The Alan Turing Institute

Data Study Group 9 – 13 December 2019

Challenge descriptions

Contents

Agile Datum

Automating the evaluation of local government planning applications

Overview

There are 3,500,000 planning applications each year in the UK, and yet 1,200,000 are rejected as the whole process is overly complex for citizens and involves too much manual admin for the councils.

The UK requires some 600,000 new homes to be built each year, to keep up with population growth, and yet in 2016/2017 & 2017/2018 only 300,000 new homes were built, and in 2018/19 this has further declined to only 176,000. There are 400 councils in the UK and some 117 are missing their housebuilding targets by upwards of 65%. This lack of new homes results in 30% higher mortgages, 30% higher rents, 30% less social housing and wider deprivation.

From building a whole estate, building a single house, to extending a house or fitting new windows on a listed building, to chopping down a tree, each project requires the submission of complex planning forms and technical drawings. Currently, each submission takes from 30-60 minutes just for a planning officer to check that all the information in the forms and drawings has been provided and is correct. This manual validation currently accounts for some 250,000 man days of admin per year.

On average it takes three weeks for a council to start looking at a planning application, as the councils are under resourced (budgets across local government have been cut by 50% over the last 10 years and yet demand for services from citizens has increased), and so the councils find themselves with large backlogs at this initial validation stage. This also leaves citizens in the dark as to where they are in the process, leading to additional calls and emails to chase progress which increases the workload on planning officers even further.

The Challenge

The aim of this data study group is to investigate how far we can go with using A.I. and Machine Learning to validate the content of forms and drawings in submitted planning applications. Can we use machine learning to speed up the review process both for the planning officers and for citizens, for example by automating the detection of certain common errors?

Our research with 20 of the 400 councils shows that 12 common errors (6 in the forms and 6 in the drawings) account for over 80% of rejected applications:

Common Errors in Form Submissions

- Has the correct form been used?
- Has the correct planning application fee been calculated?
- Have all sections been completed?
- Does the description of work match the drawings?
- Have all supporting files been uploaded?
- Are the crowns of trees shown in meters?

Common Errors in Planning Drawing Submissions

- Are floor plans provided and correct?
- Have existing & proposed elevations been provided?
- Are all scales provided and correct?
- Is True North provided and correct?
- Are Block/Site Plans provided and correct?
- Do the planning drawings match planning form descriptions?

The challenge is to not only identify the above 12 objects/errors in the planning application, but to be able to do this in context (clustered pairs). For example, a planning drawing may have multiple items (floor plans, elevations or side perspectives, site maps, scales etc.) and it could have multiple copies (e.g. existing floor plans and proposed floor plans). Therefore, there could be 3 sets of drawings with 3 associated sets of scales marked. We need to be able to identify which scale is associated with which drawing and then verify that the scale is correct.

The Data

Agile Datum is providing the Data Study Group with a library of the approx. 3500 planning applications submitted to the London Borough of Redbridge in 2019. Each application contains on average 10 supporting documents, including planning forms and planning drawings, in PDF format (which is the format councils store these files in document management systems). The whole dataset therefore contains approx. 36,000 files (16.5 GB).

We are also providing access to planning databases with data on both correct and incorrect planning forms and drawings to aid with training models.

Goals

1. Detecting and extracting relevant objects in application documents.

2. Validating whether the extracted information is correct.

The group can choose to explore the feasibility of developing a model to detect any of the common errors in applications described above depending on their background and interests. In all cases this will involve:

Each task comes with associated challenges. For example, text extracted from forms should retain key-value pairs (e.g. this block of text is the address, this is a description of work etc.), and models may have to deal with multiple form layouts. Or for drawings there is no standard symbol for true north, so a model will need to be trained to identify a large range of possible true north symbols. Validating that information is correct will generally require linking information on the forms with the drawings – for example, does the boundary marked on a site plan map match the address written on the form?

What We're Looking for

We welcome participants from all backgrounds as there is a wide variety in the challenge questions and the types of techniques and domain knowledge that could be used to solve them. However, prior experience with the following would be particularly useful:

- Computer vision (e.g. image classification, object detection, optical character recognition).
- Natural language processing.
- Supervised and semi-supervised learning (for classification).
- Unstructured data.
- Data maching/record linkage.

