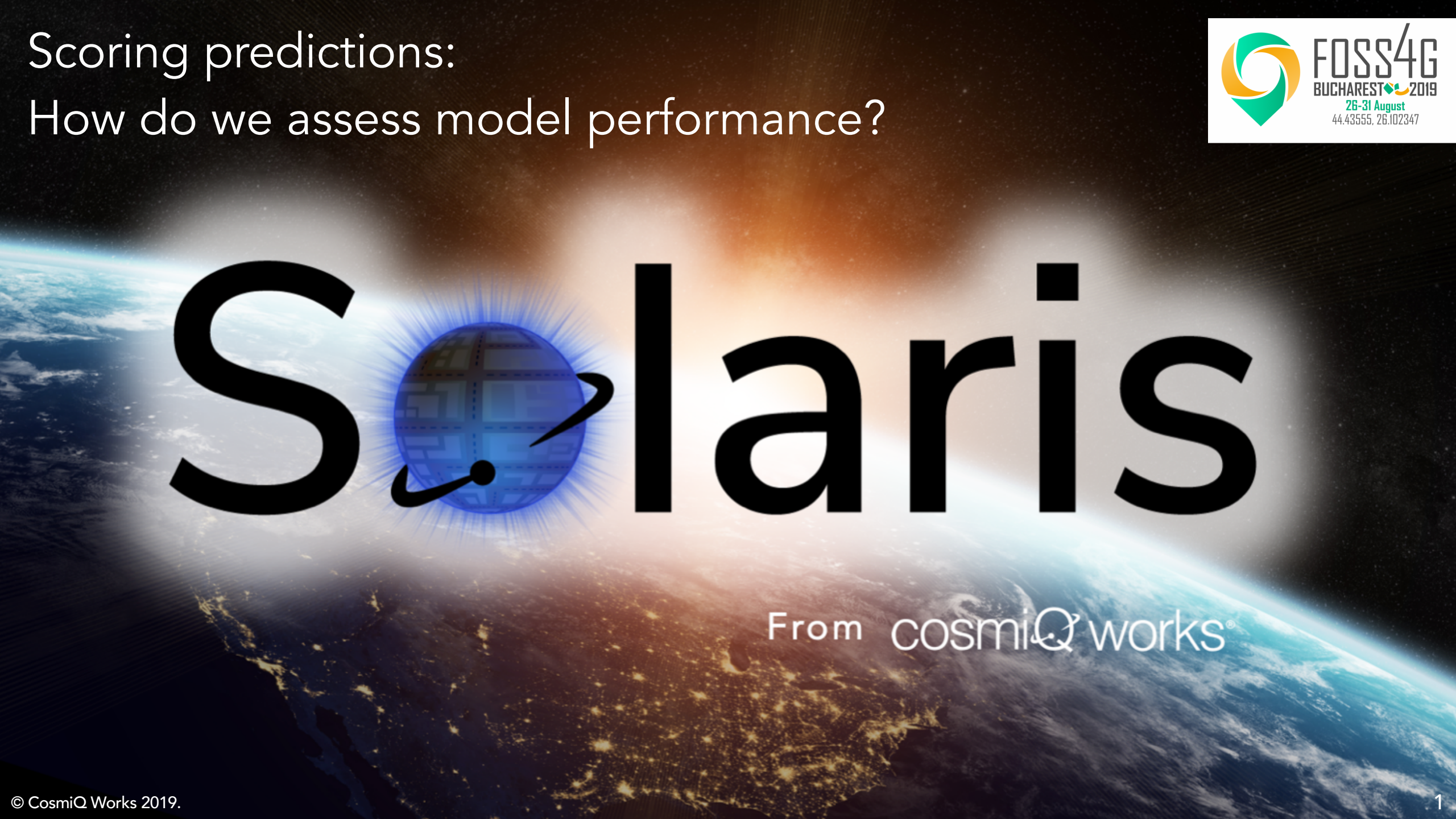
Scoring predictions:
How do we assess model performance?

