DATA+AI SUMMIT
BY databricks

THE AI OF WHERE

UNLEASHING THE POWER OF GENAI ON GEOSPATIAL DATA

Steve Kingston
Geospatial Data Scientist, Ordnance Survey

Milos Colic
Technical Lead, Databricks

WHO WE ARE

Milos Colic, CARTO

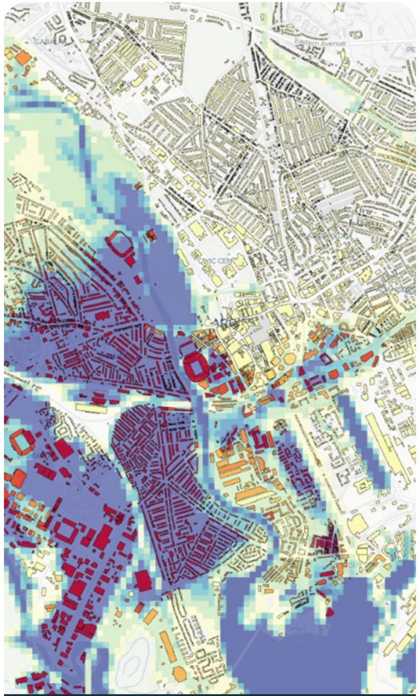
Platform APIs Lead

Steve Kingston, ORDNANCE SURVEY

Geospatial Data Scientist

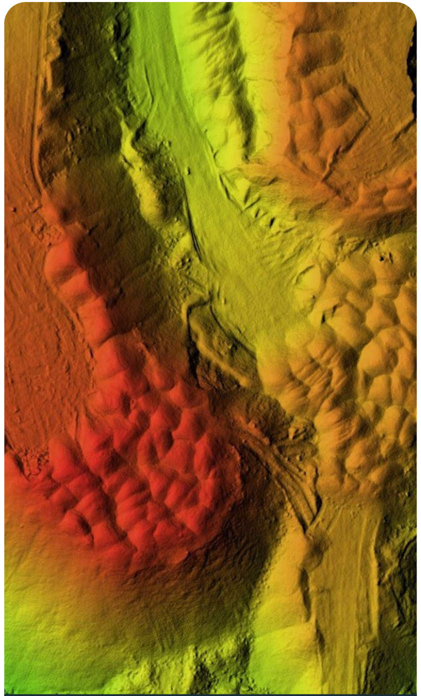
A	B
latitude	longitude
41.12887998	-73.7100635
41.13365032	-73.7099659
41.13009945	-73.70991433
41.13018385	-73.70989947
41.13043715	-73.70985008
41.1309958	-73.70976019
41.13354633	-73.70968264
41.13453456	-73.71234567
41.13987875	-73.71245671
41.13987988	-73.71267892
41.13987991	-73.714687912
41.14567889	-73.714522467

Anything with Lat/Long coordinates



Raster

Vector

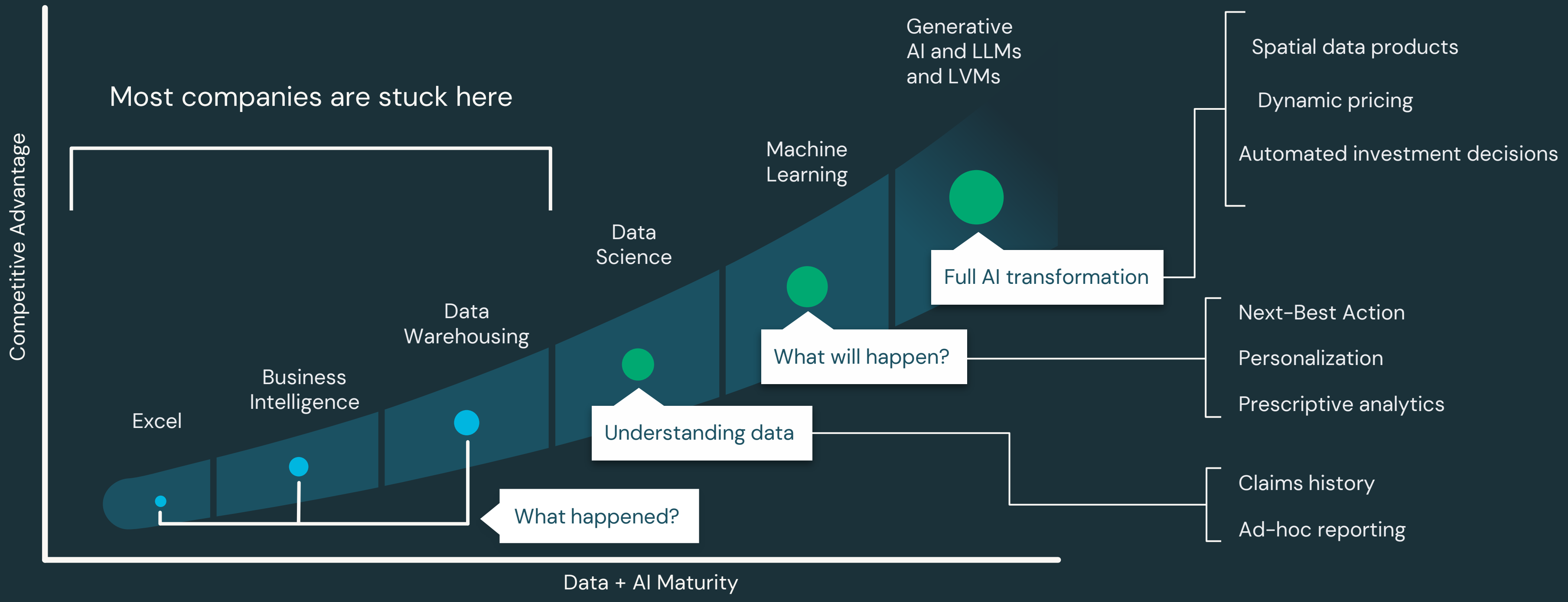


Lidar

Geospatial data is everywhere

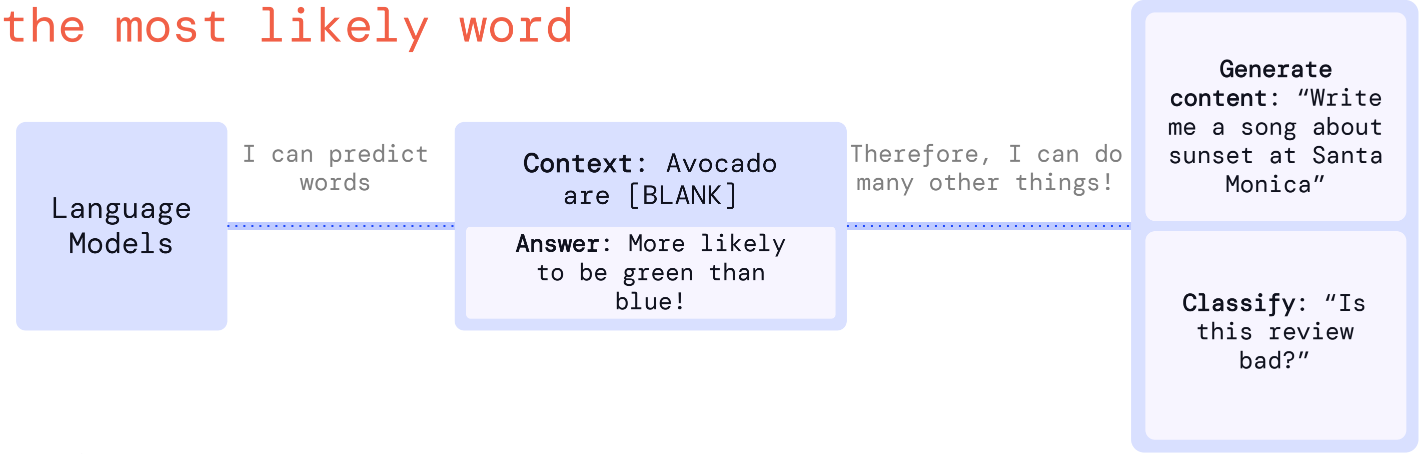
Data + AI maturity

Business driving technology, technology driving organizational changes



LARGE LANGUAGE MODELS (LLMs)

LLMs assign probabilities to word sequences: find the most likely word



Categories:

