MACHINE LEARNING UTILITY MANIFOLDS FOR NOVEL DATASETS

Adam Van Etten

avanetten@iqt.org IQT Labs August, 2020

1. EXECUTIVE SUMMARY

Data is the fundamental currency of machine learning, yet input data for machine learning projects is often treated as a nearly immutable resource. Most parties (such as academic researchers or technology startups) are not highly incentivized to spend significant effort studying the many nuances of datasets, and how those nuances inform and impact machine learning projects. On one end of the research/deployment spectrum, academic researchers are heavily incentivized to focus on novel algorithms even when added complexity may not bring an appreciable increase in performance [1, 2, 3, 4, 5]. On the other end of the spectrum, corporations and government agencies are highly focused on deploying maximally performant solutions to existing problems. There remains much to be done towards the center of the spectrum, in the underserved domain of applied research focused on the interplay of machine learning algorithm performance with dataset quality, quantity, complexity, provenance, and veracity.

Applied research organizations (such as IQT Labs) that have the ability to operate in the open and the luxury of focusing on applied research (rather than commercial product creation or academic grant acquisition) have a unique opportunity to address this gap. Conducting applied research on the interplay of machine learning and dataset facets informs a number of strategic questions, such as what type of sensor hardware is required for data collection, or how much effort is required to collect and validate novel datasets.

In the sections below, we discuss in further detail the motivation and potential impact of applied research focused on machine learning data requirements and novel datasets. Section 2 details the Satellite Utility Manifold, which provides a case study for manifold research projects by exploring remote sensing geospatial analytics performance along multiple dimensions. Section 3 discusses the many connections that this concept has to other domains, and which axes (*e.g.* data resolution) are universally important. Appendix A details the value of curating novel datasets, while Appendix B describes the SpaceNet dataset and challenge series that was a key enabler of the manifold studies detailed here.

2. THE SATELLITE UTILITY MANIFOLD

The utility of any dataset is dependent on a multitude of variables. Over the last few years CosmiQ Works (the geospatial analytics team of IQT Labs) has systematically worked to quantify the utility of satellite imagery datasets, a concept referred to as the Satellite Utility Manifold. The surface of the utility manifold has important tactical and strategic ramifications. For example one could compare the tradeoffs of sensor resolution, collection frequency, and cost (see Figure 1). By quantifying the utility surface and confidence intervals for a variety of axes, the CosmiQ Works team has sought to inform a number of important questions in the geospatial analytics domain such as: sensor quality, sensor resolution, dataset size, and resource requirements. It should be noted that the majority of the studies detailed below were only possible due to two large, high quality datasets (SpaceNet and Rareplanes) that were developed by CosmiQ Works to fill existing voids, and subsequently open-sourced.

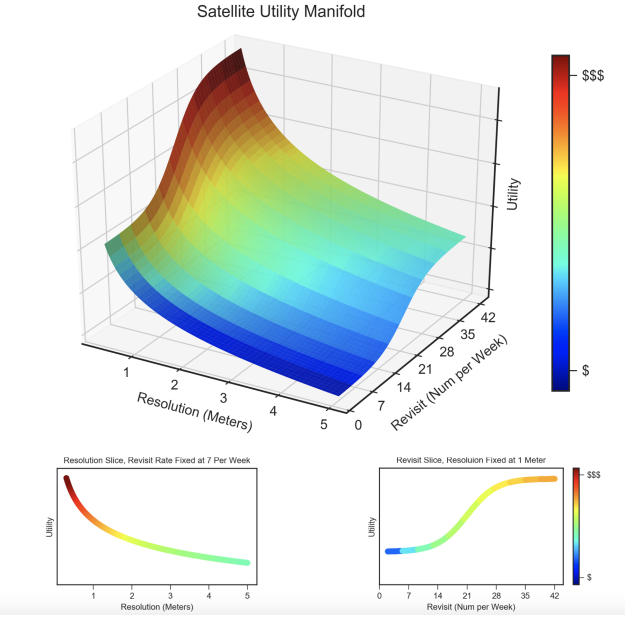

Fig. 1: Satellite Utility Manifold Example. Top: Notional utility manifold as a function of imaging resolution and revisit rate. Both utility and cost increase with constellation revisit rate and imaging resolution. **Bottom:** Cross-sectional slices of the utility manifold along the resolution (left) and revisit (right) axes.