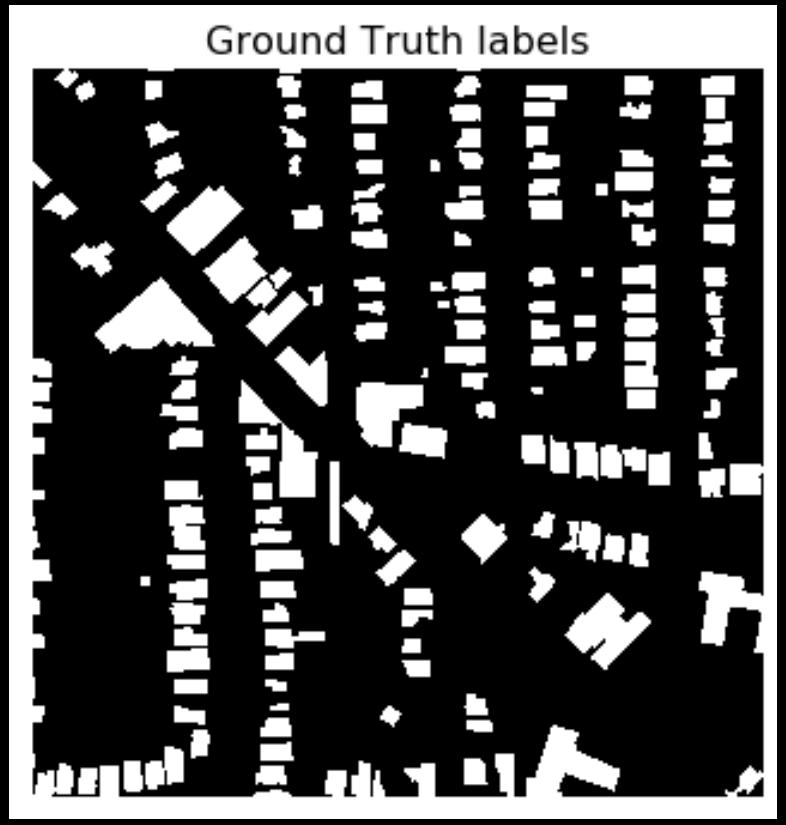


FOSS4G BUCHAREST 2019
26-31 August
44.43555, 26.102347

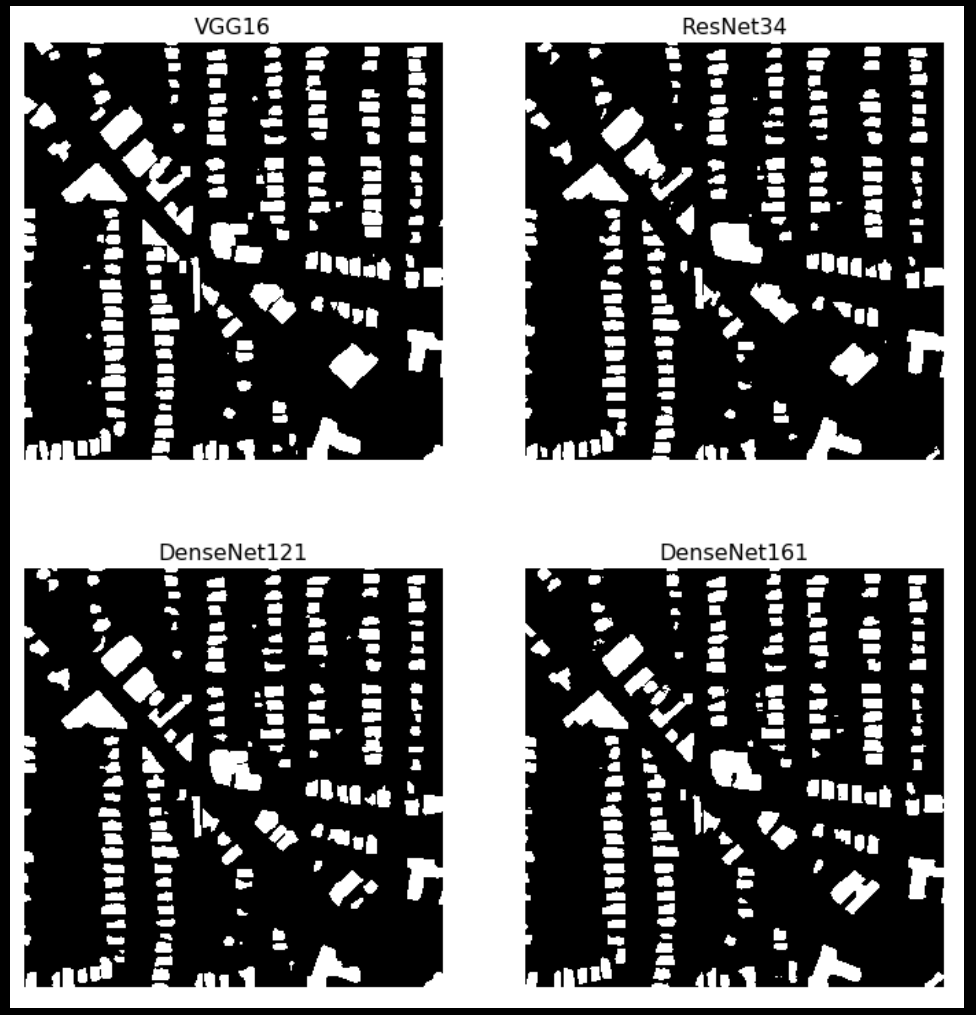
Before we get started…

Start 3_evaluating_performance.ipynb and
run the first two cells

How do we compare output quality?