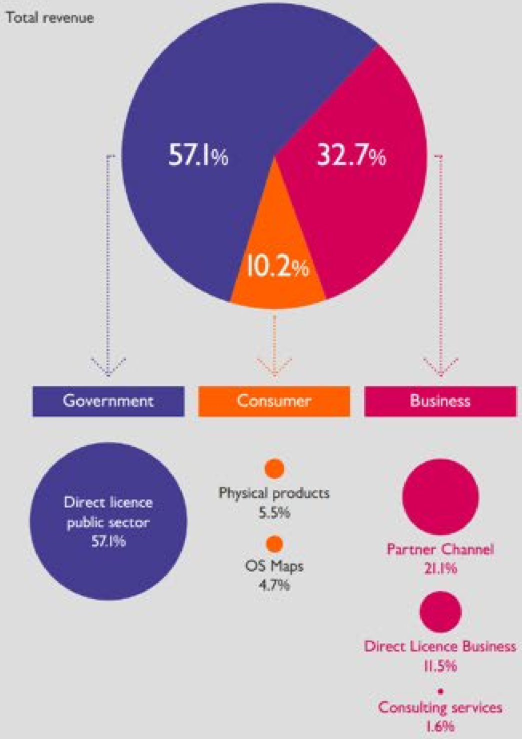
- Generative: find the most likely next word
- Classification: find the most likely classification/answer

LARGE VISUAL MODELS (LVMs)

SEGMENT ANYTHING MODEL

Ordnance Survey

As the **National Mapping Service**, Ordnance Survey (OS) creates, maintains, and disseminates consistent, definitive, and authoritative geospatial data of **Great Britain**.

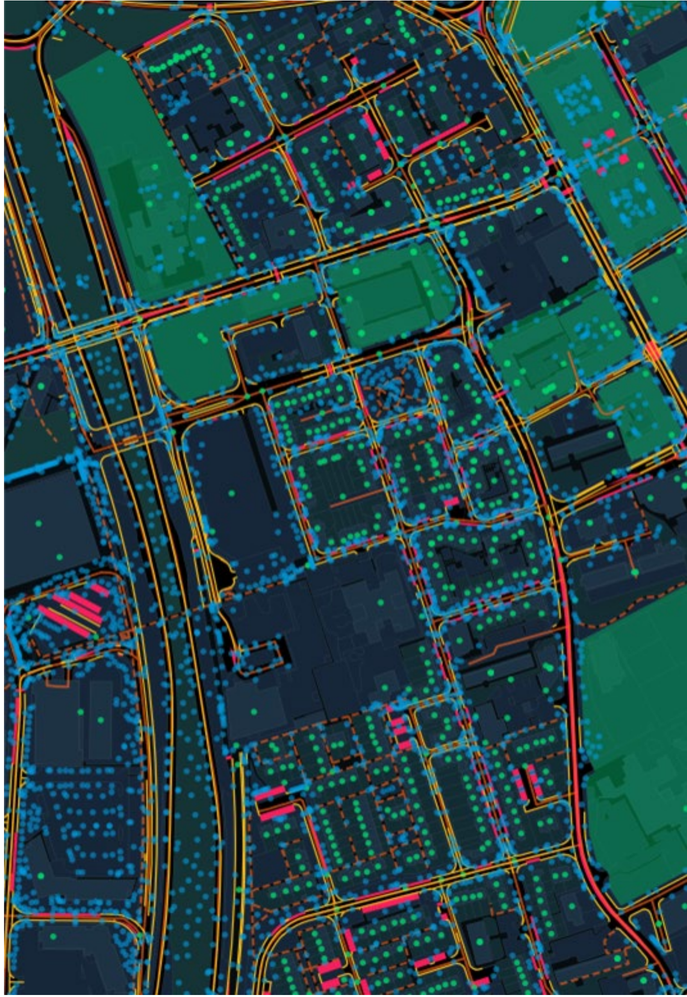


geoplace® GeoHub

OS NATIONAL GEOGRAPHIC DATABASE

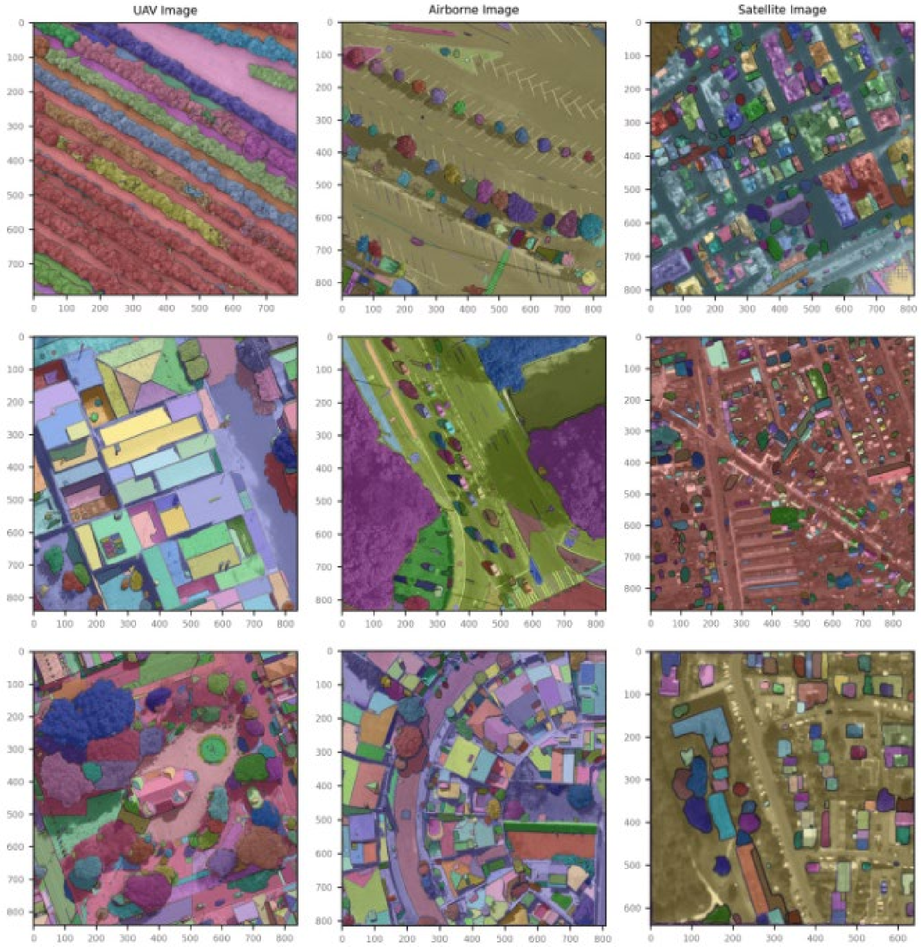
Cloud architecture evolution

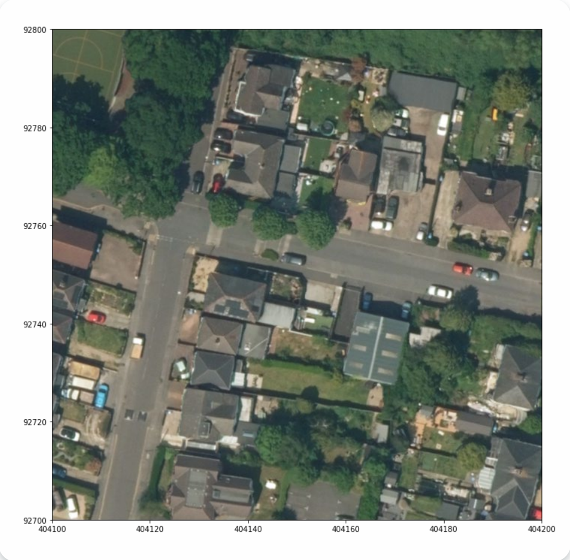
- The **Public Sector Geospatial Agreement (PSGA)** is the foundation of OS's national mapping services for Great Britain.
- The launch of the **National Geographic Database (NGD)** provides 6,000 public sector customers with location data and expertise, helping deliver more efficient public services.
- OS has been working closely with Databricks and Microsoft on the establishment and maintenance of **new cloud architecture and data capabilities to underpin the NGD**.
- The use of Azure Databricks backed by a Delta Lake storage architecture provided OS with an opportunity to re-think how to optimise both data and approach to **perform geospatial data processing and analytics at scale**.

