# Geospatial Analytics at Scale with Deep Learning and Apache Spark

Tim Hunter & Raela Wang

Spark Summit Europe 2018

databricks

# About Us

- Raela Wang
- Solutions Architect @ Databricks
- Specialist in Machine Learning solutions

# About Us

- Tim Hunter
- Software engineer & Solutions Architect @ Databricks
- Ph.D. from UC Berkeley in Machine Learning
- Very early Spark user
- Contributor to MLlib
- Co-author of Deep Learning Pipelines, TensorFrames and GraphFrames

# Outline

- Geospatial imaging at scale
- How does Apache Spark help?
- An example pipeline with Spark
  - Deep Learning Pipelines
  - Magellan

# Mapping the world

- One of the most ancient big data activities in the world
- Critical for navigation, warfare, commercial exploitation

# Mapping the world with images

- A lot of tools and companies now provide geospatial solutions
- Increasingly done with a combinations of satellites, airplanes and drones

© Wired / Planet Labs

# Mapping the world with images

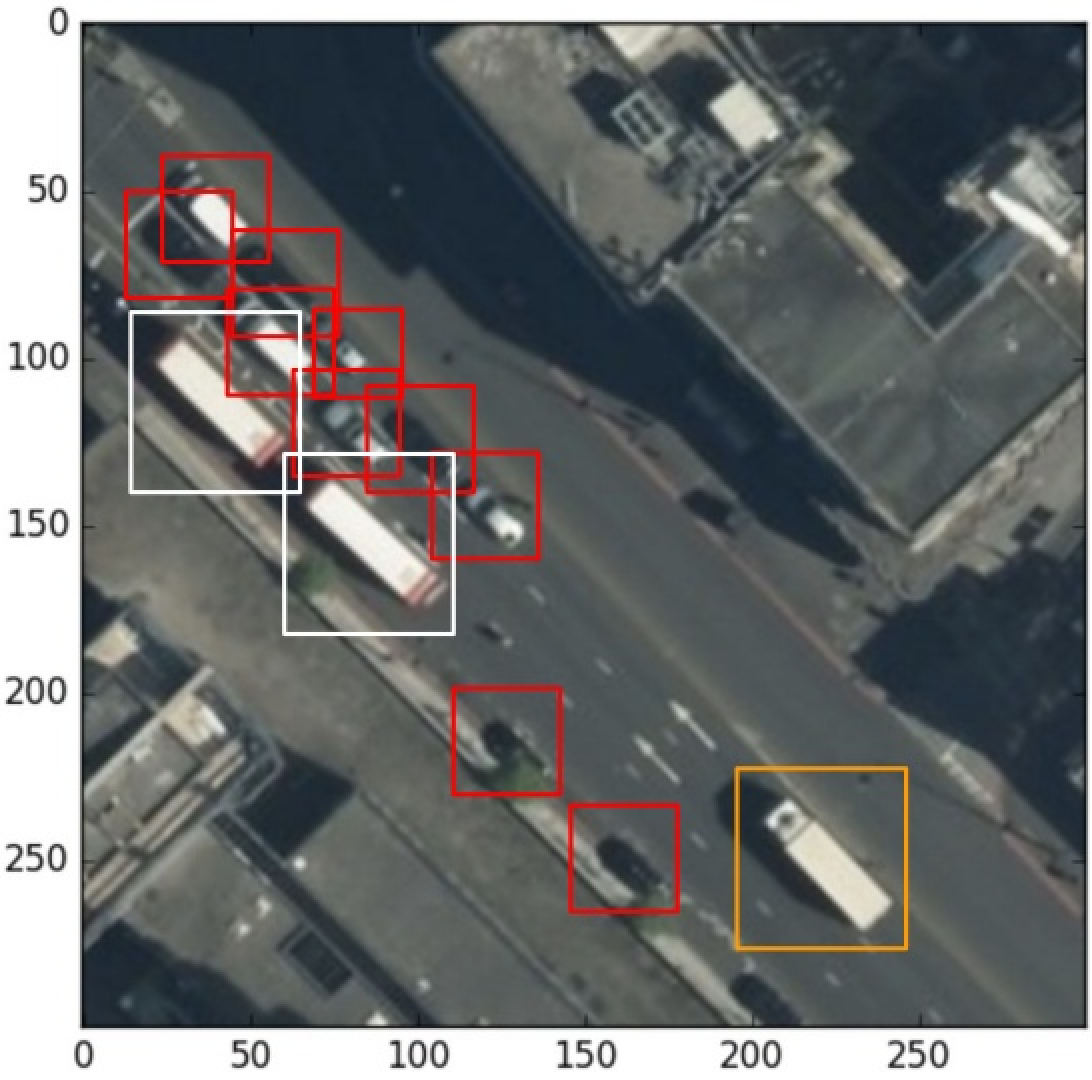
Large range of new applications

- Disaster Recovery: flood survey, fallen trees
- Infrastructure management: road damages
- Economic Intelligence: roof inclination for solar panels
- Insurance Fraud - private pools, backyard square footage

# New Challenges

- Increasing amounts of Rich Data
  - Cost effective solutions for acquiring data at scale (Drones, CubeSats)
- Difficulty Scaling
  - Traditional tools not designed for scalability: how to work at the scale of a country or a continent?
- Pipelining Challenges
  - Geospatial combines a lot of tools and problems: alignment, image corrections, object detection, …
  - All these technologies need to communicate data in a timely fashion

# An example:
# Identifying Vehicles in Aerial Imagery

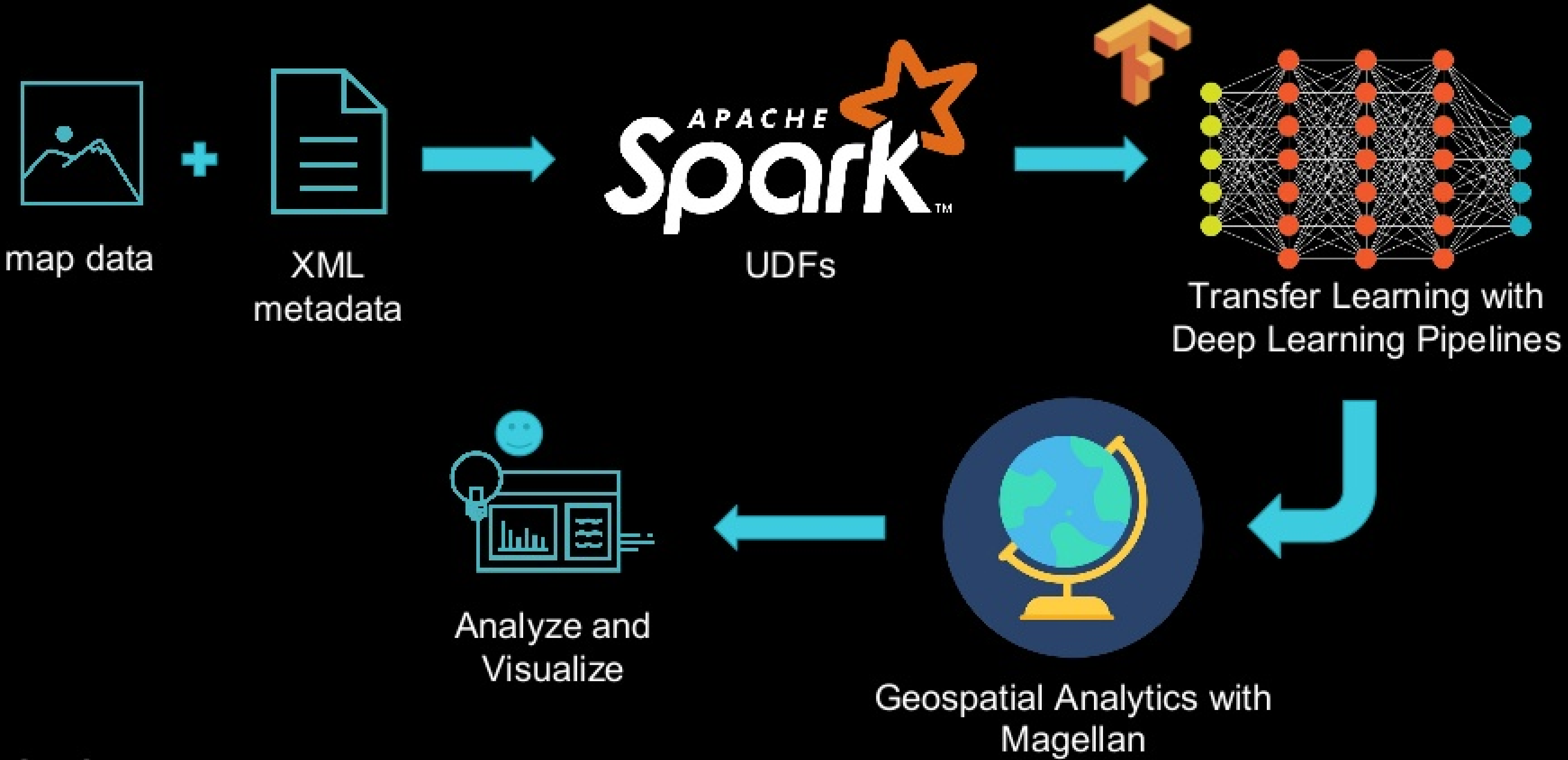


```
vehicle_classes = {
18:('car', 'red'),
23:('truck', 'orange'),
19:('bus', 'white', 0.0)}
```

