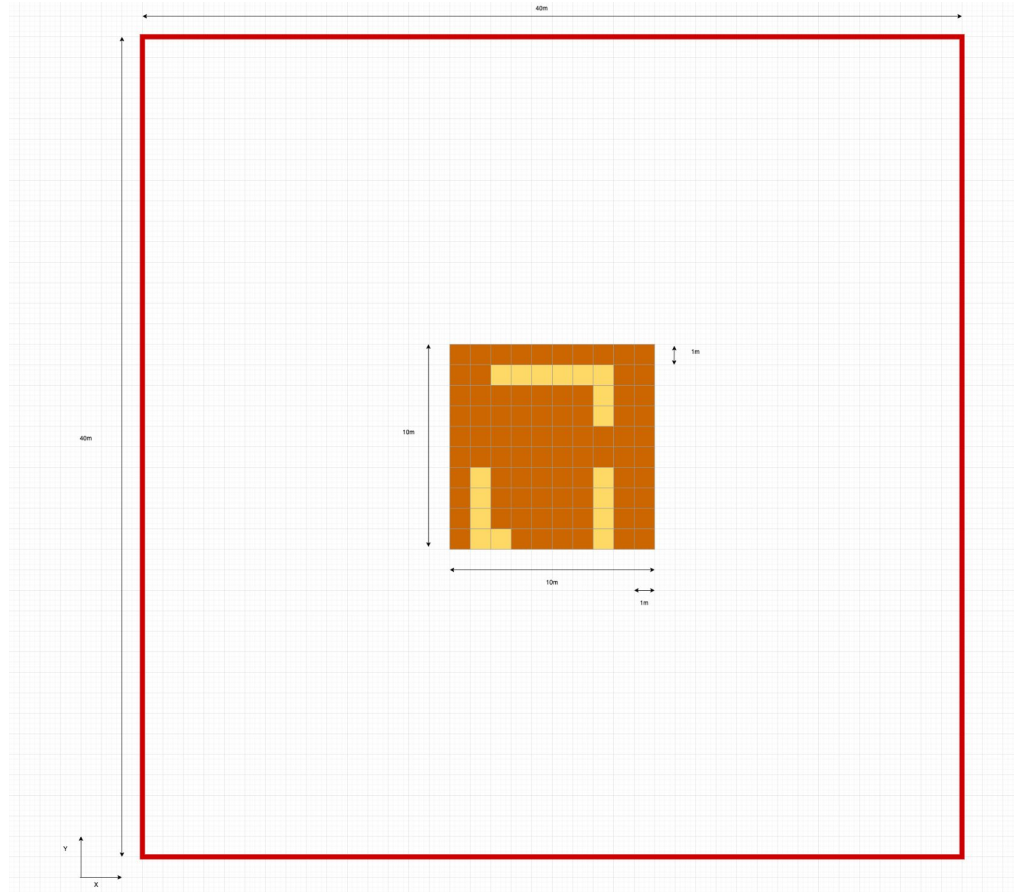
- The passive volume is a $10 \times 10 \times 10 \text{ m}$ cube
 - It is subdivided into 1000 voxels, each $1 \times 1 \times 1 \text{ m}$ in size
- Each voxel can either be soil ($X_0 \sim 0.26 \text{ m}$) or wall ($X_0 \sim 0.08 \text{ m}$)
 - The amount of muon scattering depends on the voxel X_0 , and scales as $\sqrt{\text{distance}/X_0}/\text{momentum}$
 - Muon momentum will always be 1 GeV , but the distance depends on the incoming angle



COMPLETE VOLUME SETUP: ZX

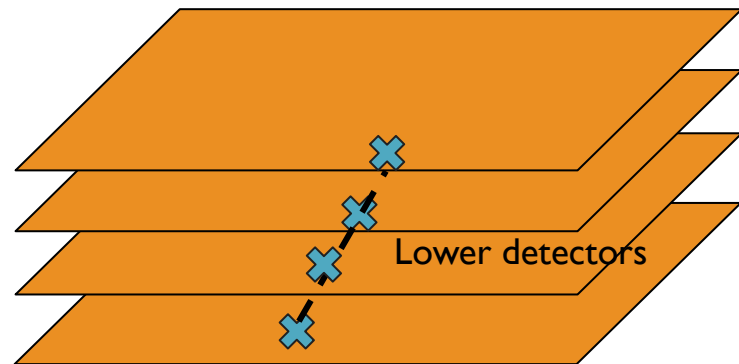
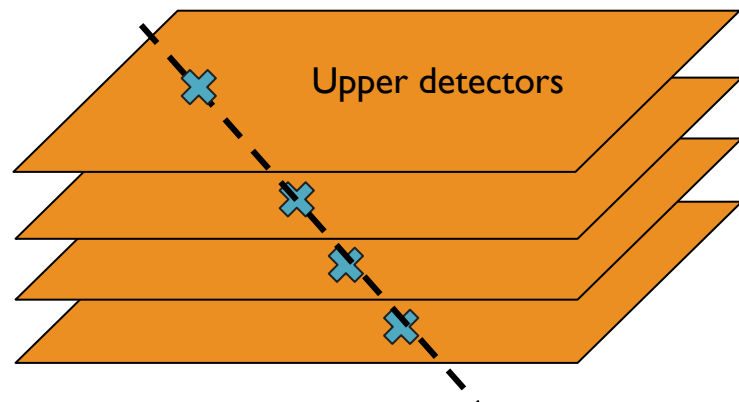