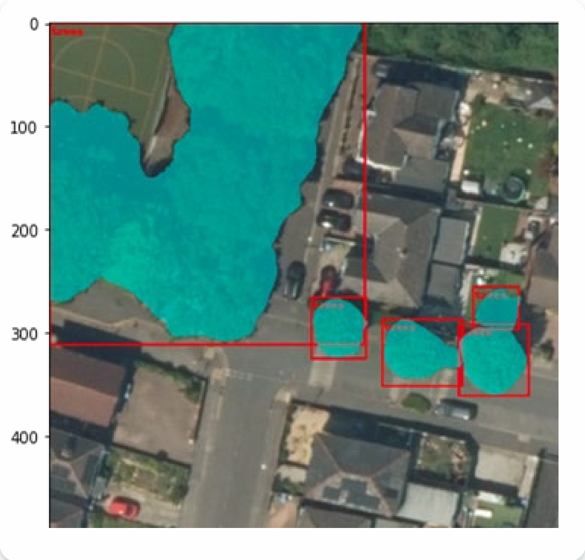


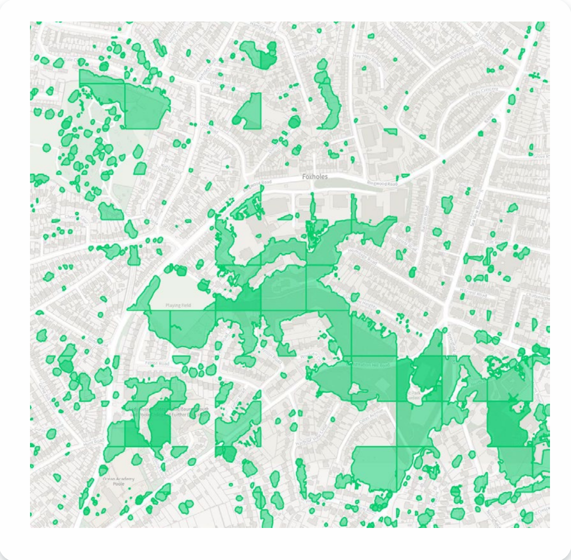
THE VIEW FROM ABOVE

APPLYING SAM
TO ORTHOIMAGERY









“Give me Trees”

Generate Geotagged vectors

LLM + LVM

Spatial Framework

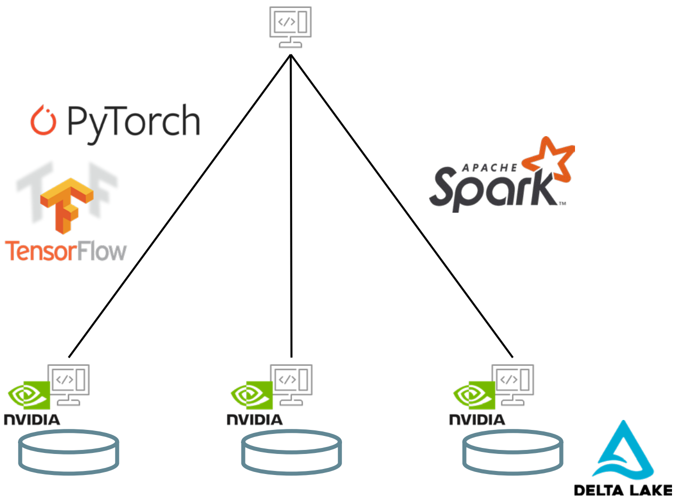
SCALING OUT

Distributed model serving

- Distributed Tensorflow
- Distributed PyTorch
- DeepSpeed
- Optionally run on Spark, Ray, etc.

Scaling pre/post processing

- Real-time: scale out end points
- Streaming and batch: Scale out pipelines, e.g. Spark + Delta Lake



SCALING BATCH INFERENCE USING RAY

- One of the **challenges OS has encountered** when experimenting with SAM on Databricks has been moving from a simple example, to **predicting at scale via distributed batch inference**.
- OS has tried, and failed*, to **tune Spark to enable a UDF-approach** - noting the size of both raster data and model.
- The availability announcement of the open-source compute framework **Ray on Databricks** presented an opportunity.
- What have we found?

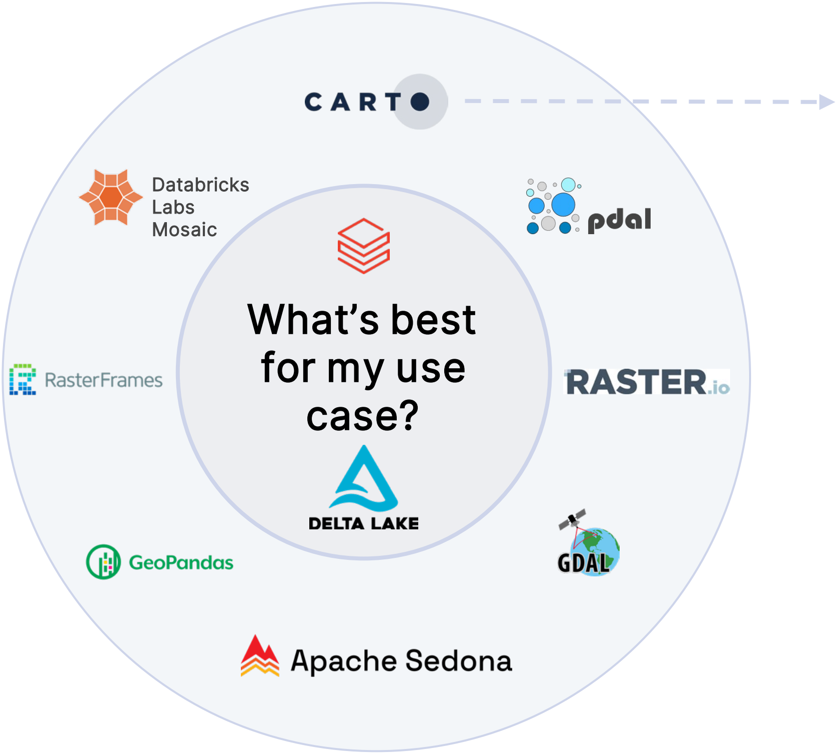
*the Spark UDF approach may well be feasible given additional Spark tuning and performance expertise.

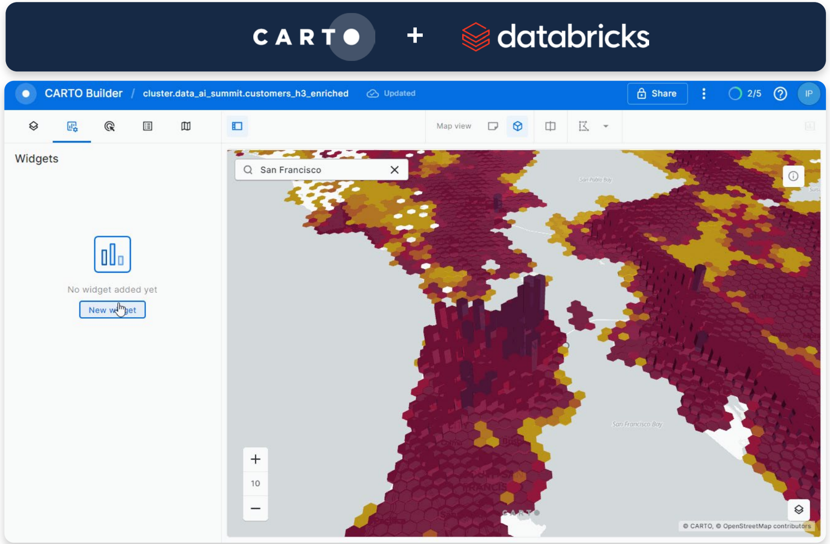
RAY - INITIAL QUALITATIVE OBSERVATIONS

- **Impressive performance 'out of the box'**, especially considering the Spark UDF approach was a non-starter.
- **Flexibility to use Spark tooling** for pre-processing and post-processing of data - see Ray as a computing step within a larger pipeline.
- More **control over batching and compute resources**, including option for **GPU-acceleration** (with minimal code changes).
- Ray provides an interesting option for testing methods and exploratory workloads – beyond ML applications.
- **New integration for Databricks** – worth noting a lack of detailed examples, tuning guidance and documentation.

