# Workshop No.1 — Kaggle Systems Engineering Analysis Report

Juan David Buitrago Rodriguez - 20242020194
David Giovanni Aza Carvajal - 20241020137

April 2025

## Competition Description

For this workshop, we have decided to analyze the competition *HuBMAP - Hacking the Human Vasculature*.

For this competition, we are assigned with the next objective:

*"The goal of this competition is to segment instances of microvascular structures, including capillaries, arterioles, and venules. You'll create a model trained on 2D PAS-stained histology images from healthy human kidney tissue slides. Your help in automating the segmentation of microvasculature structures will improve researchers' understanding of how the blood vessels are arranged in human tissues."* [1]

## Dataset structure

In order to complete the objective that we have been given, we are provided with the next Dataset structure:

- {`train|test`} - Folders containing TIFF images of the tiles. Each tile is 512x512 in size. The images are histological sections of the human kidney, obtained through digital microscopy. These images come from WSI (Whole Slide Images), which are full scans of renal tissues.
- `polygons.jsonl` - Polygonal segmentation masks in JSONL format, available for Dataset 1 and Dataset 2. Each line gives JSON annotations for a single image with:
  - `id` - Identifies the corresponding image in `train/`
  - `annotations` - A list of mask annotations with:
  - `type` - Identifies the type of structure annotated:
    - `blood_vessel` - The target structure. Your goal in this competition is to predict these kinds of masks on the test set.
    - `glomerulus` - A capillary ball structure in the kidney. These parts of the images were excluded from blood vessel annotation.
    - `unsure` - A structure the expert annotators cannot confidently distinguish as a blood vessel.
  - `coordinates` - A list of polygon coordinates defining the segmentation mask.
- `tile_meta.csv` - Metadata for each image. The hidden version of this file also contains metadata for the test set tiles.

- `Donor-Level Metadata and Potential Hidden Variables` - The dataset includes donor IDs, organ types, and other metadata, which may correlate with staining styles, acquisition settings, or tissue health.
- `Variable Organ Types and Sampling Conditions` - Images come from different organs (e.g., kidney, liver, spleen) and tissue regions, introducing biological diversity and inconsistent anatomical landmarks.
  - `source_wsi` - Identifies the WSI this tile was extracted from.
  - `Tiling Strategy: Managing Gigapixel Images` - Given the immense size of WSIs, the dataset employs a tiling strategy, breaking down each slide into smaller, manageable tiles.
  - {`i|j`} - The location of the upper-left corner within the WSI where the tile was extracted.
  - `dataset` - The dataset this tile belongs to, as described above.
- `wsi_meta.csv` - Metadata for the Whole Slide Images the tiles were extracted from.
  - `source_wsi` - Identifies the WSI.
  - `age, sex, race, height, weight,` and `bmi` - demographic information about the tissue donor.
- `sample_submission.csv` - A sample submission file in the correct format.