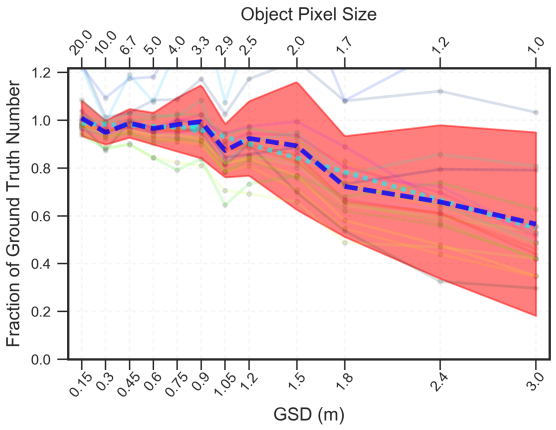

Fig. 2: Object detection performance (object enumeration) as a function of sensor resolution (from Figure 8 here).

2.1. Sensor Resolution

One method for measuring the utility of satellite imaging constellations is object detection performance. Back in early 2017, CosmiQ Works' first formal foray into exploring the utility manifold quantified the effects of image resolution on vehicle object detection, with the goal of providing a cross-section of the manifold and informing tradeoffs in hardware design. This study demonstrated that for the selected dataset, detection performance was extremely high for objects ≥ 5 pixels in extent.

This early work holds up well against more recent (July 2020) work conducted by industry: compare the object enumeration performance of Figure 2 with the final figure of this Maxar blog. At 60 cm resolution the recent Maxar analysis records a recall of only 0.03, whereas our analysis in Figure 2 has $35\times$ better performance, with a recall of 0.97. The salient point here is that while manifold studies are starting to interest industry, clearly much work remains to be done. Further information is available in the arXiv paper, and blogs on The DownLinQ [6, 7, 8].

Expanding upon the initial work on resolution, CosmiQ undertook a detailed study on the application of super-resolution techniques to satellite imagery, and the effects of these techniques on object detection algorithm performance applied to terrestrial and marine vehicles. Using multiple resolutions and super-resolution methodologies, this work showed that super-resolution is especially effective at the highest resolutions, with up to a 50% improvement in detection scores. Further information is available in the CVPR EarthVision paper, as well as a series of blogs [9, 10, 11].

2.2. Imaging Band Cardinality

Along with resolution, the number of imaging bands is an important aspect in the design of imaging sensors. CosmiQ Works undertook a study to quantify how object detection performance varies between grayscale, standard RGB, and multispectral imagery [12], and showed that while RGB provides a boost over grayscale, multispectral data often provides diminishing returns. Another study sought to "multi-spectralize" lower cardinality imagery to add imaging bands as a means of studying whether detailed spectral information could be extrapolated from simple sensors [13, 14]. These studies help inform sensor design, as well as providing a baseline for algorithm comparison among different data types.

2.3. Limited Training Data

In most machine learning applications, training data is a precious resource. With this motivation, CosmiQ Works undertook a Robustness Study to determine how training dataset size affects model performance in the geospatial domain, specifically: the task of finding building footprint polygons in satellite imagery. The study indicated that model performance initially rises rapidly as training data is increased, with diminishing returns as the amount of training data is increased further. In fact, compared to using the full data set, using just 3% of the data still provides 2/3 of the performance (see Figure 3). Irregardless of domain, a better understanding of the relationship between dataset size and predictive performance has the potential to help guide decision-making surrounding data collection and analysis approaches. Further details are available as a booklet, and a series of blogs [15, 16, 17, 18].