Vs.



How do we compare output quality?

Pixel-wise metrics:

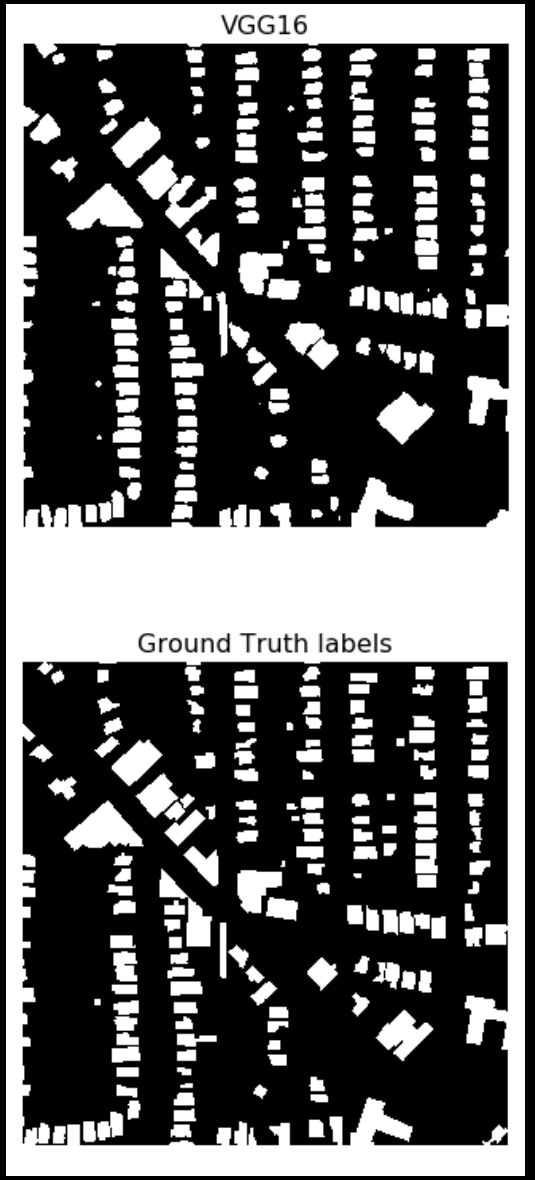
$$Accuracy = \frac{Correct\ pixels}{Total\ pixels} \qquad Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \times \frac{P \times R}{P + R}$$

(there are others)

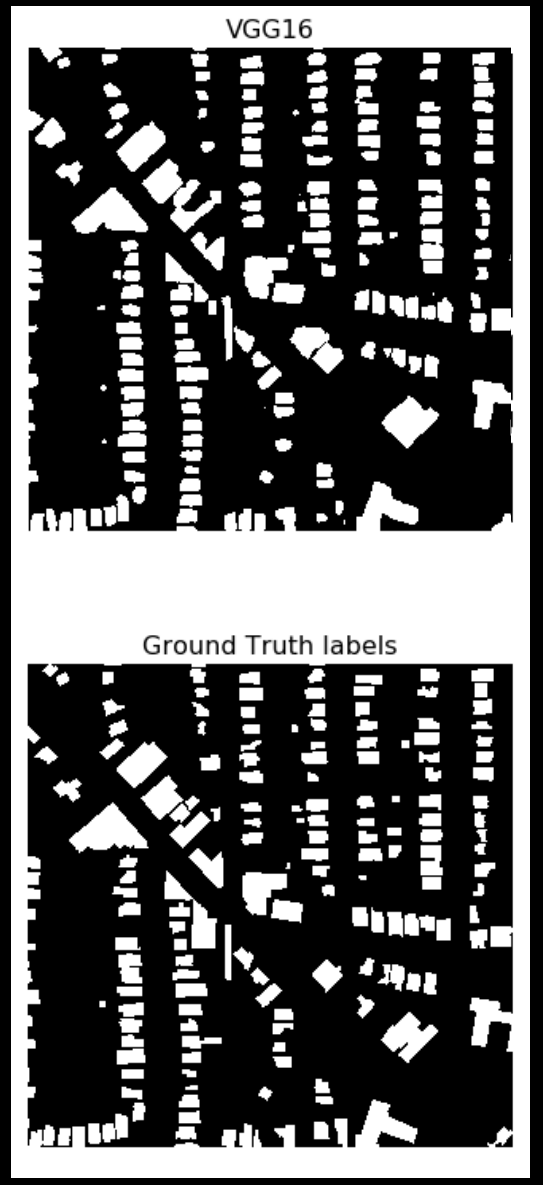
What questions can we answer with counting pixels?