COMPLETE VOLUME SETUP: XY



TRACK FITTING

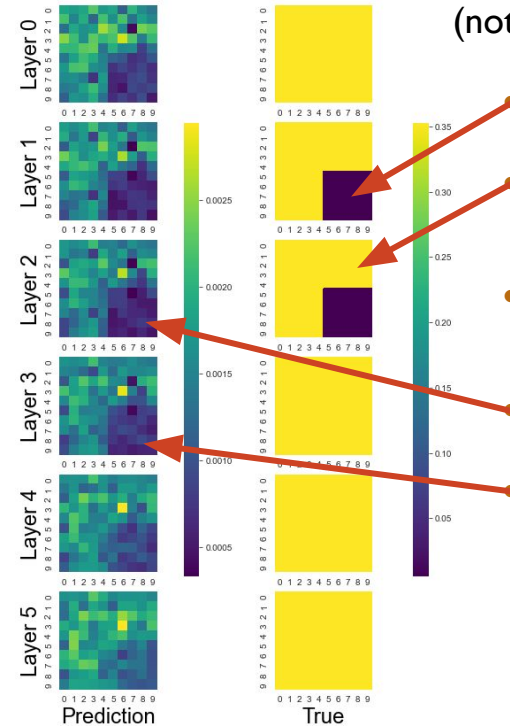
- Fit linear trajectories to the hits
- Can then compute variables about the muon scattering, e.g.:
 - Incoming and outgoing angles
 - Changes in trajectory



PoCA INFERENCE

- PoCA method assigns the entirety of the muon scattering to a single voxel
 - The voxel is chosen by extrapolating trajectories inside the passive volume to find the Point of Closest Approach
 - X_0 predicted by inverting the scattering model to get X_0 as a function of total scattering, and then averaging over many muons
 - We made a slight modification: X_0 predictions are applied to every voxel, but in an average weighted by the probability of the scattering having occurred there
 - Computed using the uncertainty on the PoCA location
 - This provides a dense set of voxelwise X_0 predictions
- PoCA is a very standard algorithm in tomography, but produces biased and blurred results

PoCA Example
(not related to our dataset)



Block of lead
($X_0=0.005612\text{m}$)

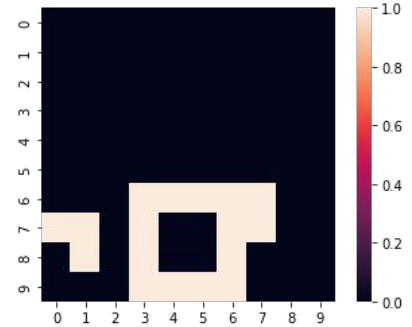
Surrounded by
beryllium
($X_0=0.3528\text{m}$)

- Predictions highly
biased to
underestimate X_0
Lead block clearly
visible

but high z uncertainty
in scatter location
causes 'ghosting'
above and below

TARGET

- The passive volumes are randomly generated to simulate stone walls buried underground
 - Each volume contains at least one wall, surrounded by soil
- All walls in a given sample begin on the same z position, but can vary in height
 - The starting z position of the walls can vary between samples