Currently, the standard approach for training deep neural network models is to use pre-trained weights as a starting point in order to improve performance and decrease training time, an approach called transfer learning. Given its ubiquity, quantifying the boost provided by transfer learning is therefore of great importance. A transfer learning study undertaken by CosmiQ Works showed that while pre-trained weights yielded abysmal results when applied to a new testing corpus, transfer learning using these weights allowed the model to rapidly (*i.e.* in the span of ~ 5 minutes) train on the new domain, yielding marked improvements in performance; such findings not only quantify the utility of transfer learning,

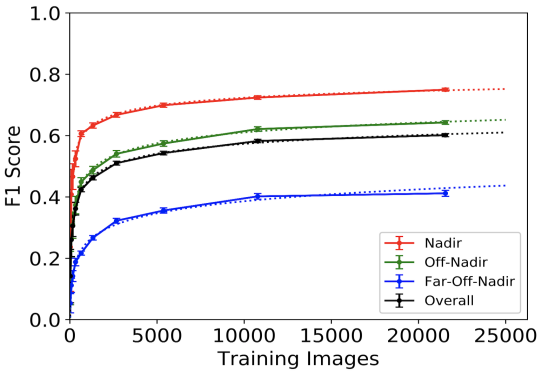

Fig. 3: Foundational mapping performance as a function of training size over multiple data collection paradigms (from Figure 2.1 here).

but the amount of data required to adapt pre-trained weights to new environments.

The Rareplanes synthetic data study looked at another important computer vision problem: how much data is required to detect rare objects (airplanes, in this case) in a large dataset. See Section 2.5 for further discussion.

2.4. Data Diversity

While understanding the amount of training data required is critical for machine learning projects, dataset diversity is another crucial factor. Geographic diversity is an important aspect in many datasets, particularly the geospatial domain. CosmiQ Works investigated data diversity as part of the aforementioned building footprint robustness study, with recommendations on whether to increase diversity or targeted data depending on the scenario. CosmiQ Works also investigated the geographic diversity required for road extraction from overhead imagery, quantifying the surprisingly small performance delta in unseen areas [19].

2.5. Synthetic Data

Rareplanes is a machine learning dataset and research study that examines the value of synthetic data to aid computer vision algorithms in their ability to automatically detect aircraft and their attributes in satellite imagery. CosmiQ curated a large labeled dataset of satellite imagery (real data), to go along with an accompanying synthetic dataset generated by an industry partner (AI Reverie). Along with the dataset, the Rareplanes project provided insight into a number of axes, such as the performance tradeoffs of computer vision algorithms for identification of rare aircraft that are infrequently observed in satellite imagery using blends of synthetic and real training data (see Figure 4). The Rareplanes model of developing a novel dataset in conjunction with addressing fundamental questions about the utility of the new dataset is a paradigm that merits replication. Further details are available in the arXiv paper and blog series [20, 21, 22, 23].

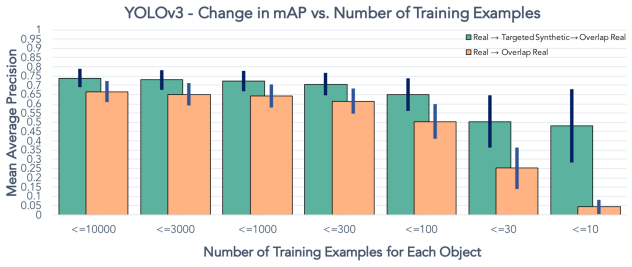

Fig. 4: Effects of synthetic data for various training dataset sizes in the Rareplanes dataset. The green (synthetic augmented) models clearly outperform the orange baseline (real data only) models for rare objects (from Figure 5 here).

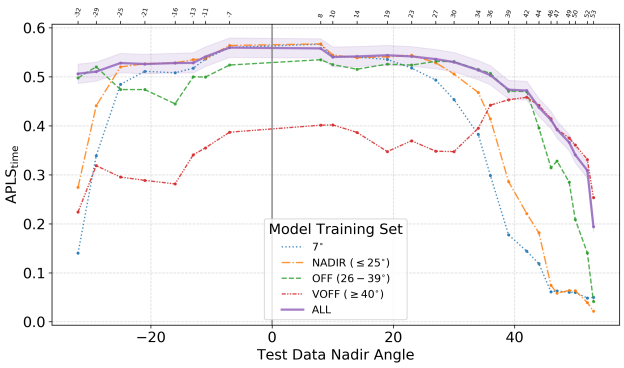

Fig. 5: Road network extraction performance as a function observation angle (from Figure 5 here).