- What fraction of pixels did we get right?
- How many non-building pixels did we ID as part of buildings?
- What portion of the image is composed of buildings?



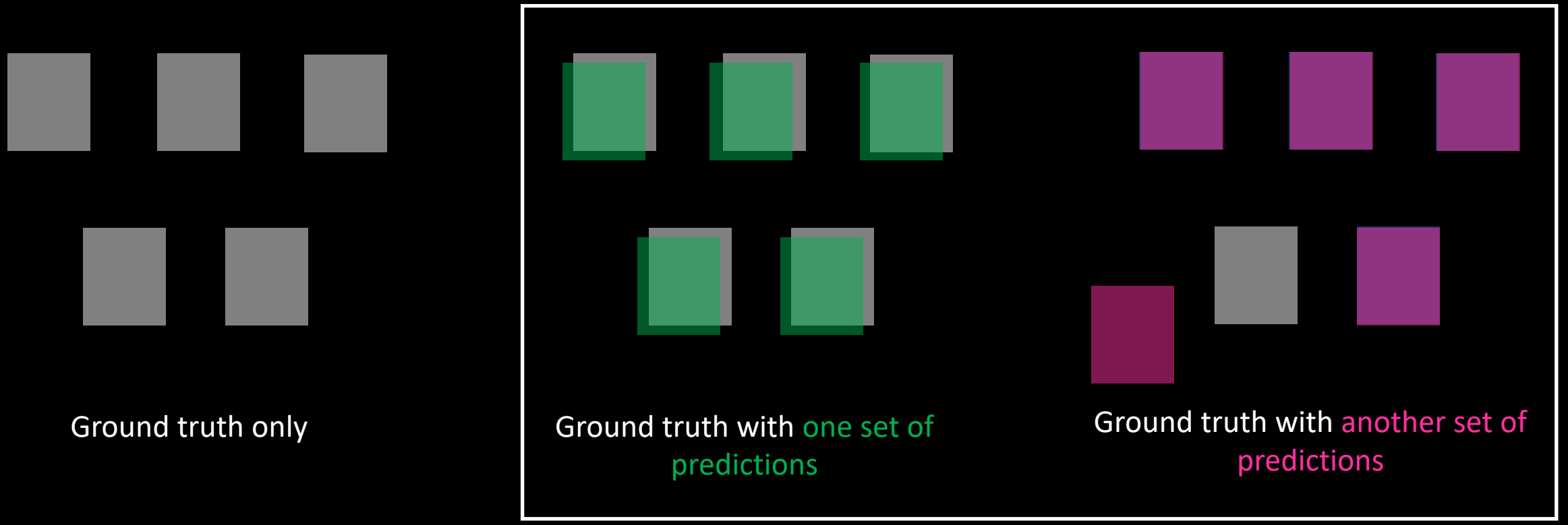
What *can't* we say from a pixel-based metric?

- How many buildings are there?
- What fraction of buildings did our model find?
- How many buildings did we find fewer than half of the pixels for?
- How many extra buildings did we predict?
- Did we ID smaller or bigger buildings better?



Pixel-wise metrics can introduce confusion

The Green and Pink predictions have the **same** pixel-wise F_1 , Precision, and Recall scores.