Dstl

Anthrax and nerve agent detector identification of hazardous chemical and biological contamination on surfaces using spectral signatures

Overview

The assessment of surfaces for potential contamination by biological (e.g. anthrax pathogen Bacillus anthracis) and chemical (e.g. nerve agents such as VX) hazards is relevant for a range of military and civilian applications. To this end, Dstl and the Defence and Security Accelerator (DASA) are providing a dataset collected using a range of different sensor modalities that have measured various surfaces contaminated with surrogate bacteria, hazardous chemicals and relevant control materials. Both un-mixing and identification of the contaminant contribution from that of the underlying surface is non-trivial.

Participants are invited to explore how data science and machine learning techniques can be applied to recognise and discriminate between the various contaminants based on data from individual sensors or fusion of multiple data sources, and how models can be applied to characterise contamination on new surfaces without re-training.

About the Challenge Owner

The Defence Science and Technology Laboratory (Dstl) is the UK's leading Government agency in applying Science and Technology (S&T) to the defence and security of the UK. As part of the Ministry of Defence (MOD), Dstl responds to the S&T direction set by the UK Government's National Security Strategy and Strategic Defence and Security Review. Dstl brings together the defence and security S&T community, including industry, academia, wider Government and international partners, to provide sensitive and specialist S&T services to MOD and wider Government.

Background

Current methods to detect, locate and report hazardous biological and chemical materials incur operative, logistic and temporal burdens when various factors (which may include sampling, removal of the sample to laboratory infrastructure, sample processing, analysis) are taken into consideration. Additionally, many biodetection systems are large and require mains power, regular maintenance and a constant supply of consumables (for example, reagents) to operate. These systems are typically complex to use and are only operable by skilled end users. The time taken to analyse samples by such methods can reduce operational tempo.

In order to address these issues, Dstl and the Defence and Security Accelerator (DASA) conduct research into prototype systems with industrial and academic partners. These efforts have focused on the development of innovative sensor technologies that could ultimately lead to fieldable systems to provide rapid, high-confidence detection, location and identification of biological and chemical hazards deposited over a wide area. Successful detection technologies require the consideration, and ultimately the optimisation, of a range of parameters. These include speed of response and low false alarm rate in a range of environments (for example the system does not alarm to natural and anthropogenic background microbiomes). Some of these parameters may compete (such as speed of response versus limit of detection), plus the real-world deployment of these sensors further complicates this endeavor.

The Challenge

This challenge will address the application of machine learning to identify surface deposited bacterial species and chemical hazards from their spectral signatures, even in cases with significant contributions from the background surface. Hazardous chemical and biological materials pose an invisible threat and are technically challenging to detect, but doing so has the potential to save lives. olecular spectroscopy techniques provide information rich chemical signatures which have been shown in the literature to discriminate between different bacterial species and chemical hazards.

From a biological sensing perspective, we are currently co-ordinating a combined research effort (funded by UK Ministry of Defence) in a competition run by the Defence and Security Accelerator, DASA. our institutions, with underpinning research at Dstl, are developing and applying different spectroscopic technologies to generate data from a standardised sample set produced by Dstl. These sample sets comprise four different surfaces and two bacterial species (closely related to each other and the anthrax pathogen Bacillus anthracis). Likewise, spectral signatures of a range of deposited chemical hazards have been collected through assessment of prototype systems at Dstl. This sample set comprises three different surfaces, with three deposited hazards across two spectroscopic techniques. This will generate a unique multi-technology dataset from a standardised sample set. The challenge is the combined analysis of a large dataset from multiple technologies and will be an opportunity to investigate the specific sub-challenges of data fusion and technique agnostic analysis. This presents a rare opportunity for researchers to access chemical analysis data which would require specialist lab facilities and equipment to generate.

The Data

We will provide a data set of surrogate biological hazards that have been generated under a DASA competition involving multiple institutions with additional underpinning Dstl lab work. This endeavour is based upon different optical techniques to generate spectra, which are information rich and can be used to discern the chemical constituents within a sample. Each institution has generated spectra from a test set of bacterial samples generated by Dstl under a standardized protocol to ensure consistency. The test set consisted of spore samples of two bacterial species (B. atrophaeus & B. thuringiensis) which are close relatives of B. anthracis (the causative agent of anthrax), loaded onto four different surfaces at three different concentrations. Controls included blank surfaces, clean media, spent media and 1 μm diameter polystyrene beads. All permutations were repeated three times, and multiple spectral measurements generated for each repeat. Lab-grade reference spectra with a high signal-to-noise ratio are available for the two bacterial spore types independent of any background substrate and also from ideal substrates designed for spectroscopic analysis.