## Relationships among elements

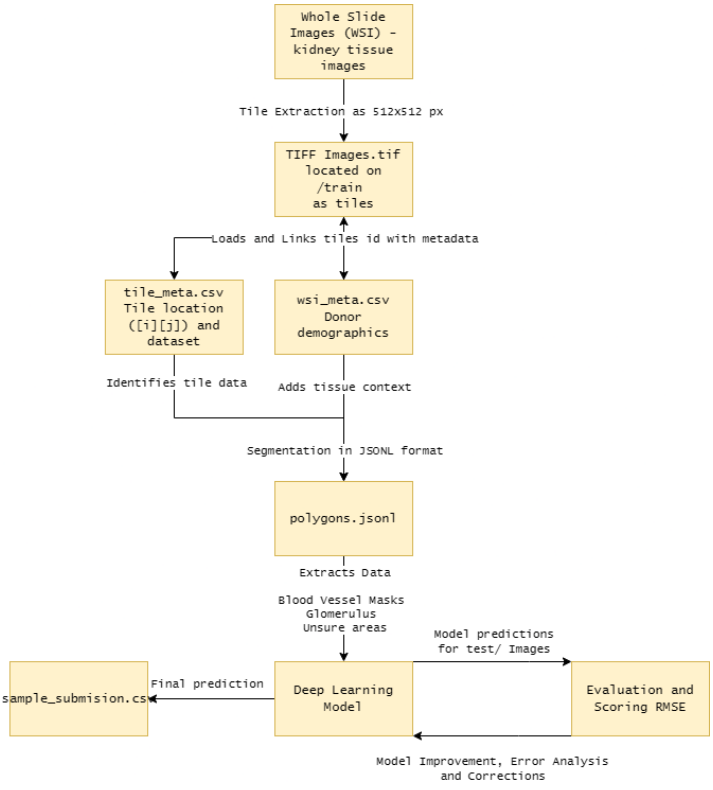


Figure 1: Interactions among elements

This data flow represents the problem presented by the competition in a more visual way. Whole Slide Images (WSI) are divided into 512x512 pixel tiles and stored as TIFF images in the /train directory. Metadata files (`tile_meta.csv` and `wsi_meta.csv`) link each tile to its position within the original WSI and provide donor demographic information. Segmentation masks in `polygons.jsonl` mark the coordinates of blood vessels, which are our main target, glomerulus regions (meant to be excluded), and areas that are classified as uncertain. A deep learning model is trained using these annotations to generate predictions for test images. The model's performance is evaluated using RMSE, followed by improvements based on error analysis. Finally, the refined predictions are formatted and stored in sample_submission.csv for submission.

## Complexity & Sensitivity

The HuBMAP competition presents a complex data ecosystem in which multiple variables interact in potentially nonlinear ways. Several factors introduce sensitivity into the system:

- **Image Variability:** Even though the tiles extracted maintain the same dimensions (512x512) the variation on the stain colors, tissue density and the glomerulus size. Are some of the variations to consider that might make it harder for the model to analyze.
- **Competing Optimization Objectives:** Participants must balance several trade-offs: precision vs. recall, generalization vs. specialization, runtime vs. accuracy..
- **Metric-Induced Fragility:** Metrics like IoU and Dice penalize even minor misalignments between predicted and ground truth masks. This sensitivity demands surgical-level precision, especially in segmenting thin and branching vessels that occupy very little area but are crucial to the outcome.
- **Annotation Ambiguity:** Analyzing the data structure on `polygons.jsonl` the presence of the `"unsure"` data structure (A structure the expert annotators cannot confidently distinguish as a blood vessel) leaves a space for ambiguity. the more `unsure` data the more possibles miss interpretations we might encounter on our prediction.
- **Demographic Diversity:** Scoping on `wsi_meta.csv` file we can find variations on age, sex and BMI from different tissue donors, these differences are linked to the variability on the blood vessel structure me might encounter.
- **Mask Complexity:** The variation on complexity on each tile may differ, due to the overlapping of blood vessels, shape and density variations represent an extra layer of complexity.
- **Feedback Loops in Model Tuning:** The constant use of the same data in all the training stage may influence the model to perform poorly when tested on new data.

## Chaos and Randomness

Some elements of chaos theory are observable or potentially emergent in this system:

- **Initial Condition Sensitivity:** Under initial learning stages, any small changes on the model settings can lead to very different results.
- **Data Augmentation Effects:** Feeding the model with several types of WSI's differing by color, resolution, more detailed blood vessels, rotations and others may help the model generalize and process different inputs but may also lead to unpredictable outputs.
- **Random Tile Sampling:** The random selection of image tiles during training can expose the model to different structures in each run, leading to small variations in performance between training sessions.

- **Inter-Class Visual Overlap:** Non-vascular structures such as tubules, lymphatic channels, or tears can resemble vessels visually. This leads to semantic confusion.
- **Artifacts Introduced During Image Acquisition:** Technical noise (including scanner calibration errors, blur, pixelation, or overexposure) adds unpredictable disturbances. These errors often distort the very vascular structures that must be segmented.
- **Tile Variability:** Tiles from the same WSI may have variations going from stain color, consistency, unsure annotations or even scanner errors, this breach of non-static inputs can lead to a unpredictable curve of learning.
- **Temporal and Environmental Variability:** Tissue samples may be collected over long periods, from different hospitals, under varying protocols. This adds temporal noise and uncontrolled variables.

## Conclusion

Along the analysis of the HuBMAP competition, we have encountered various interconnected key elements. Going through each element and proactively understanding how they interact we get a better understanding on the systems data flow. Thanks to the rich dataset on `polygons.jsonl` and detailed metadata represent a great resource for deep learning, but also a challenge due to the variability on the `.tif` files. Regardless of the challenge the variations on the `.tif` might represent the extensiveness of the annotations on `polygons.jsonl` and the inclusion of demographic data enhances the possibility of more robust results.

However, some visible weaknesses can be highlighted. The presence of the `"unsure"` data structure leaves a space for ambiguity, and complexity on each mask, like the overlapping of blood vessels, are behaviors that might result in a degree of randomness. These factors may lead to different strategies with the objective of minimizing these behaviors.
Overall, this analysis highlights the need for robust data pre-processing, careful model validation, and strategies that embrace rather than eliminate any possible variability or randomness.

## References

[1] Kaggle. (2025). *HuBMAP - Hacking the Human Vasculature Competition*. Retrieved from `https://www.kaggle.com/competitions/hubmap-hacking-the-human-vasculature`