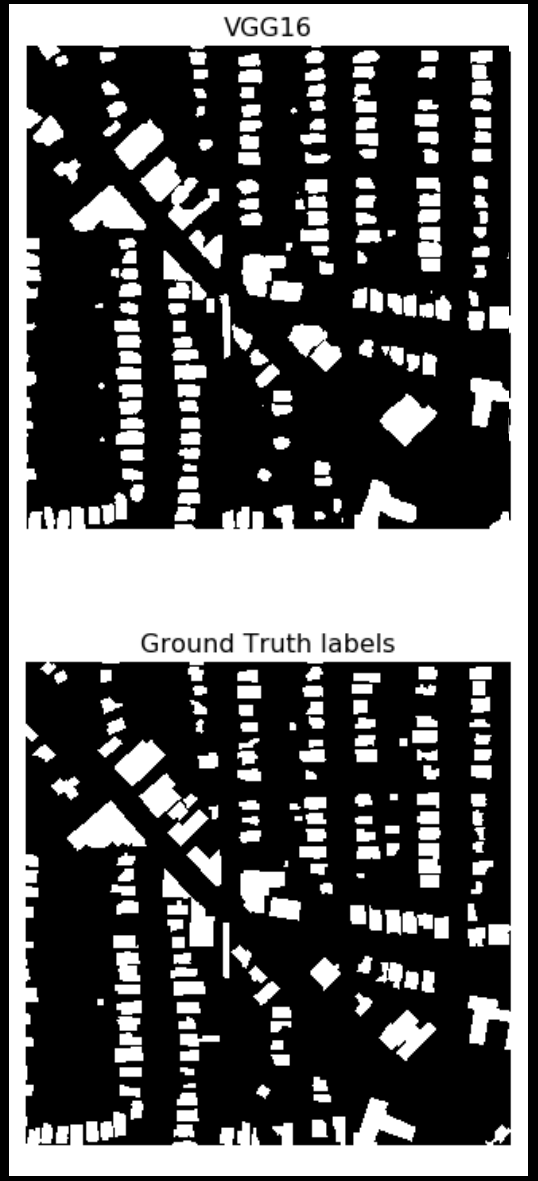
Ground truth only

Ground truth with one set of predictions

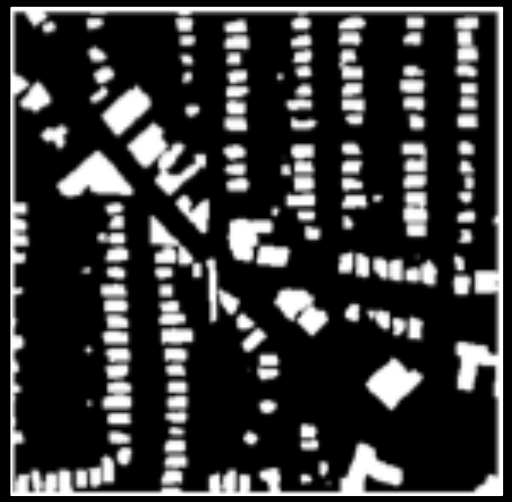
Ground truth with another set of predictions

Measuring object identification directly

- Rather than measuring *pixel-wise* success, measure *object-wise* success
- Two requirements:
  - A method to determine if you got a given object correct
  - A summary metric to aggregate across objects



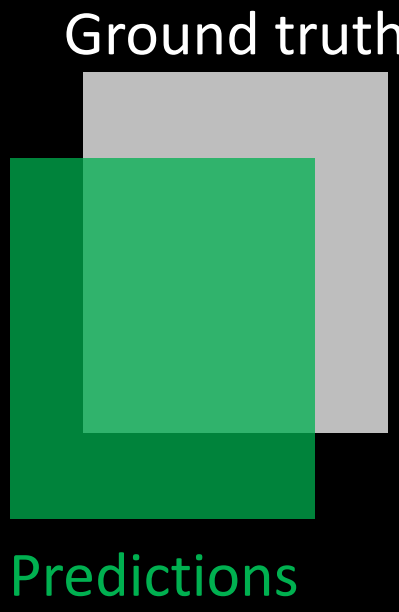
Going from pixel masks to object polygons



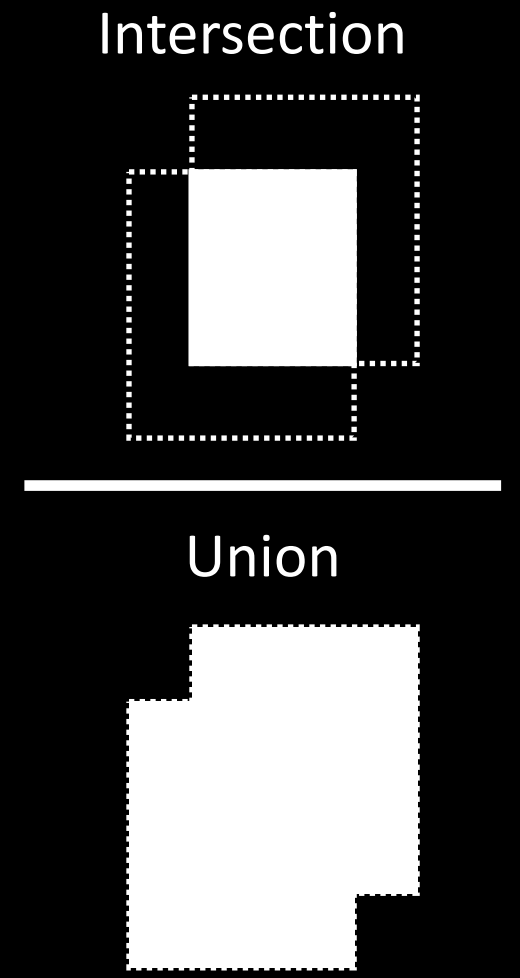
POLYGON ((745804.5 3726489, 745812.5 3726489, ...
POLYGON ((746007 3726470, 746008.5 3726470, ...
POLYGON ((745825.5 3726436.5, 745826.5 3726436, ...
POLYGON ((745940.5 3726377, 745941 3726377, ...
POLYGON ((746192.5 3726256, 746201 3726256, ...
POLYGON ((745829.5 3726234, 745842 3726234, ...
POLYGON ((746078 3726202.5, 746080.5 3726202.5, ...
POLYGON ((745865 3726103, 745868 3726103, ...
POLYGON ((745880.5 3726096.5, 745883.5 3726096, ...
POLYGON ((745763 3726056.5, 745764.5 3726056.5...