Broad Platform Ecosystem

"Flexibility" to choose your own geospatial processing

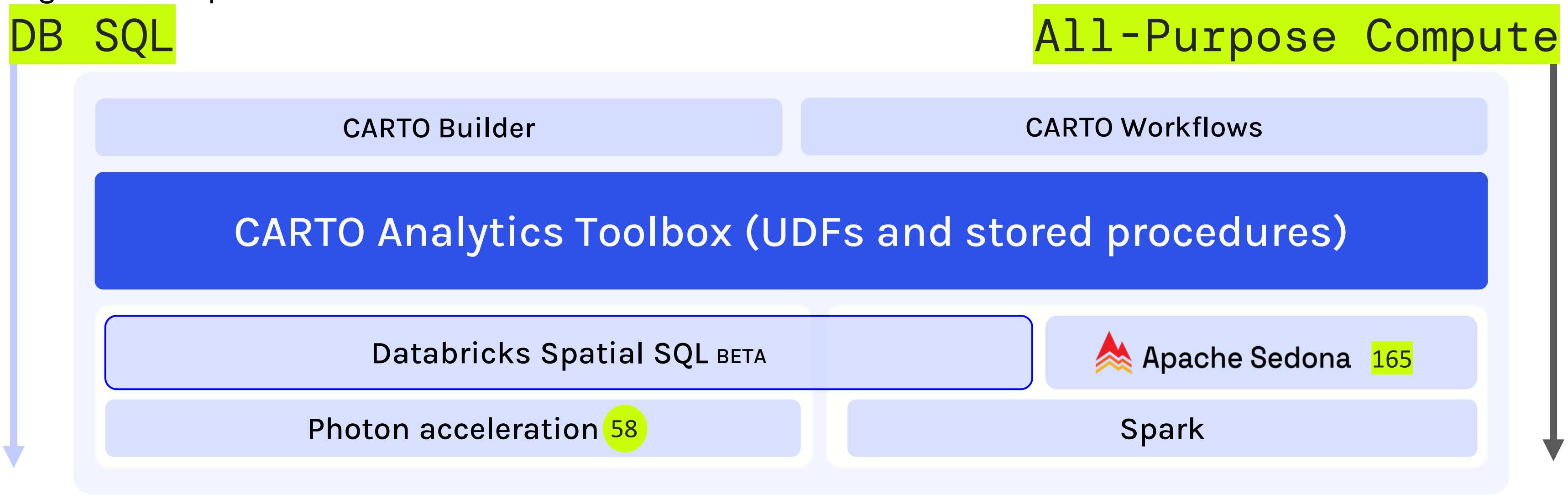




- H3 integration
- Analytics Toolbox
- pydeck-carto visualization
- Spatial Features for enrichment
- Databricks Data Marketplace

SUPPORTING THE DATABRICKS/SPARK ECOSYSTEM

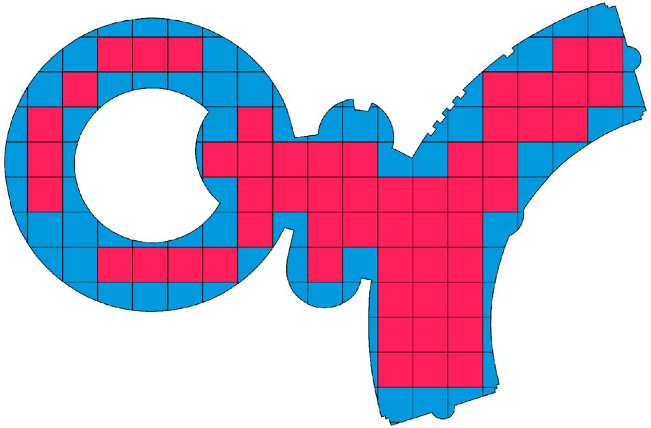
Fastest in Databricks with Spatial SQL and Photon acceleration, and available in generic Spark



GRID INDEXING AT SCALE

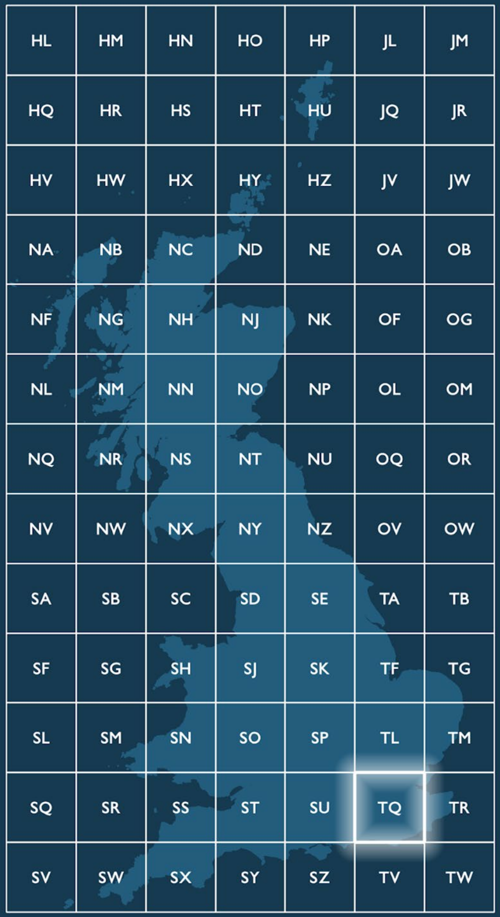
Co-development of an indexing strategy

- OS, Databricks, and Microsoft co-developed a strategy that disaggregates geometries into simplified representations bounded by their presence in each index – implemented for both BNG and H3 in **Databricks Mosaic.**
- **Classifies 'core' indices** contained by the feature geometry which support join optimisations.
- Use the index identifier attribute as a **join key to collocate rows** and then only test a spatial predicate within those collocated rows.
- Also **serves as a subdivide function** to transform large and complex geometries into simpler 'chips'.



MOSAIC

databricks × Microsoft Azure × Ordnance Survey


HL HM HN HO HP JL JM
HQ HR HS HT HU JQ JR
HV HW HX HY HZ JV JW
NA NB NC ND NE OA OB
NF NG NH NJ NK OF OG
NL NM NN NO NP OL OM
NQ NR NS NT NU OQ OR
NV NW NX NY NZ OV OW
SA SB SC SD SE TA TB
SF SG SH SJ SK TF TG
SL SM SN SO SP TL TM
SQ SR SS ST SU TQ TR
SV SW SX SY SZ TV TW


SP
TL
TM
TQ09 TQ19 TQ29 TQ39 TQ49 TQ59 TQ69 TQ79 TQ89 TQ99
TQ08 TQ18 TQ28 TQ38 TQ48 TQ58 TQ68 TQ78 TQ88 TQ98
TQ07 TQ17 TQ27 TQ37 TQ47 TQ57 TQ67 TQ77 TQ87 TQ97
TQ06 TQ16 TQ26 TQ36 TQ46 TQ56 TQ66 TQ76 TQ86 TQ96
TQ05 TQ15 TQ25 TQ35 TQ45 TQ55 TQ65 TQ75 TQ85 TQ95
SU
TQ
TR
TQ04 TQ14 TQ24 TQ34 TQ44 TQ54 TQ64 TQ74 TQ84 TQ94
TQ03 TQ13 TQ23 TQ33 TQ43 TQ53 TQ63 TQ73 TQ83 TQ93
TQ02 TQ12 TQ22 TQ32 TQ42 TQ52 TQ62 TQ72 TQ82 TQ92
TQ01 TQ11 TQ21 TQ31 TQ41 TQ51 TQ61 TQ71 TQ81 TQ91
TQ00 TQ10 TQ20 TQ30 TQ40 TQ50 TQ60 TQ70 TQ80 TQ90
SZ
TV
TW