2.6. Observation Angle

In many scenarios where timeliness is key, remote sensing imagery cannot be taken directly overhead (nadir), necessitating data collection from an off-nadir perspective. The vast majority of remote sensing datasets and models are solely nadir, however, leaving a significant gap in understanding as to how state-of-the-art algorithms perform in non-ideal scenarios such as high off-nadir imagery. To address this question CosmiQ (along with its SpaceNet partners) launched the SpaceNet 4 dataset and challenge (see the ICCV paper for further details). The dataset consisted of multiple collections of the same location (Atlanta, Georgia) as a satellite passed overhead, yielding 27 different nadir angles in the dataset. This allowed the CosmiQ Works team (and SpaceNet 4 competitors) to quantify the drop in building footprint detection performance as the "quality" of imagery degrades as imagery becomes more and more skewed.

Subsequently, the CosmiQ Works team performed a similar analysis of the Atlanta dataset, this time extracting road networks from highly off-nadir imagery, demonstrating the ability to identify road networks and features at off-nadir angles (see Figure 5). Somewhat surprisingly, road networks appear easier to extract at high off nadir angles than buildings, see the arXiv paper for further details. The performance benchmarks established by both the buildings and roads studies have the potential to inform collection management procedures, as well as satellite constellation design.

2.7. Label Quality

High quality labels are critical for machine learning, even if quantifying the effects of low-quality labels can be a challenge. This is exactly what the CosmiQ Works team set out to do in comparing road network extraction performance for models trained on crowd-sourced OpenStreetMap labels, versus highly curated SpaceNet labels ($\geq$ 60% improvement). See the WACV paper for greater detail.

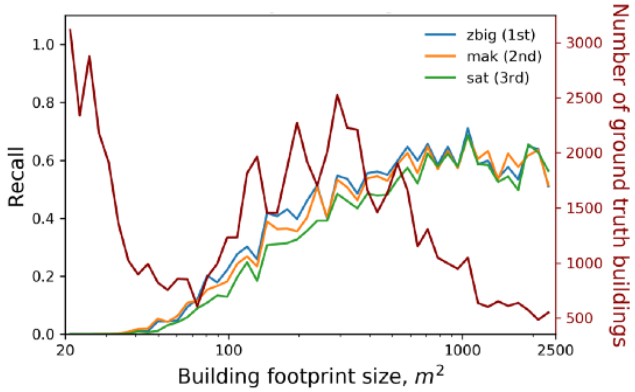

Fig. 6: Detection performance as a function of building footprint area for the top SpaceNet 6 models (from Figure 2 here).

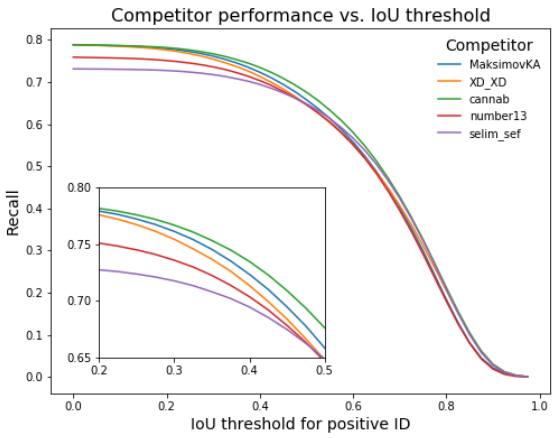

Fig. 7: Detection performance as metric the IOU metric parameter is varied (from Figure 2 here).

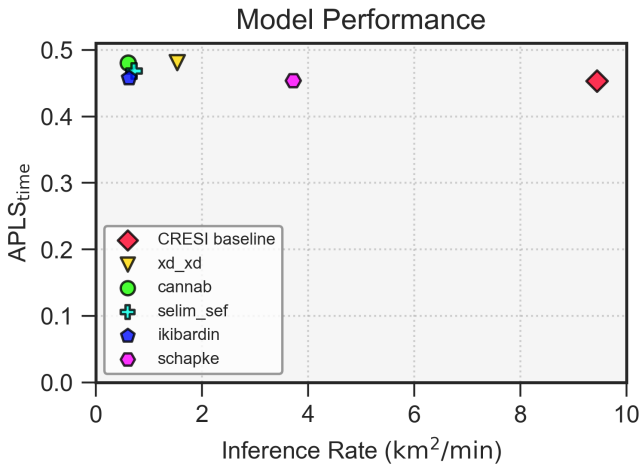

Fig. 8: Speed/performance tradeoff analysis for various road detection models (from here).