1. Merge contiguous labels
2. Discard inappropriate-sized objects
3. Georegister
4. Produce outputs in standardized format

Measuring object-wise success: Intersection-over-Union (IoU)



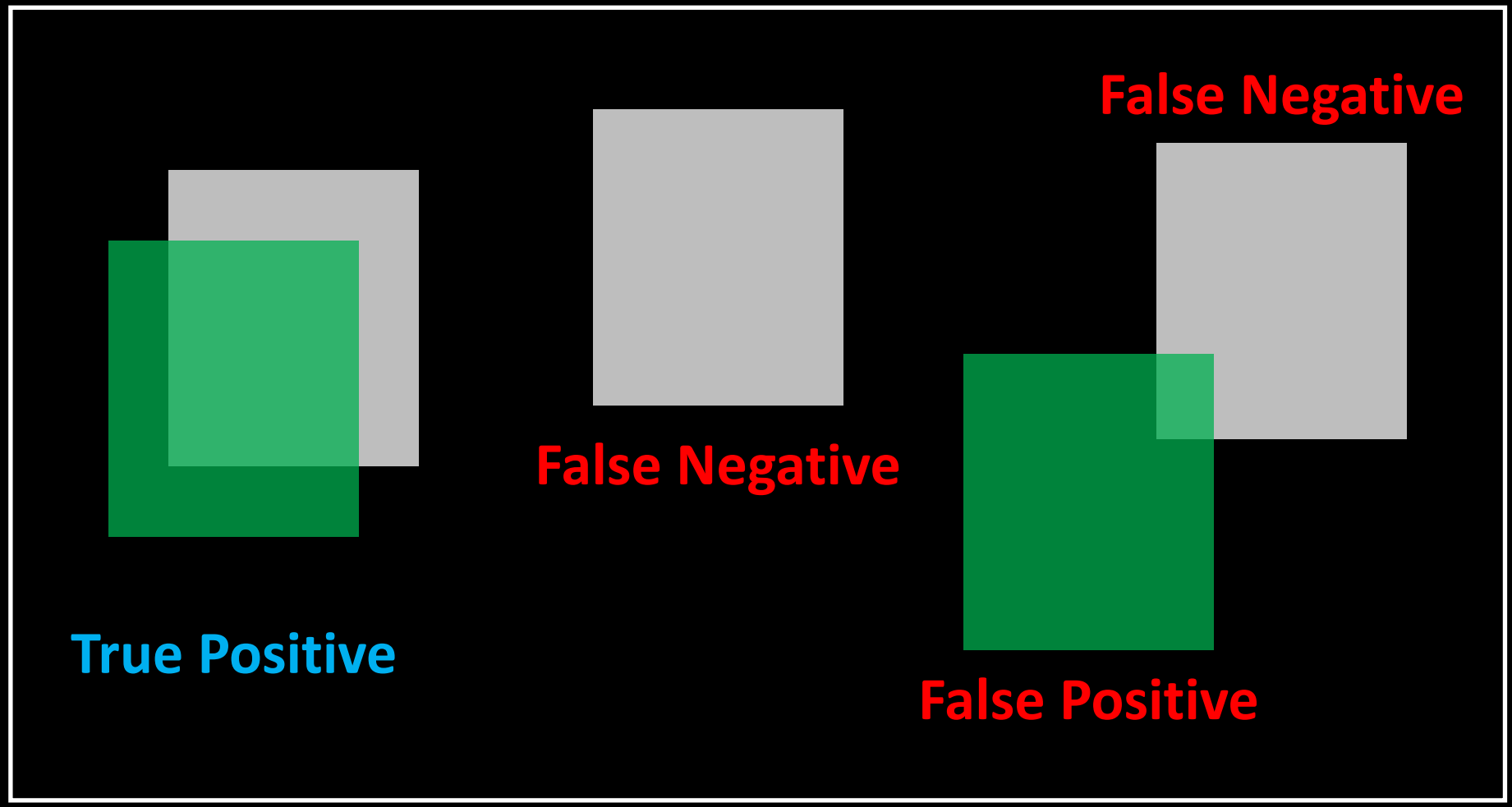
$$IoU = \frac{Area(intersection)}{Area(union)} =$$



If IoU > your threshold, this counts as a positive ID (True Positive).