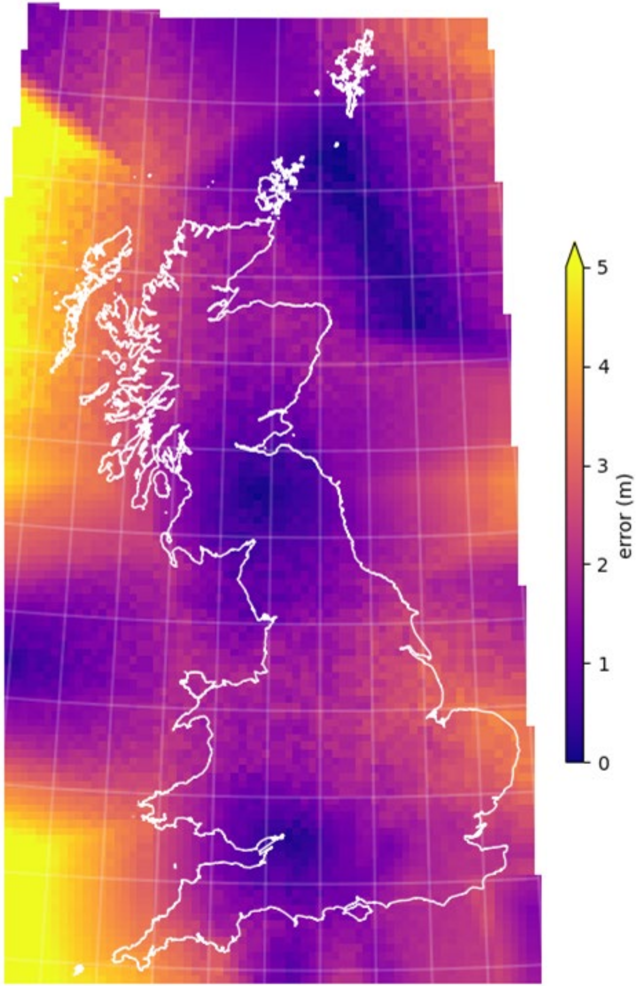
Source: DALL·E

WHY THE BRITISH NATIONAL GRID?

Whilst there are alternative global index systems that OS could adopted, we **choose a BNG-first approach** because:

1. The BNG system is **native to OS's geospatial data collection**, with almost all OS data referenced against the BNG CRS.
2. Using BNG avoids the (very) **costly reprojection to WGS84** (or ETRS89) CRSs via the OSTN15 transformation grid.
3. Square tiles are naturally more suited to the indexing of raster data.

For OS BNG is a natural choice for GB-local use cases, whereas H3 is a more suitable global and European alternatives.



A RASTER MEDALLION CLOUD DATA ARCHITECTURE

Bronze
(Raw)

- Landing zone for **'raw'** raster data.
- Little to no data architecture investment.

Silver
(Base)

- Raster tiles stored using **GeoTIFF** format.
- Accompanying **Delta metadata table** providing acquisition date and time, resolution, bands etc.

Gold
(Indexed)

- **Indexed raster tiles** subdivided by BNG/H3 indexing strategy.
- Raster represented using **binary type** (or other universal type).

Platinum
(Derive and Enrich)

- Rasters clipped by the **'highest value'** vector geometries.
- Derived from indexed raster tiles via collocate, clip, and merge steps.

DELTA LAKE

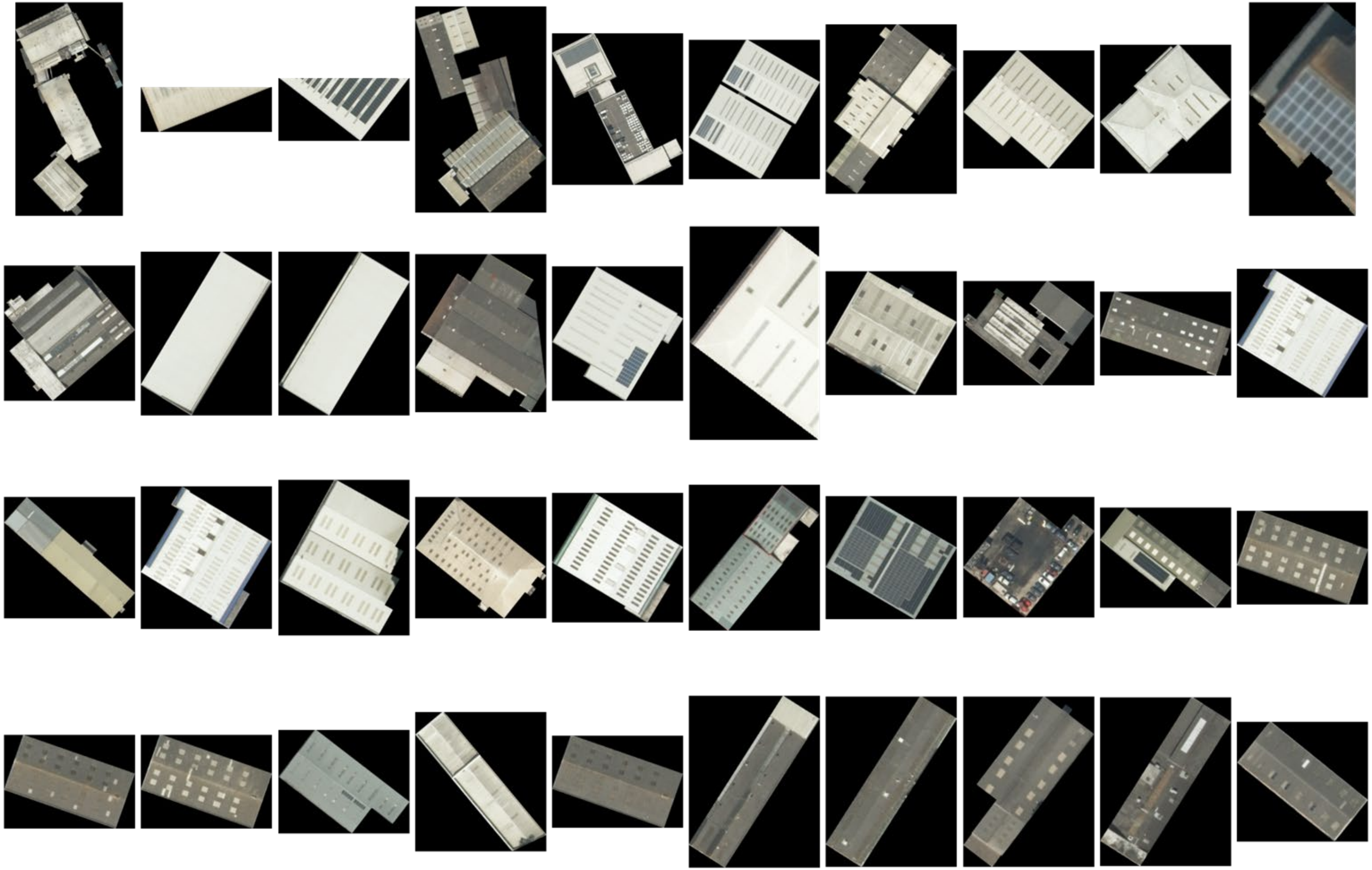
RASTER CLIP - MERGE

Databricks Mosaic PySpark Code

PYTHON

```
# Collocate indexed raster and indexed vector on index identifier
clip = (
    raster.alias("a")
    .join(
        vector.alias("b"),
        on=F.col("a.tile.index_id") == F.col("b.index_id"),
        how="inner",
    )
# Clip indexed raster to indexed vector geometry
    .withColumn("tile", mos.rst_clip(F.col("a.tile"), F.col("b.wkb")))
)

# Merge clipped raster components by vector feature identifier
   merge = clip.groupBy(F.col("fid")).agg(
    mos.rst_merge_agg(F.col("tile")).alias("tile")
)
```



GeoParqet & RasQuet

Bringing RASTER into the Data Cloud via PARQUET. An Open format looking for feedback and collaboration.

Geoparquet 1.1

Vector data encoded in WKB or GeoArrow.