2.8. Object Properties

Breaking down performance according to the properties of objects within the dataset is another axis worth exploring. One example of this is exploring road network extraction performance based upon features such as road length, maximum speed, and intersection density [24]. Building identification performance as a function of building area and volume is also quite informative, and was explored in both SpaceNet 4 [25], and SpaceNet 6 [26], see Figure 6.

2.9. Metric Parameters

Determining the correct measure of utility is critical for gleaning meaningful insights into machine learning manifolds. Even simple measures such as the true positive rate are often subject to multiple parameters, as how one defines "true positive" often varies depending on the required prediction fidelity. Accordingly, CosmiQ explored how metric parameters influence metric scores (*i.e.* utility) in the geospatial domain. For example, the road extraction APLS curve can be quantified as a function of allowed buffer [27]. Figure 7 displays object detection scores for the SpaceNet 4 dataset as building outline IoU is increased; curves such as these help inform expectations and requirements for end users.

2.10. Speed Performance Tradeoffs

The final piece of the utility manifold we will discuss is the tradeoff of algorithm performance and runtime. Frequently, the state-of-the-art in machine learning is advanced by adding layers of complexity to existing models, thereby netting a $2-5\%$ improvement in the metric of choice (and an academic paper), but at the expense of increased runtime. Such "advances" sometimes turn out to be detrimental for real-world use cases due to slower speeds and increased fragility. These tradeoffs were analyzed for the SpaceNet 5 road network and travel time challenge [28[(see Figure 8), as well as the SpaceNet 6 synthetic aperture radar (SAR) building extraction challenge [29]. These analyses allow potential users of the open source code provided by these efforts to benchmark performance and to determine the appropriate algorithmic approach based on their performance requirements and computational environment.

3. RECOMMENDATIONS

Many of the research questions and lessons learned from CosmiQ Works' geospatial analytics discussed above translate readily to new domains. For example, one might ask: what is the label fidelity required for natural language or audio data (such as the VOiCES dataset)? Or, can poorly labeled cyber datasets be successfully augmented or pruned to improve dataset quality and prediction confidence, and if so what dataset axes matter most? Or, how does edge computing device sensor performance very with resolution, frequency, and weight? Irrespective of domain, myriad questions can be asked once a suitable measure of utility is decided upon, and quite often the least interesting (though most commonly pursued) research topic is building a machine learning model that maximizes utility without addressing the underlying feature space that determines performance. Studying the performance curve along various axes yields far more insights than just a single datapoint denoting maximum performance. Combining multiple facets together also permits quantification of the complex multi-dimensional utility manifold. This concept is illustrated in Figure 9. Some axes are universally important such as data resolution, label resolution, and dataset size. Measures of utility will of course vary across domains, but it is possible that manifolds may hold predictive power across domains.

Creating new datasets is often a critical piece of manifold studies, as discussed further in the Appendices. For context, precisely zero of the examples above had the luxury of analyzing existing datasets; in all cases answering the pertinent question necessitated the curation of a new dataset with the requisite features.

This freedom from hyperfocus on incremental algorithmic improvement (*i.e.* academia), or software development in support of marketable products (*i.e.* industry) allows applied research organizations (*e.g.* IQT Labs) to focus on underserved research areas. This document posits that such organizations will have maximal impact by rigorously exploring the feature space that determines machine learning performance on high quality datasets. Accordingly, projects should be structured around dataset creation and the subsequent study of the machine learning utility manifold. Quantifying the extent to which machine learning performance depends upon and influences dataset size, fidelity, quality, veracity, and provenance has the potential to positively impact the complex missions of a multitude of customers at both the tactical and strategic levels.