Interactive example: Determining TP, FP, and FN

If the IoU threshold for a true positive is 0.5…

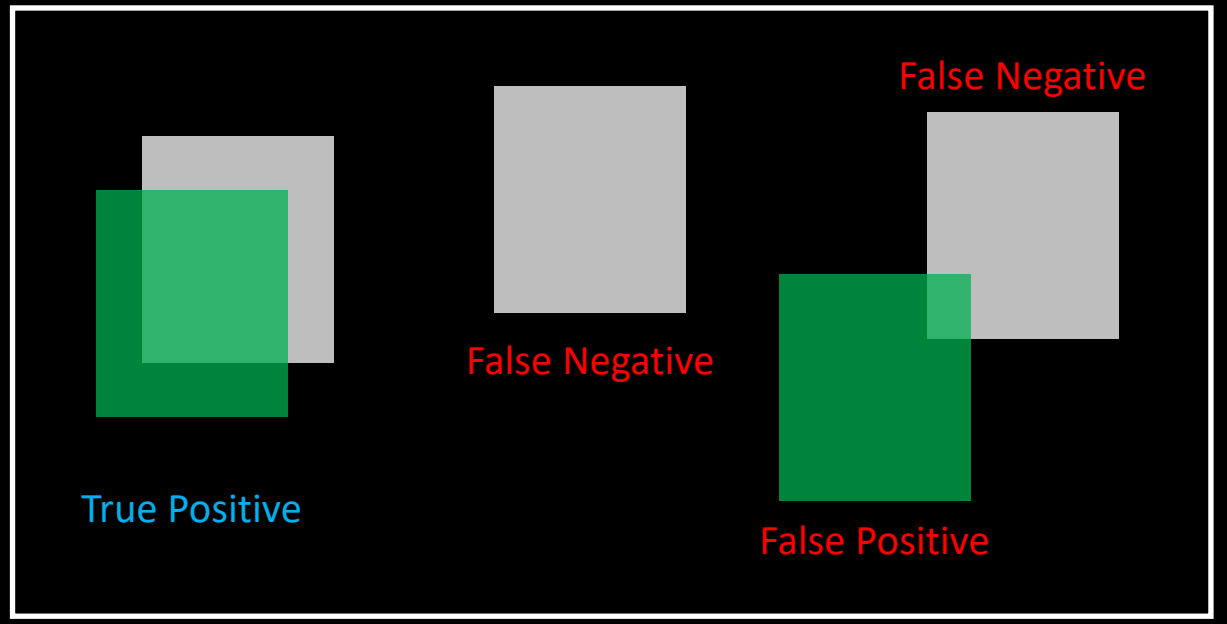


Calculating precision, recall, and F_1

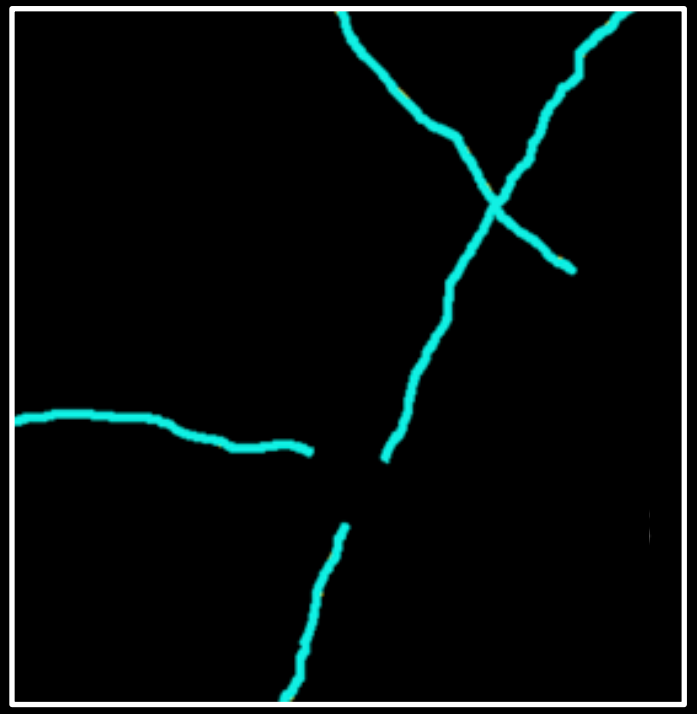
$$Precision = \frac{TP}{TP + FP} = \frac{1}{1 + 1} = 0.5$$

$$Recall = \frac{TP}{TP + FN} = \frac{1}{1 + 2} = 0.33$$

$$F_1 = 2 \times \frac{P \times R}{P + R} = 2 \times \frac{0.5 \times 0.33}{0.5 + 0.33} \approx 0.397$$