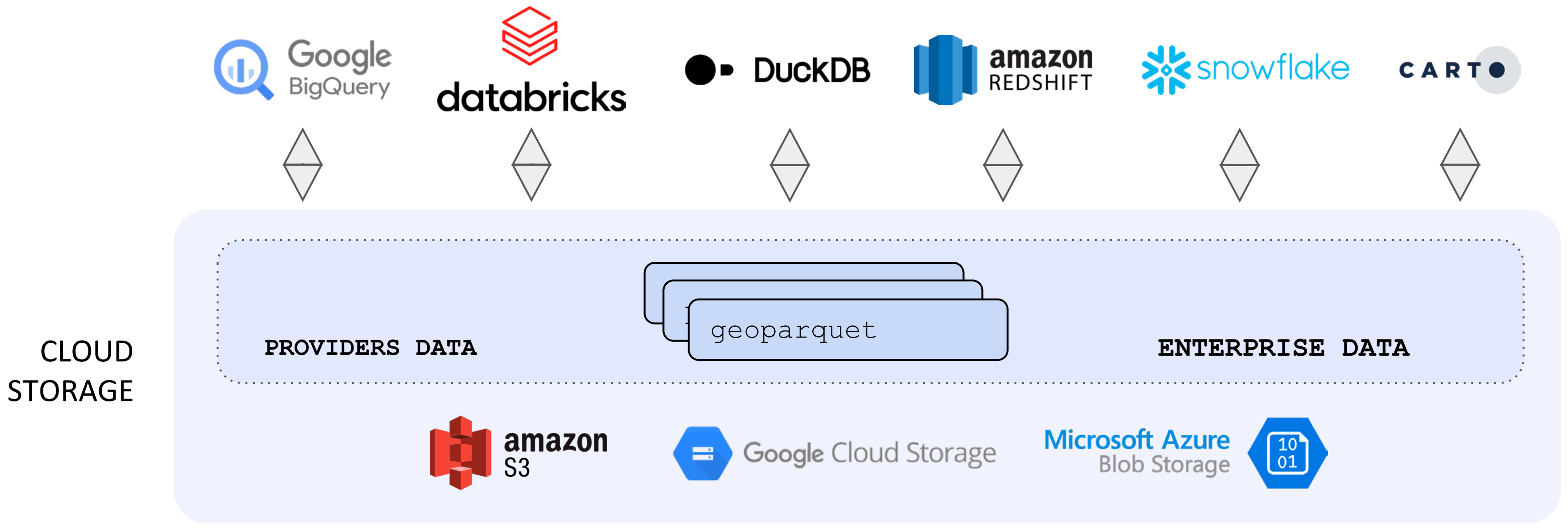
On its way to become OGC standard.

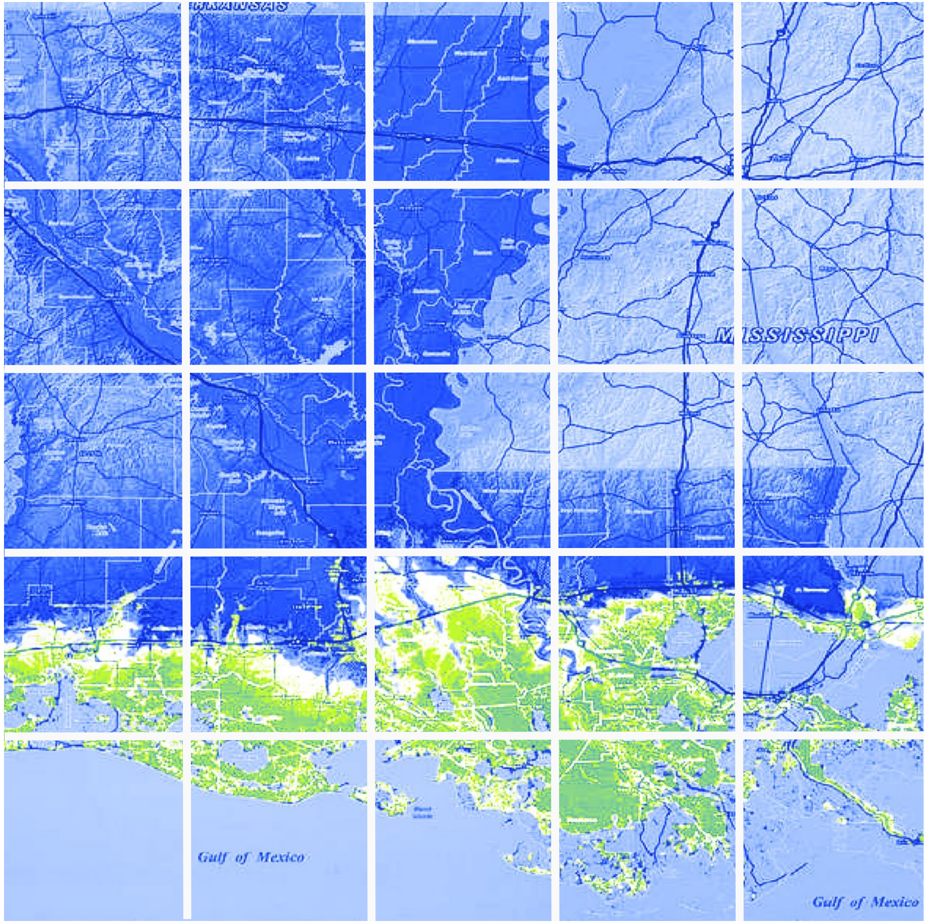
RasQuet (Parquet Raster) 0.1

Raster data encoded in native arrays or Zarr

- Support for projections
- Different encodings
- Multidimensional

ENABLE INTEROPERABILITY BETWEEN DIFFERENT SYSTEMS FOR GEO DATA





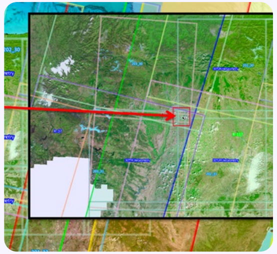
tile_id	time	variable	raster	metadata
5211548776216395775	2024-05-14	RiskScore	> [3.32,4.52,4.07,4.39,3.72,6....	> {units=m ...
5211548776216395775	2024-05-15	RiskScore	> [3.72,4.86,4.16,4.91,5.35,3....	> {units=m ...
5211548776216395775	2024-05-16	RiskScore	> [3.79,4.36,3.89,3.15,6.24,5....	> {units=m ...
5211548776216395775	2024-05-14	elevation	> [1053.0200683952853,1090...	> {long_na...
5211548776216395775	2024-05-15	elevation	> [1009.894736543188,1014....	> {long_na...
5211548776216395775	2024-05-16	elevation	> [1095.5421888278183,1056...	> {long_na...
5211447621146640383	2024-05-14	RiskScore	> [9.19,8.64,6.66,8.97,9.11,9....	> {units=m ...
5211447621146640383	2024-05-15	RiskScore	> [8.2,8.24,8.92,9.92,6.95,9.5,...	> {units=m ...
5211447621146640383	2024-05-16	RiskScore	> [8.98,7.14,7.15,8.88,6.75,7....	> {units=m ...
5211447621146640383	2024-05-14	elevation	> [1038.77122032916,1042.2...	> {long_na...
5211447621146640383	2024-05-15	elevation	> [1098.5488845815894,1111...	> {long_na...
5211447621146640383	2024-05-16	elevation	> [1029.40142420468,1102.0...	> {long_na...
5211355262169907199	2024-05-14	RiskScore	> [6.57,7.77,8.03,8.6,9.01,8.9...	> {units=m ...
5211355262169907199	2024-05-15	RiskScore	> [9.14,8.22,9.17,9.33,8.38,8....	> {units=m ...

```
SELECT
 RASTER_VALUE(
   raster, (11, 23))
FROM FloodRiskRasquet
WHERE date = '2024-05-16'
AND variable = 'RiskScore'
```

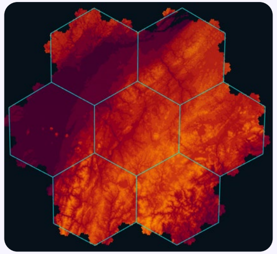
Unified Data Model for all GIS Data

Vector + Raster for "All Spatial" Analysis

Earth Observation Data / Planetary Computer

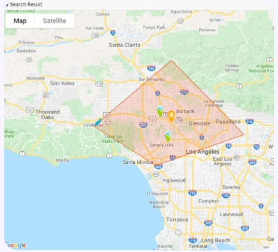


Earth Observation Data



Raster Grid Tiles

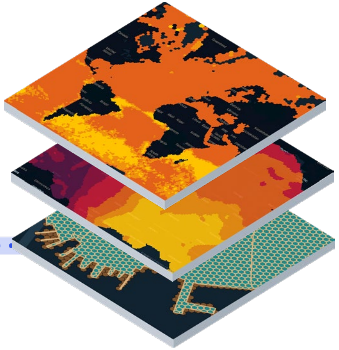
Areas of interest



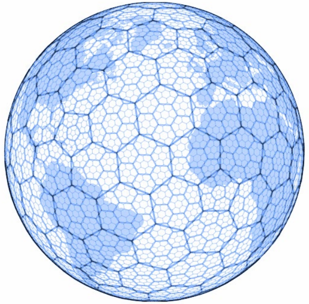
Areas of Interest as Open Spatial Formats