# Apache Spark - the glue of big data

- Technologies exist in isolation
  - OpenCV - Image manipulation
  - Tensorflow - Deep Learning
  - PostGIS, GeoMesa, Magellan - Geospatial Analytics
  - Leaflet.js/OpenStreetMap - Visualization
- Apache Spark - ties all these libraries together
  - At scale
  - Allows pipelining
  - Easily move data from 1 technology to another without having to think about data representation

# High-level View of the Pipeline



map data

XML metadata

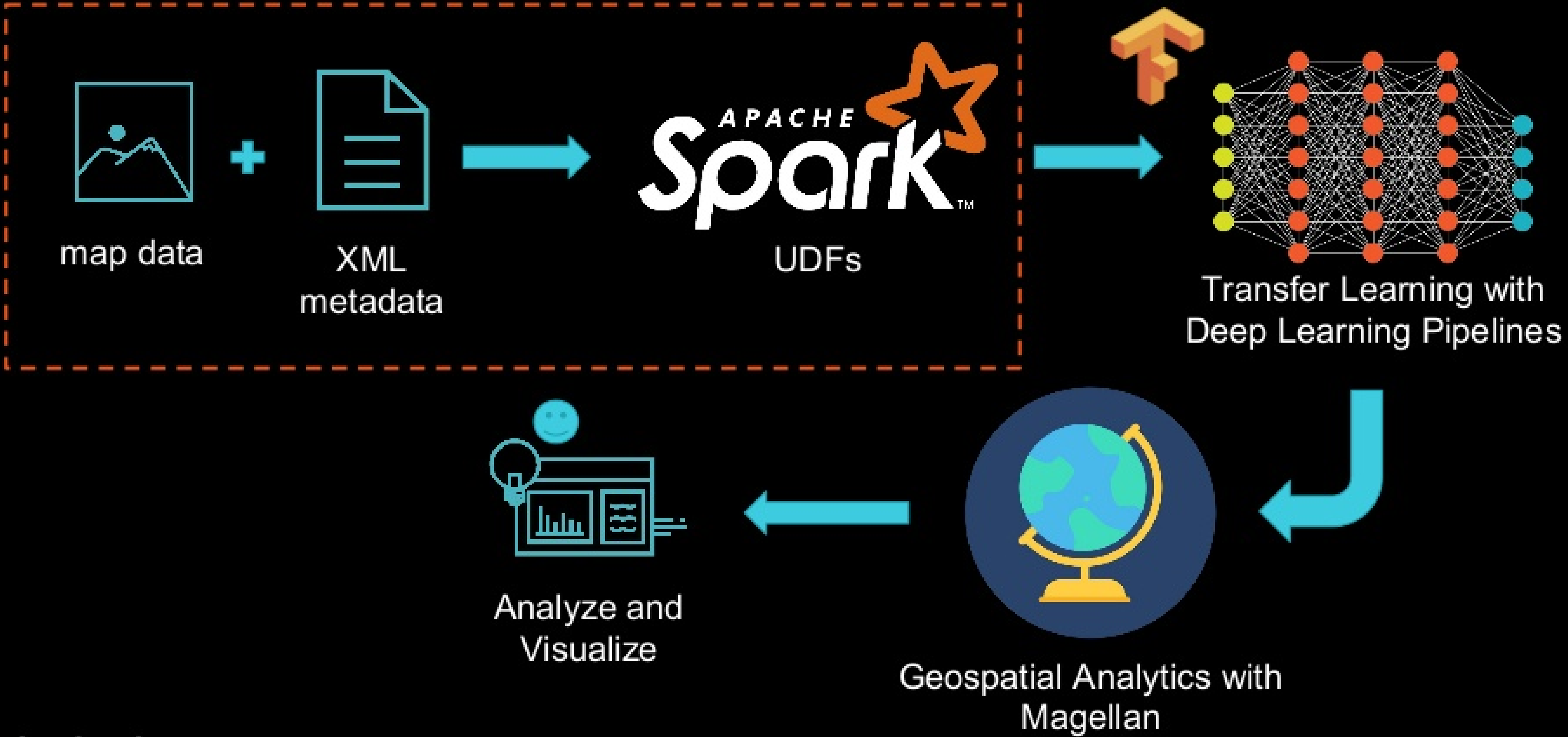
UDFs

Transfer Learning with Deep Learning Pipelines

Geospatial Analytics with Magellan

Analyze and Visualize

# Parsing Image Data



map data

XML metadata

UDFs

Transfer Learning with Deep Learning Pipelines

Geospatial Analytics with Magellan

Analyze and Visualize

# Ingesting Images

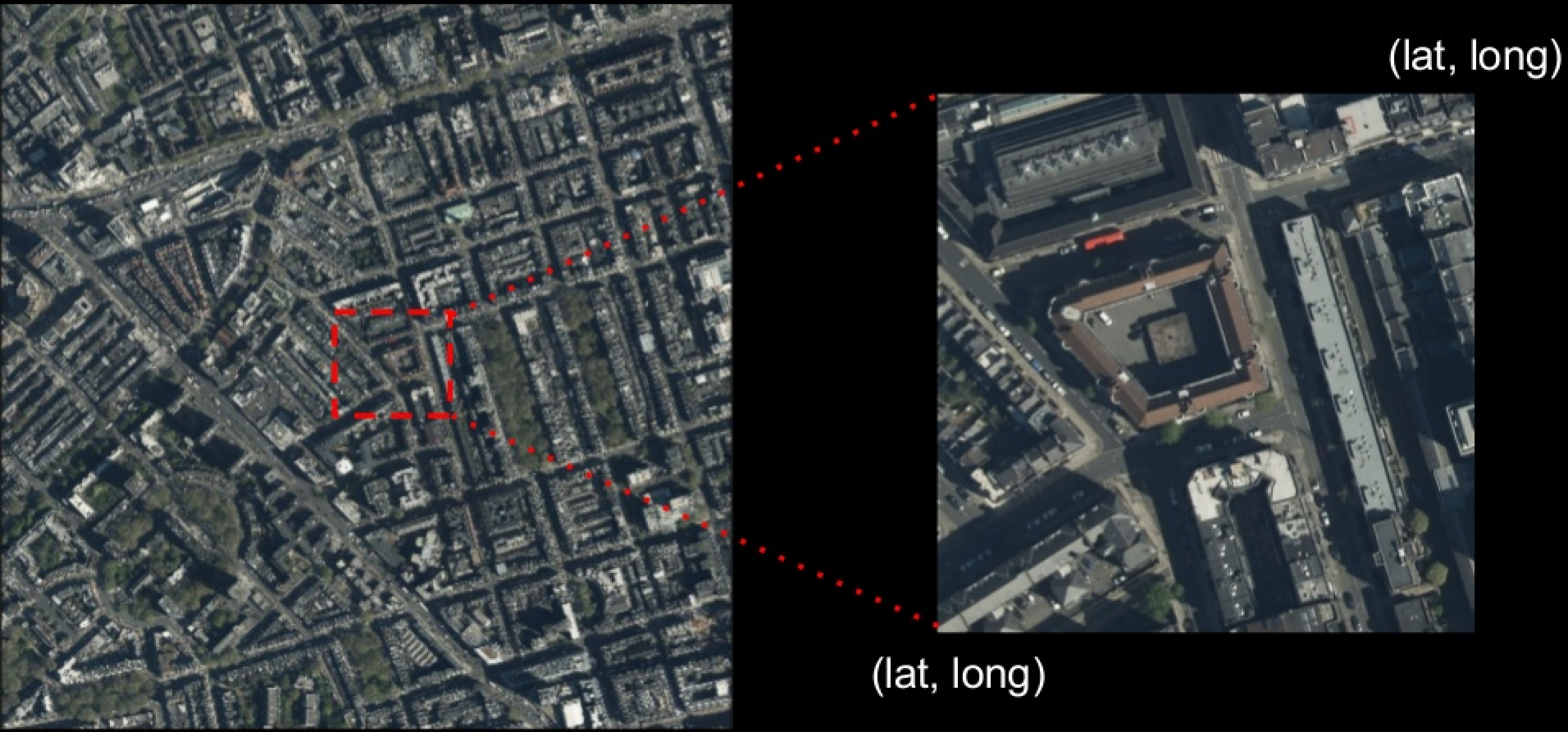
Spark 2.3 -- ImageSchema to Read/Write Image data