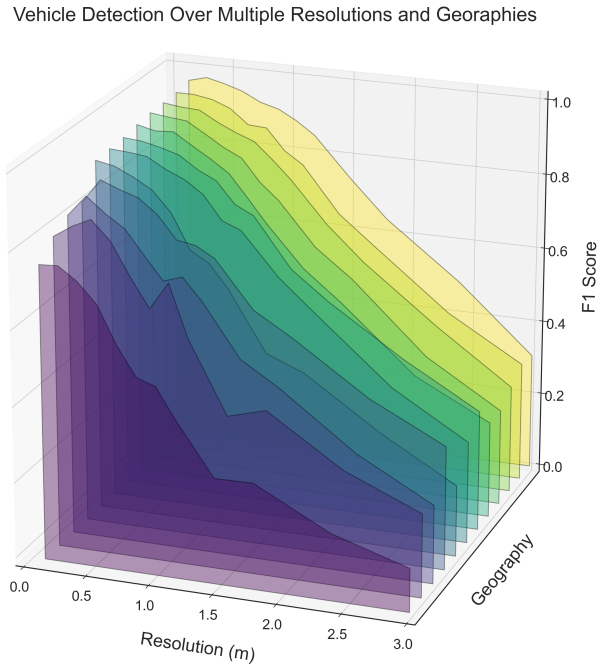

Fig. 9: Computing vehicle detection performance over multiple resolutions and geographies illustrates the utility manifold surface for these parameters (adapted from here).

Acknowledgements

The work summarized here stems from research conducted by Ryan Lewis, Jake Shermeyer, Daniel Hogan, Nick Weir, Dave Lindenbaum, Lee Cohn, and the author.

Appendix A: Dataset Creation

The majority of recent AI advances have been in the realm of supervised machine learning, meaning that high quality *labeled* datasets are necessary. While gathering raw data is often relatively simple, ensuring that this data is of high quality can be a significant challenge. Furthermore, assigning labels to the raw data is often a very time consuming and expensive exercise. Yet despite the challenges, dataset creation remains something of a low hanging fruit in the machine learning workflow for a number of reasons.

First of all, relatively few open source datasets exist that are appropriately labeled and structured for machine learning. Many commercial entities understandably regard labeled data as a resource to be zealously protected, as a means to give their researchers an advantage in the quest to create ever more performant algorithms. Yet even entities with large corpora of proprietary data find significant value in open datasets. A canonical example of this is the ImageNet dataset and competition, which over the course of many years posed a number of computer vision challenges to the research community. The staggering amount of resources entities like Google and Microsoft poured into competing in ImageNet underscores the utility provided by open source datasets. Data collection in small commercial markets (*e.g.* sparsely populated geographies) is sporadic at best in both academia and industry, even though such markets may be of great interest to many parties.

Secondly, new datasets have the potential to launch entirely new avenues of research. This provides applied research labs the opportunity to help spur research in areas currently underserved by existing academic or industry projects. An example of this is the increasing quantity of academic research devoted to road network extraction from satellite imagery, spurred on in part by the SpaceNet 3 and SpaceNet 5 datasets and public data science challenges.

Third, open source curated datasets are necessary for product comparison and evaluation. Evaluation of new products, both from industry and our customer base, is often done on proprietary datasets, thereby preventing meaningful comparison to alternative or competitory solutions. Quantitative comparison of algorithmic performance requires evaluating on a shared common test set. Establishing such gold standard test sets serves not only the research community, but peripheral institutions as well (*e.g.* strategic investors seeking to assist and evaluate technology startups).

Finally, datasets have the potential for outsized outreach, marketing, and publication impact. Properly curated datasets often have a shelf life far longer than algorithms. While academic publishing is not the foremost priority of many organizations, it is nevertheless often important from a reputation standpoint; the potential for citations within academia, industry, and government tends to be higher for dataset papers than algorithm papers given all the subsequent work that uses said dataset.

It is important to note that research labs will likely need to be heavily involved in any dataset project in order to ensure quality, even if data collection and labeling is contracted out. Labeling schemas and tolerances are not easily altered after the fact, necessitating thoughtful input and feedback from research personnel prior to and during data collection and labeling efforts. While a high quality dataset has the potential for significant impact given all reasons discussed above, creation/curation of a low quality dataset would actually prove counterproductive. For further specifics, see Appendix B for a dataset creation case study.