Vector Grid Tiles

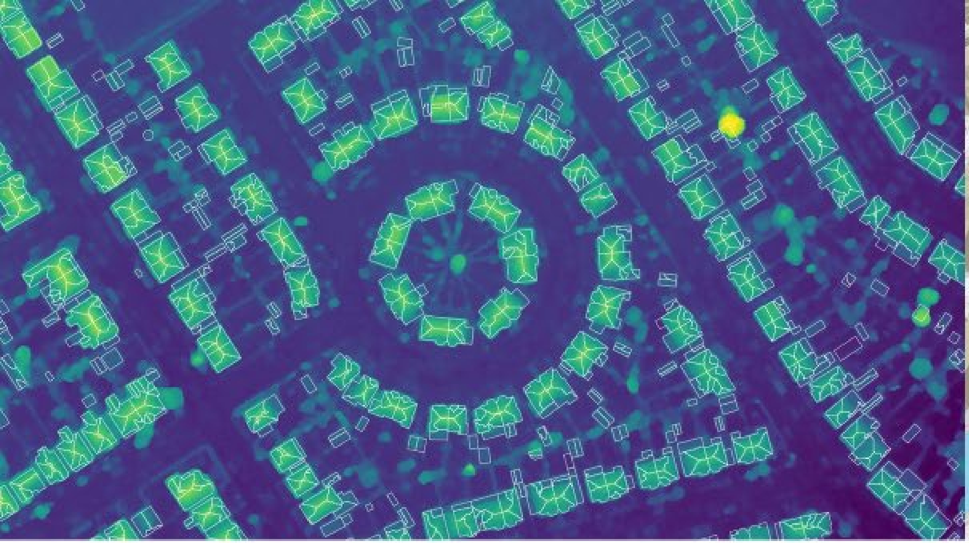
Unified Data Model

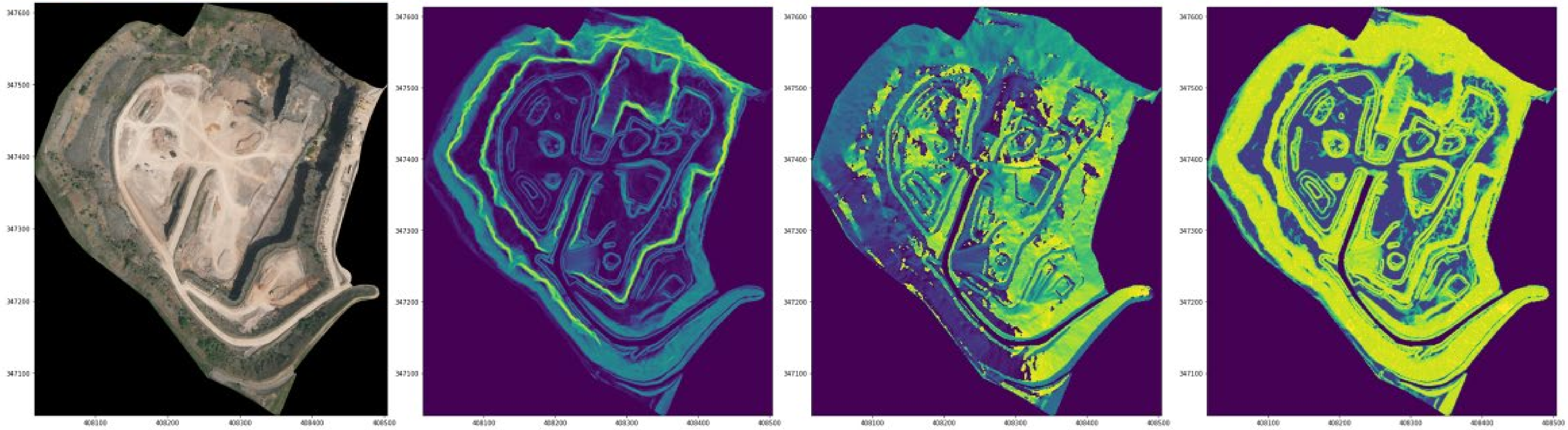


Observable Earth Model



- All data uses same grid w/CRS abstracted
- Layer joining + stacking is just idx ==
- Stored as Delta; metadata + data in the same format; Delta Sharing ready
- Marketplace ready
- AI ready




347600
347500
347400
347300
347200
347100
408100
408200
408300
408400
408500

EXPERIMENTATION USING FOUNDATION MODELS

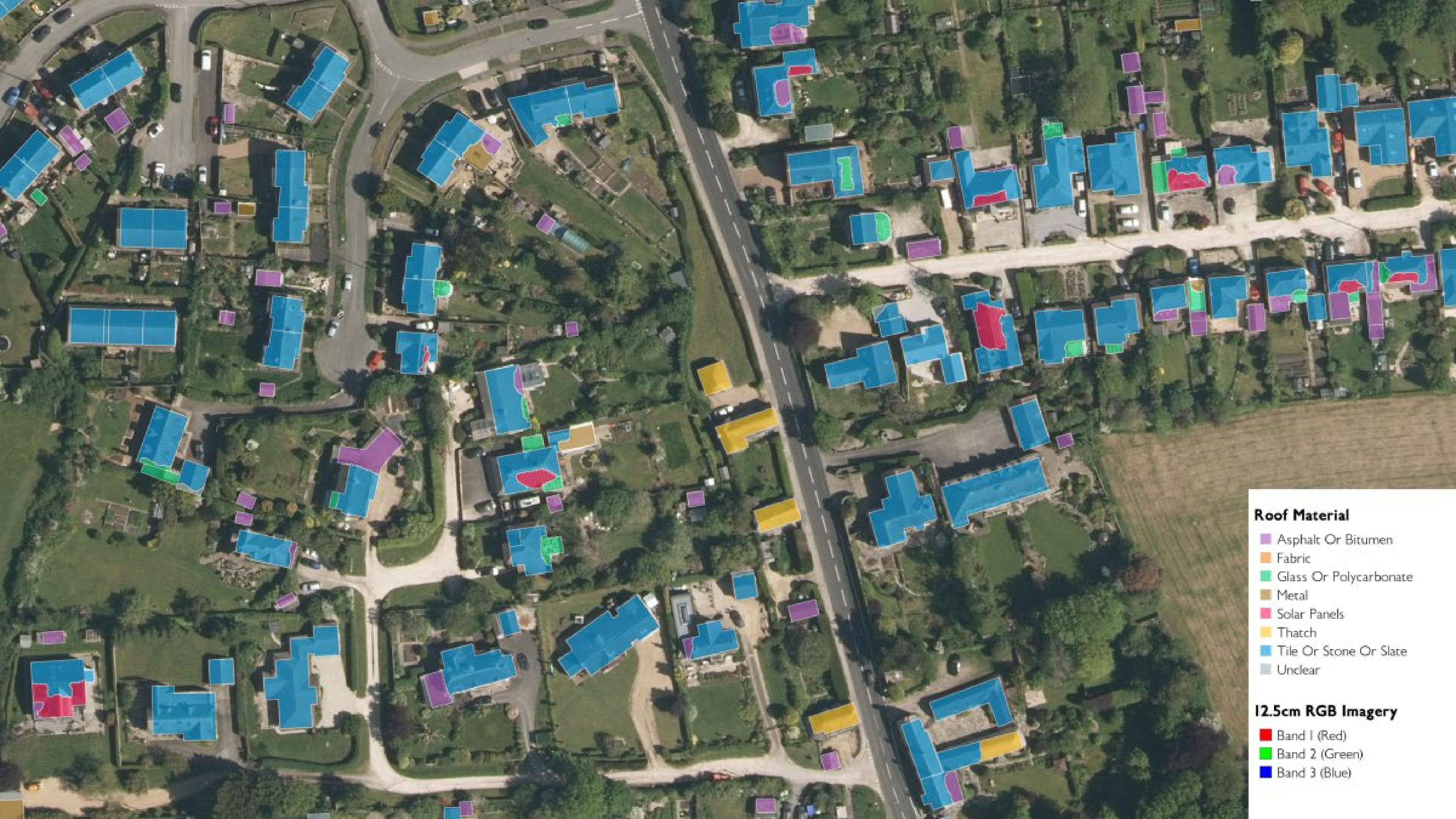
- **Zero-shot** application of Meta's foundation Segment Anything Model (SAM) provides an opportunity to **accelerate use case exploration** – enables rapid proof-of-concept and experimentation with no training data requirement.
- Object detection and instance segmentation working with text prompts **e.g. 'Car'** via Language SAM model.
- Expecting future developments **tailored specifically to aerial and satellite imagery**:
- OS has attempted to refine the SAM model for certain use cases ('personalised-SAM').
- Further evaluation of methods and potential required.