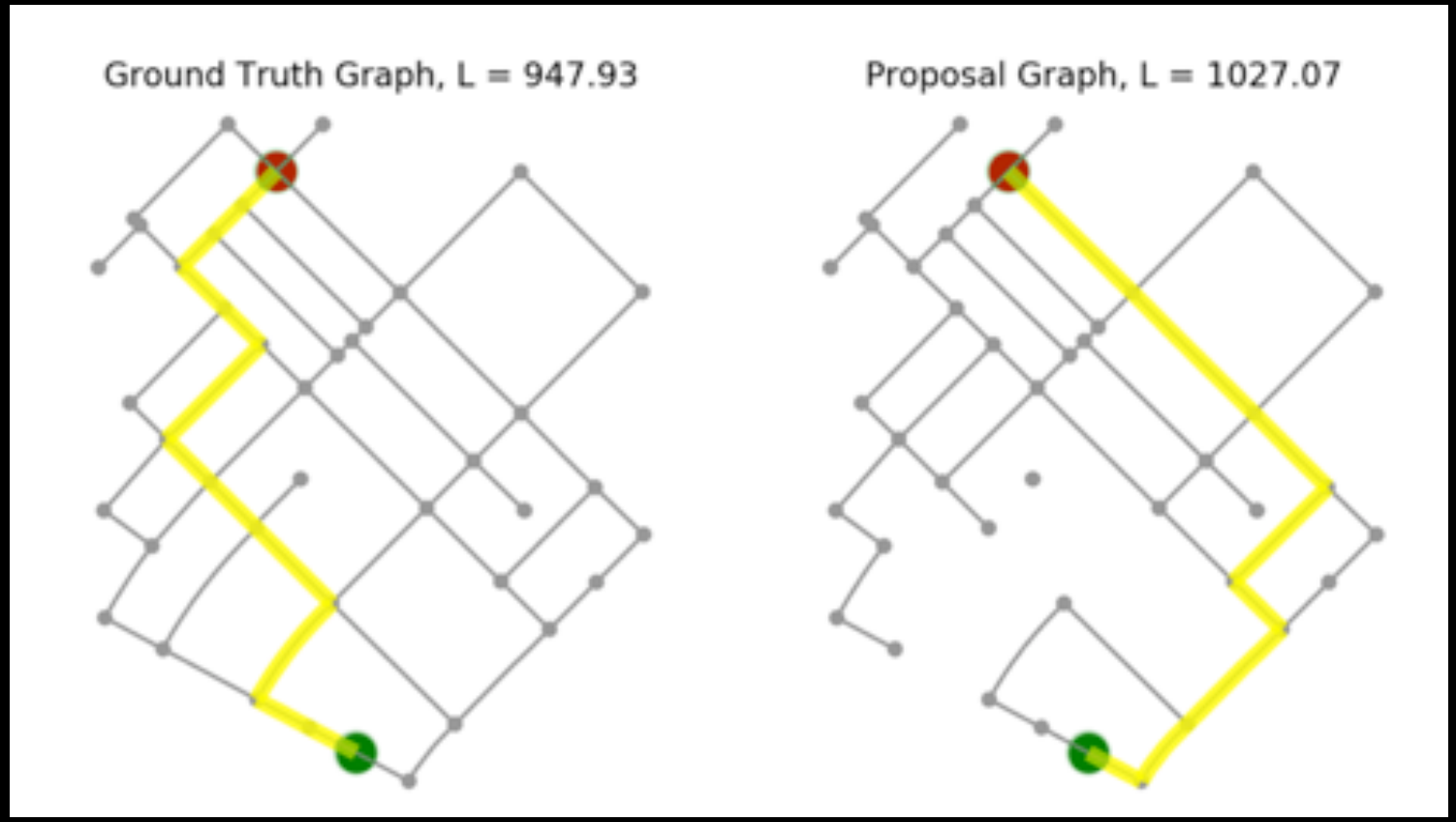


What if you wanted to score road network predictions?



Scoring road network quality: Average Path Length Similarity (APLS)

- Average Path Length Similarity (APLS) was developed by Adam Van Etten (CosmiQ Works) for SpaceNet Challenge 3
- Both Logical and Physical Topology Are Important for Road Detection
- Sum the Difference in Paths between Ground Truth & Proposals
- Betweenness Centrality is Fundamental to the APLS Metric



www.github.com/cosmiq/apls

Back to notebook 3_evaluating_performance for a use case!