- Use the same schema across packages
  - Scikit-image, MMLSpark, OpenCV, PIL, Deep Learning Pipelines, ...

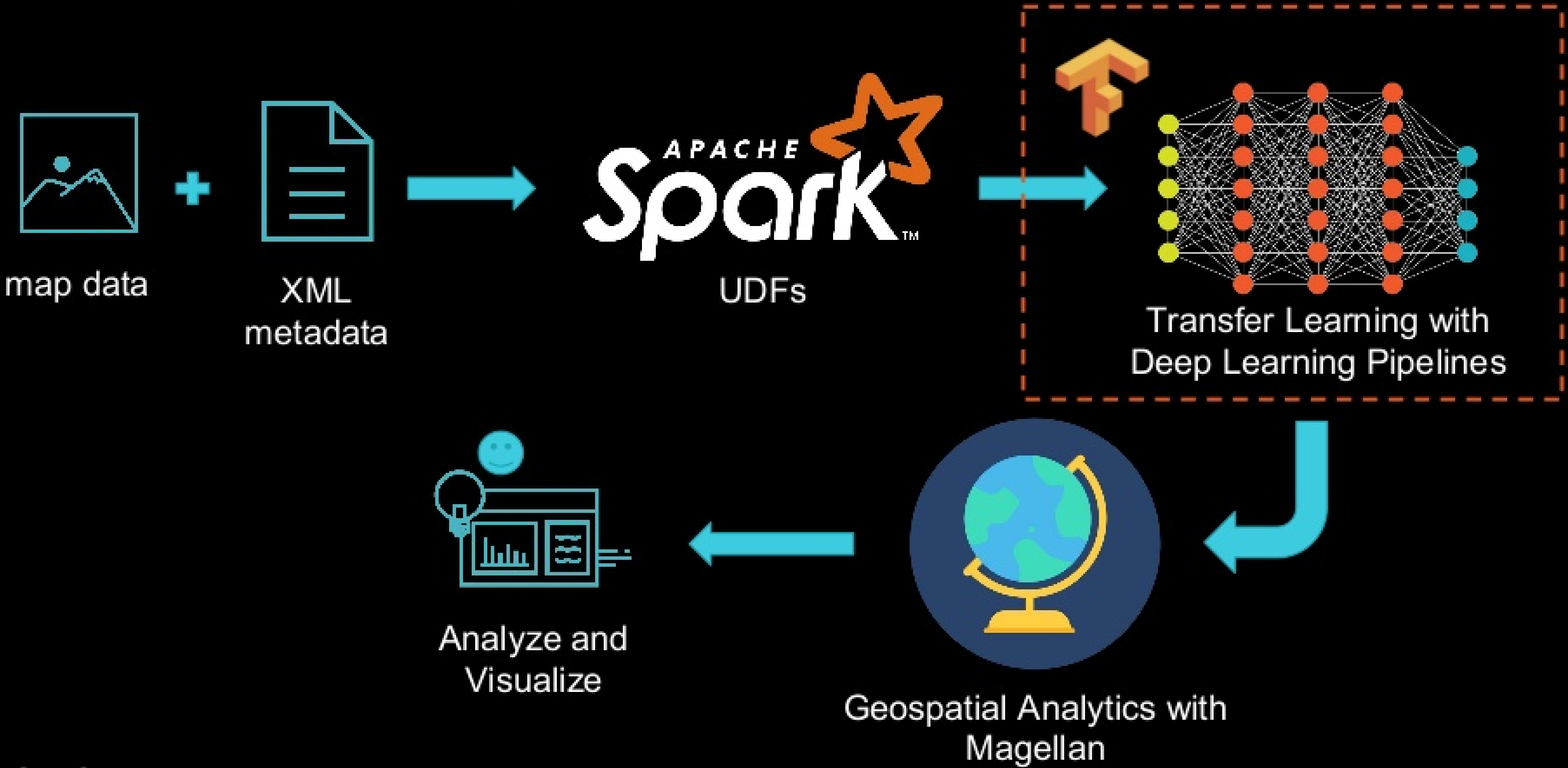
```
images = spark.readImages(img_dir,
  recursive = True,
  sampleRatio = 0.1)
```

# Image Transformations  with Spark

- Spark Joins
  - combine images with XML metadata
- Spark UDFs
  - Eastings and Northings → Latitudes and Longitudes
  - Creating Image chips and respective coordinates


(lat, long)
(lat, long)