We will also provide a data set derived from surface deposited chemical hazards. This sample set consists of samples of three chemical hazards and one control chemical, deposited onto three different surfaces and measured by two spectroscopic technologies. All of the aforementioned data is available fully labeled within an accompanying metadata file.

Each spectrum will consist of intensities over a given range of wavenumbers (x-axis, equivalent to spectral frequency) note that spectra could have different x-axis limits and increments (ergo spectral resolution). The spectra will be provided with labels indicating the surface (up to   classes), bacteria chemical hazard present (Boolean), deposited material (bacterial, chemical hazard, controls), concentrations (classes), and spectroscopic technology (1   classes). The aim will be to develop classifiers to recognise the presence of bacteria or chemical hazard, and identify the type of bacteria or chemical hazard, despite other variables.

Goals

The overall aim is to explore how data science and machine learning techniques can be applied to recognise and discriminate between various contaminants, using data acquired from multiple technologies. The ultimate goals from this challenge can be summarised in the following questions:

1. Can the deposited bacteria and chemical hazards be robustly differentiated identified from
   - a. each other
   - b. control samples
   - c. various surfaces

2. Is the approach flexible enough to allow recognition of known bacteria and chemical hazards on surfaces that were not previously encountered during model development training   deally, performance should not be compromised using models that were trained with limited or noisy datasets.

3. Can we combine spectral data from more than one technology to enhance confidence in identification

What we re looking for

Potential approaches reported in the literature for the discrimination of biological/chemical hazards by spectroscopic techniques include multivariate analyses such as principal component analysis (PCA) and Linear Discriminants; regression analyses such as partial least squares (PLS); classification and regression trees (CARTs); ensemble methods; support vector machines (SVMs); clustering methods; neural networks; and rule based methods. Participants with experience in these techniques are therefore encouraged to take part, but we also suspect that successful solutions to this challenge may lie outside of what has already been reported in the literature, so also highly encourage participants who can suggest novel approaches to the challenge to also take part. Prior knowledge of spectral data is helpful, but not required, as scientists from Dstl will be available for discussions throughout the week.

Impact

If successful, this challenge has the potential to enhance current research by providing new analysis techniques, which will inform future laboratory research work planned in-house and through DASA in terms of types of data, number of samples and focus of research (e.g. exploring pre-processing methods which could enhance sensor performance. The development of data science and machine learning techniques capable of working across different instruments and technologies would be a significant step forwards in the wider spectroscopy field for many different applications (such as medical imaging). There is potential that work based on this challenge could therefore be reported in the scientific literature.

The basis of a flexible system that could rapidly and robustly analyse unknown samples and detect the presence of biological or chemical hazards in the field at potential stand-off distances, would ultimately limit casualties in the event of a release. Identification of contaminated areas avoids exposure to the hazard, but also allows rapid application of specific medical countermeasures to personnel at risk of exposure and supports decontamination efforts.

Dstl

Bright-field image segmentation

Overview

Confocal microscopy is a tool many bio-medical researchers use to gather data about their field of interest, including study of disease and infection. Automated analysis of these images typically starts with cellular segmentation (the process of identifying individual cells within images). This is routinely done using high contrast fluorescent labels for either the cytoplasm or plasma membrane. Segmentation using label free modalities such as transmitted light/bright-field microscopy is advantageous because it is less phototoxic than fluorescent imaging and removes the need for labels which may affect the function of the cells.

To the human eye cells are easy to identify on a transmitted light image, however due to the similarities in pixel values between the background and cell events segmentation by computational analysis is still a real challenge. This DSG challenge invites you to utilise the power of AI to design methods to segment cells from confocal microscopy datasets of human/murine immune cells infected with various pathogens.