Appendix B: SpaceNet Dataset and Challenge Series

SpaceNet is a nonprofit LLC managed by CosmiQ Works dedicated to accelerating open source, artificial intelligence applied research for geospatial applications, specifically foundational mapping. Over the course of 7 challenges from 2016-2020, the SpaceNet partnership has released open source permissively licensed satellite imagery and labels over 100+ cities across all six inhabited continents. This imagery comprises both electro-optical (EO) imagery, as well as synthetic aperture radar (SAR) returns. The dataset encompasses over 10,000,000 building footprint labels and over 20,000 km of road labels, and has been downloaded over half a billion times worldwide. Imagery formats and locations are determined by the CosmiQ Team, thus ensuring data collection aligns with the desired goals of each challenge. Data labels are rigorously validated by multiple expert parties, with the labeling schema defined and verified by the CosmiQ Works team; heavy involvement in the data collection and labeling efforts has proven essential to establishing SpaceNet as a gold standard dataset. To push the state of the art in geospatial analytics, 6 public data science challenges have been run with the ever increasing dataset, as summarized in Figure 10. The winning algorithms are open sourced (33 in total as of August 2020), with $300,000 USD in prize money distributed. Scores of academic articles have cited SpaceNet publications, and over 30 publications have used SpaceNet data in their research.

The open source dataset and winning algorithms have been ingested by multiple organizations, and evaluation metrics implemented for the SpaceNet challenges have become the de-facto standard for multiple applications. The seventh SpaceNet challenge will incorporate the temporal dimension into the challenge for the first time, and has been accepted into the competition track for the prestigious NeurIPS 2020 conference.

More information is available at spacenet.ai, the SpaceNet academic papers (SpaceNet 1-3 (arXiv), SpaceNet 4 (ICCV), SpaceNet 5 (WACV), SpaceNet 6 (CVPR EarthVision), as well as a multitude of blogs on The DownLinQ.

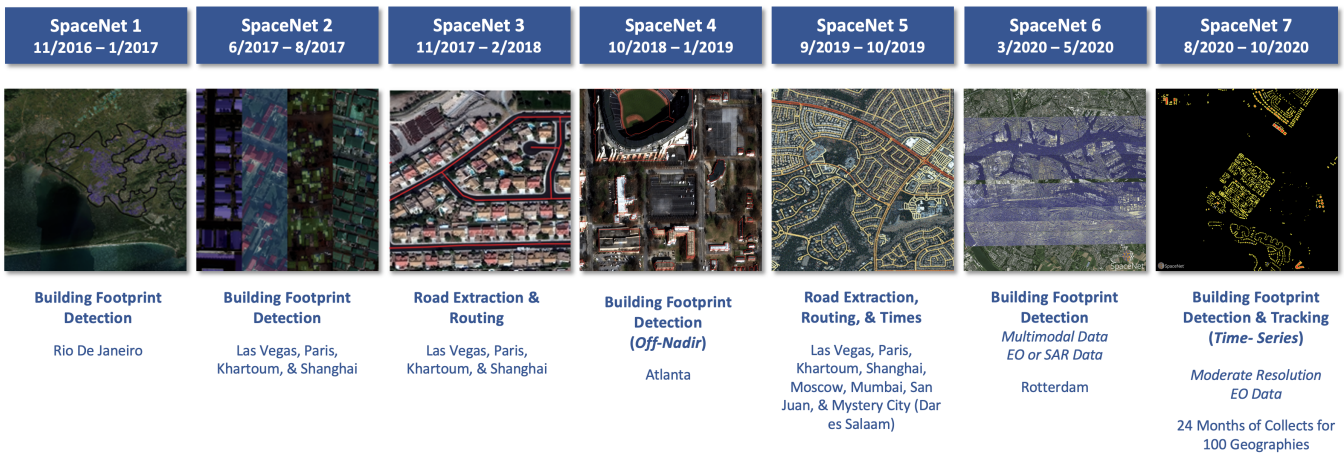

Fig. 10: SpaceNet Challenge series details.