# Deep Learning



map data

XML metadata

UDFs

Transfer Learning with Deep Learning Pipelines

Geospatial Analytics with Magellan

Analyze and Visualize

# Success of Deep Learning

Tremendous success of image-based applications

*Increased availability of pre-trained models*

Quickly building domain-specific models using transfer learning

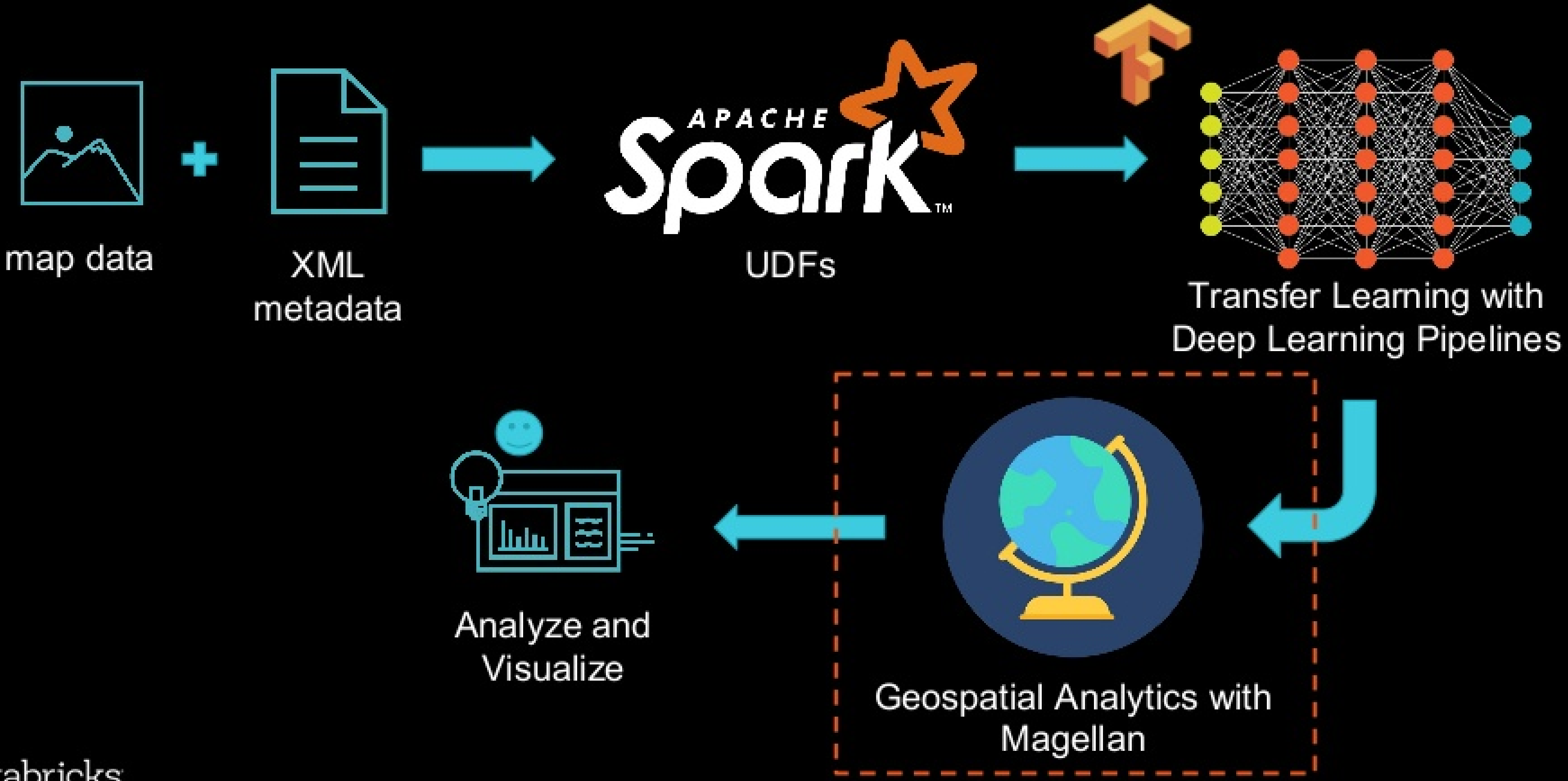
# Existing frameworks

- Mostly Python
- Google's TensorFlow is the most popular (easy to install/use)
- PyTorch popular in research
- Others: MXNet, Theano, Caffe, Keras, DeepLearning4J (java)

# Spark Deep Learning Pipelines: Deep Learning with Simplicity

- Open-source Databricks library
- Focuses on *ease of use* and *integration*
  - without sacrificing *performance*
- Primary language: Python
- Includes APIs to transform images

# Geospatial Analytics with Magellan



map data

XML metadata

UDFs

Transfer Learning with Deep Learning Pipelines

Geospatial Analytics with Magellan

Analyze and Visualize

# Common geospatial tasks

- Find all objects within an area
- Build geometries
- Cluster and aggregate similar objects
- Infer geometries (roads, buildings, etc.)

# Magellan

- Open-source library for geospatial analytics with Spark

- Understands various formats (geojson, …)

- Performs basic geometric operations at scale (polygon intersection, joining, … )

- Integrates into Spark SQL engine and builds indices for high performance

# Demo

# Recap

1) Read images with Spark
2) Parse image data with OpenCV and Spark UDFs
    a) Slice images into smaller image chips
    b) Generate respective coordinates for image chips
3) Pass data into a pre-trained tensorflow model and extract predictions with Spark Deep Learning Pipelines
    a) Model was trained on the xView dataset
    b) Model classifies objects identified in images
4) Visualize identified vehicles on a heatmap
5) Cross-check with Magellan

# Thank you

databricks