About the Challenge Owner

Dominic Jenner has worked at Defence Science Technology Laboratory (Dstl) in biological defence science for over 15 years. He is the technical specialist for the cytometry equipment within the Chemical, Biological and Radiological Division (CBR Division). Dominic works within the microbiology and aerosol science group, working on projects that aim to develop medical countermeasures to toxicological and biological insults, along with projects that aim to define the immunological response to traumatic injury. Dominic is a chartered member of the Royal Society of Biology and in 2018 was awarded a 5 year Marylou Ingram scholarship with the International Society for Advancement of Cytometry for his work in the field of imaging flow cytometry.

Background

We use confocal microscopy to identify novel targets for anti-microbial therapies and to understand the effect of medical countermeasures. Confocal microscopy is a form of light microscopy which scans a laser point by point to build up an image. This is combined with a pinhole in the light path to remove out-of-focus light. We have a range of assays where cells are grown in-vitro (outside of the body) and imaged with a confocal microscope to study the effects of infection, or chemicals, on host cells such as macrophages. Macrophages are a type of white-blood-cell which fight infection by engulfing and digesting pathogens (bacteria, virus, fungi etc). We use fluorescent bacteria to infect the cells which allows us to see where the bacteria go within the cell and what they do to the cell during infection (see image below). Fluorescent labels have the special property that they absorb the excitation light, here a laser beam, and then emit light at a lower wavelength. By using a selective emission filter we can then produce high contrast images of the bacteria. However florescent labelling and imaging of the macrophages can dramatically change the behaviour of these sensitive cells and should be avoided if possible. Therefore to identify the location of the cells a transmitted light image is captured (see image below). This simply captures the light which passes through the sample and produces a low contrast image of the cells which is much harder to process than florescent data.

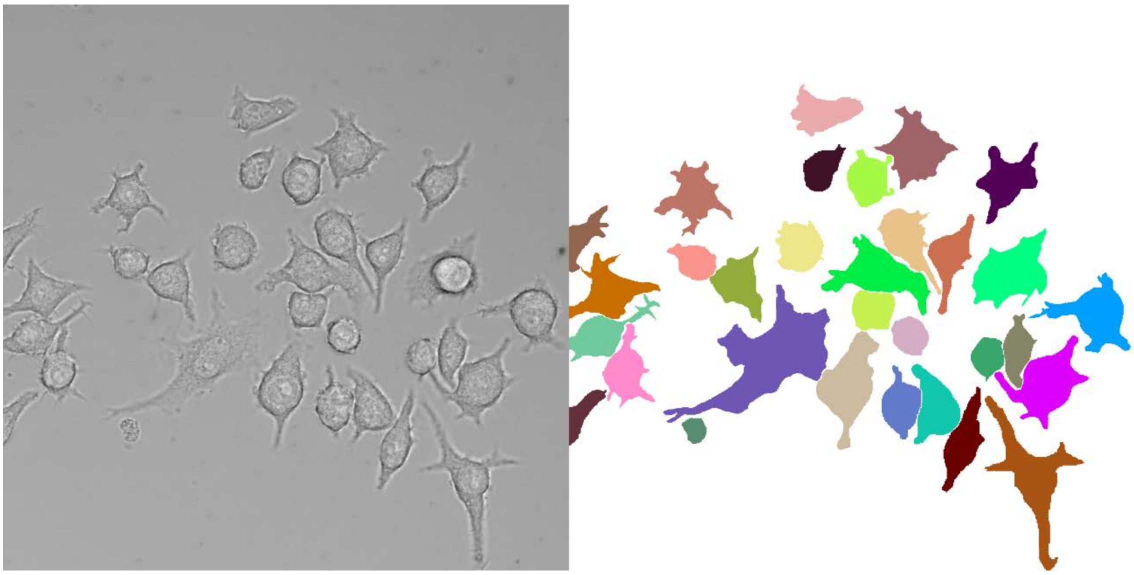
Challenge & Goals

The principle challenge is to develop segmentation protocols to identify individual macrophages from transmitted light, or bright-field images. This is a relatively simple task when cells are well separated but very difficult when they are densely packed together. Assuming a segmentation protocol is established we also have the following secondary goals:

- Count and detect the location of individual bacteria which have infected the macrophages. This will performed by correlating the florescent bacteria images with the cellular segmentations of the macrophages.
- Using time-lapse movies of macrophages quantify and model how infection spreads between cells and changes their behaviour.