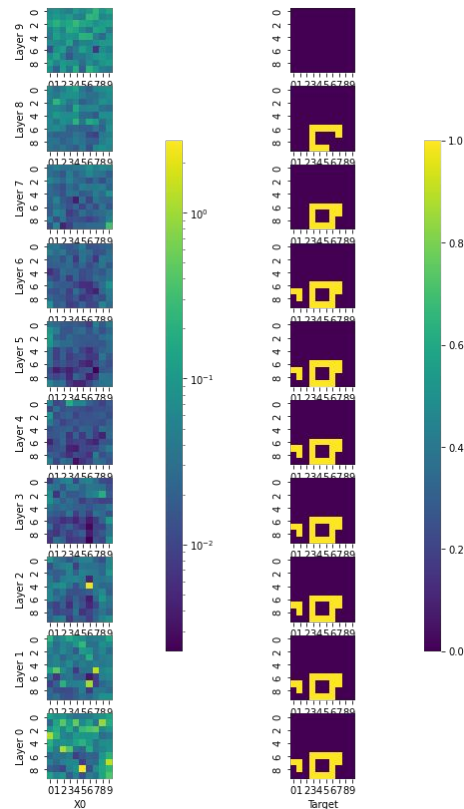


Top-down image

XY slices in layers of z

DATASET

- A sample is created by scanning a newly generated passive volume:
 - 10,000 muons are recorded
 - Incoming angle and initial xy position can vary
 - Momentum is fixed at 1 GeV
 - The scanning process uses the PoCA inference method described on slide 9
 - This results in a biased PoCA image of voxelwise X0 predictions (float32)
 - The target is a map of the wall voxels (int, 0 = soil, 1 = wall)
- Approximately 100k labelled samples are provided along with 30k unlabelled testing samples
 - Your task is to provide predictions on the unlabelled sample



Example sample:

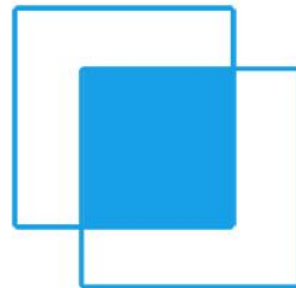
Left = voxelwise
X0 predictions

Right = map of
voxels which
contain wall

METRIC

- Both predictions and targets will be a voxelwise 0 or 1
- Performance evaluated using the “intersection-over-union” (IOU) metric
 - Intersection is number of wall voxels correctly predicted to be wall
 - Union is the sum of the number of voxels that are either predicted to be wall or are actually wall
- Behaviour:
 - Predicting all walls gets high intersection but also high union = low IOU
 - Predicting no walls gets zero intersection = zero IOU
 - Accurate wall predictions get high intersection and low union = high IOU
- IOU is between 0 and 1, and higher values mean better performance

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

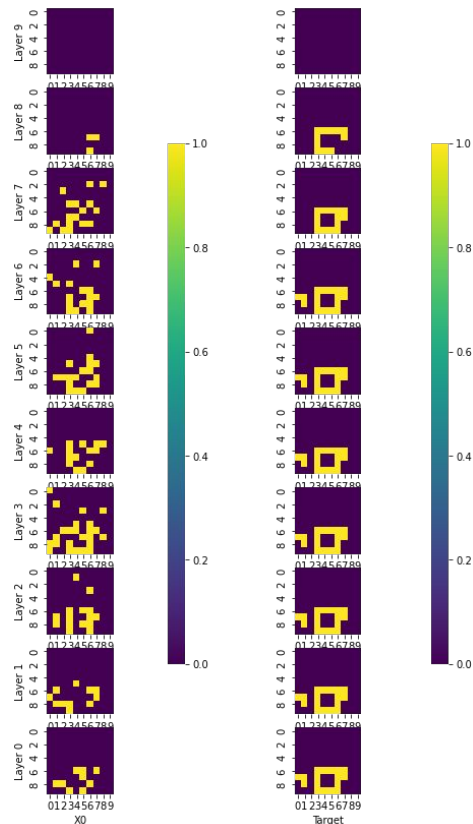


STARTER PACK

- Starter notebook covers:
 - Loading the data and viewing the data
 - Applying a simple threshold-based approach
 - Evaluating performance using the intersection-over-union metric
 - Creating a submission with predictions for the test data
- Threshold method: prediction based solely on X_0 values:
 - X_0 below threshold, predict wall
 - X_0 above threshold, predict soil
 - Optimise thresholds by maximising the IOU

Left:
predictions
using the
optimised
threshold

Resulting
IOU is 0.63



SUBMISSIONS

- Submissions must be uploaded in HDF5 format to <https://cernbox.cern.ch/index.php/s/ylsOYg9q7hcRk4l>
 - Format must be a (sample, z, x, y) matrix of integer values stored in a dataset called 'preds'
 - Submissions not in this format will be ignored
 - Order of predictions must be the same as the samples in the testing dataset
 - The starter pack contains instructions on how to do this
 - File names should include your full name and an optional ID number for the submission
 - In case of multiple submissions, I will use the one with the highest ID number
 - Files without a person's name will be ignored
- Deadline is 23:59:59 CEST on 22/09/04
- Submissions can be made anytime before the deadline, and multiple submissions are ok
 - Final results will be based on the latest submission
- Teams are allowed, but we can only offer one prize to each team
- Every Friday I will announce results of performance on a fixed, random subsample (5000 samples) of the test data using everyone's latest submissions
 - I will also let anyone know if their submission format was invalid
- The final results will be computed using the remaining samples (25,000) from the test dataset