8cm resolution

4cm resolution

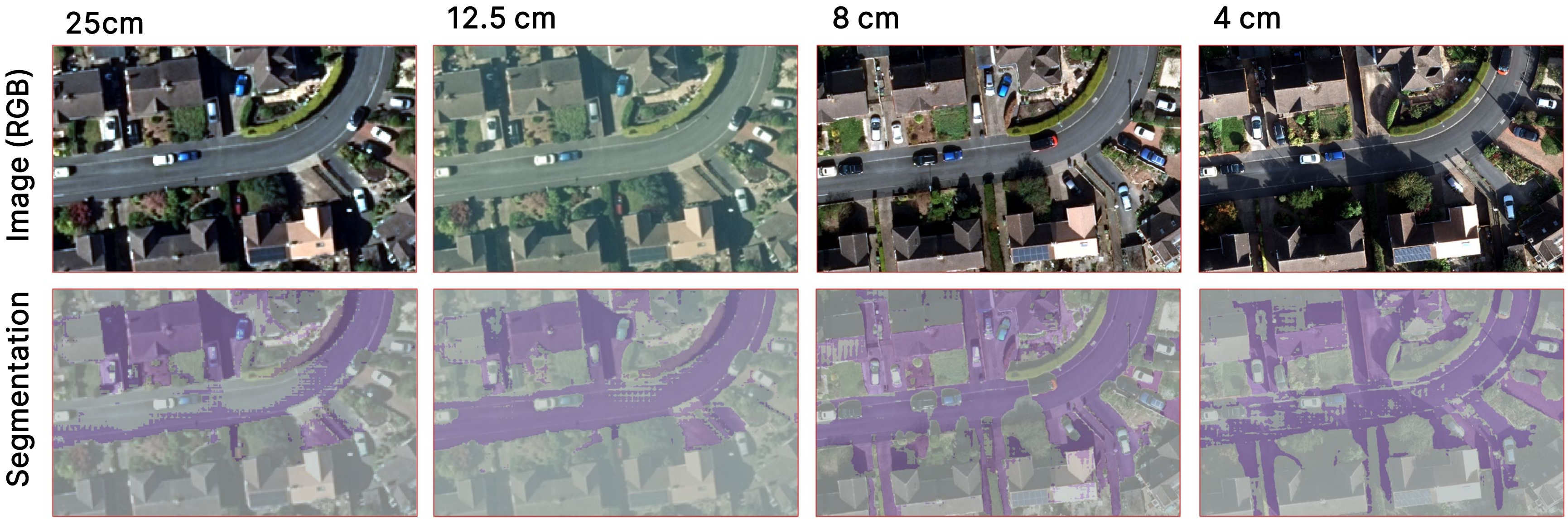
FINE-TUNING SAM

- In addition to zero-shot application, SAM can be **fine-tuned** to target a more specific segmentation task using prompts in the form of a point dataset.
- Potential to **automate prompt point creation** from other geospatial reference data.
- This example explores point placement against **‘driveways’** – related to use cases including off-street EV charging capacity.
- Considerations (and limitations) related to **use case specificity**.


Roof Material
Asphalt Or Bitumen
Fabric
Glass Or Polycarbonate
Metal
Solar Panels
Thatch
Tile Or Stone Or Slate
Unclear
12.5cm RGB Imagery
Band 1 (Red)
Band 2 (Green)
Band 3 (Blue)

SEGMENTATION CAPABILITY ACROSS DIFFERENT SPATIAL RESOLUTIONS

Exploring the segmentation capability of a tuned-SAM against different resolutions of OS aerial imagery*:

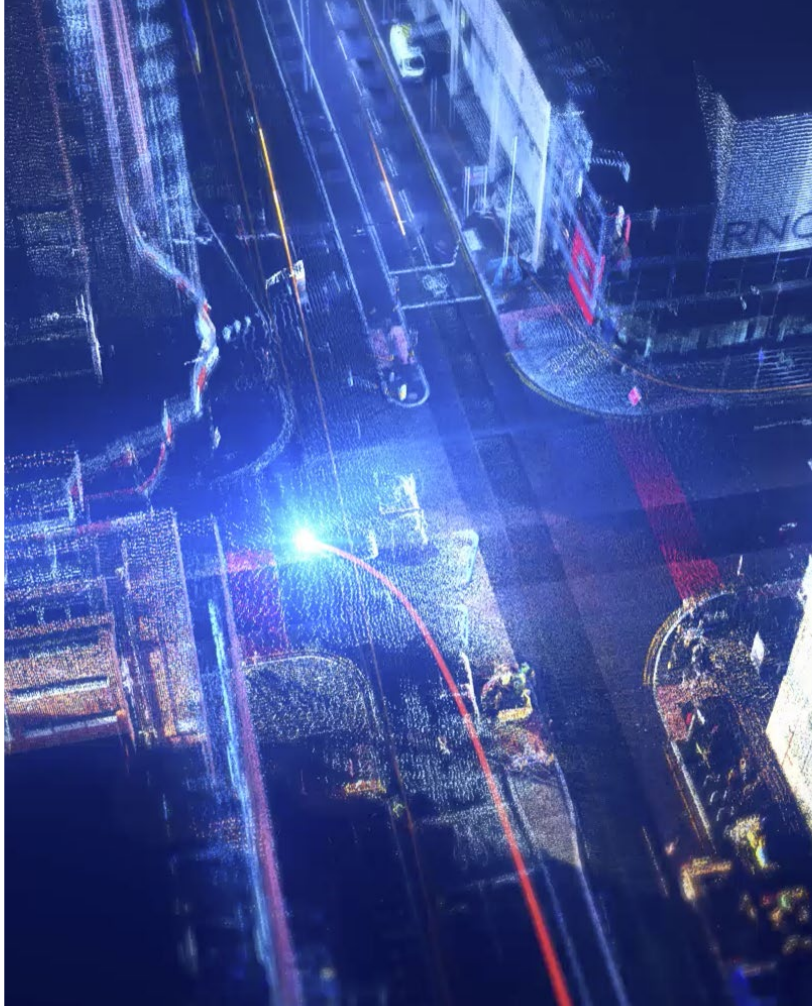


*Note, imagery captured over different date and time ranges.

DIFFERENT SPATIAL RESOLUTIONS

Trade-offs

- Experimenting with segmentation capability against higher resolution imagery shows **opportunities for improved delineation of small features**.
- However, there's likely a **'sweet-spot'/compromise** to be found considering trade-offs/factors including:
    1. Flying programme – currency, time of day, and seasonality (leaf-on vs leaf-off).
    2. Data management and storage costs.
    3. Processing costs.
    4. Prediction/detection quality.



ALTERNATIVE VIEWING ANGLES

- **Oblique imagery** has the potential to **increase the level of detail** and detect otherwise obscured features. For example, oblique angles could be used to derive insights around the facades of buildings, window and door layout, and detail underneath canopies and overhangs.
- Presents **new challenges** of georeferencing segmentation masks.
- Requires **processing multiple frames** and combining detections to target complete coverage.

THANK YOU

Any questions?

Milos Colic, CARTO
Platform APIs Lead

Steve Kingston, ORDNANCE SURVEY
Geospatial Data Scientist

Stop by to chat at Booth #MP3 (Marketplace Area)

WHO WE ARE

Stop by to chat at Booth #MP3 (Marketplace Area)

Milos Colic, CARTO

Platform APIs Lead

Steve Kingston, ORDNANCE SURVEY

Geospatial Data Scientist