The Data

The data being used in the study has been generated specifically for this challenge. The data are images of the RAW 264.7 macrophage cell line grown on 35mm tissue culture dishes, seeded at different densities and imaged with confocal microscopy (transmitted light). The data has been generated to take the challenge back to the very basics of segmentation; find and outline the cells within the images. There are approximately 370 images of RAW cells. Along with this there are a range of images taken from bacterial infection assays, where GFP bacteria are also present and imaged in a separate fluorescent channel. A subset of the data has been annotated manually to facilitate the use of machine learning approaches and to provide a ground-truth for performance evaluation (see, example below).



What we are looking for

Data computational scientists with some experience of working with images and a desire to apply their skills to answer bio-medical research questions. Experience of image processing analysis and or machine learning alongside strong programming skills is ideal.

Impact

If successful the developed segmentation strategy will be incorporated into DSTL s assays and used to further our understanding of infection. Additionally segmentation of cells from bright-field images is a common problem faced by bio-medical researchers in many domains. Therefore the results of this challenge will be interest to a much wider audience and could easily be developed into a larger pro ect, or publication. This pro ect is a really exciting opportunity to gain interdisciplinary experience and apply you skill set to a problem with clear real life applications.

The National Archives

Discovering topics and trends in the Government Web Archive

Background

The UK Government Web Archive (UKGWA) is a treasure trove of information and the history of government since 1996. Leveraging cutting edge data science, collaborate with us to understand what we have and how we can enable others to explore this open dataset. The challenge is to improve access, discoverability and comprehension of the UKGWA, to open the resource to new forms of research at scale, to individual focussed enquiry, and to understanding its contents macroscopically, through semantic approaches.

Users currently explore the UKGWA either through browsing, or by keyword searching. Research into users of the UKGWA indicates that they expect an "intuitive" search experience, which would enable them to navigate this massive dataset, with items of relevance to their interests presented more readily.

Many of the challenges we face with the UKGWA are similar to those of other digital records - scale, complexity, changes over time, archival description and resource discovery. External and internal search engines do not function optimally over the UKGWA due to the very particularity and complexity that make it the rich information source it is. This Data Study Group is an exciting opportunity to develop a more sophisticated, nuanced and assistive approach to search and discovery of this unique archive.

About the Challenge Owner

As the official archive of UK government, The National Archives has created the UK Government Web Archive (UKGWA), a vast archive of UK government websites and social media containing data from 1996 to the present. In contrast to other web archives, the UKGWA is a fully open and publicly accessible service that gives continued access to government records and information originally published online, as well as providing context for other digital and analogue records, such as those from public enquiries and government departments. The existing search service is effective at search across the full text of the more than 350 million documents that make up the UKGWA. However, this search relies on keyword matching and leveraging some domain knowledge, which most users do not have.

The Challenge

How can we open the UKGWA for interrogation and exploration by researchers and the general public? The challenge is to identify, or create, approaches to understanding the UKGWA as a whole, and how it changes over time, allowing researchers to explore the archive by themes, not just keywords. The UKGWA is a diachronic corpus covering 23 years. The dataset comprises select websites from different points in time over this period, and the Data Study Group will work to enable the discovery and exploration of emergence and decay of topics over time, across different domains and different government websites. The proposed solutions therefore need to cope with high levels of duplication and growth in the amount of information published on the web year on year.

Experimenting and developing methods of semantically enhanced search and discovery of the UKGWA will bring benefits to the global web archiving community, archivists and users alike, for restricted-access and open web archives. The methods explored during this Data Study Group will be applicable to filtering, sorting, and comprehending digital collections at scale across industries and sectors.

Domain-specific expertise from the team at The National Archives will be present during the week.

The Data

The data consists of three datasets. Each is a plain-text collection derived from HTML pages in the UKGWA:

- Dataset 1: TNA Government Hub Websites 2006–2019: pages from direct.gov.uk and www.gov.uk, throughout this period.

- Dataset 2: TNA Thematically Sampled Websites 2006–2019: pages from approximately 450 government websites, which have been pre-selected through high-level topic modelling.

Supplementary Data: Home pages 1996-2019 and Blogs 2013-2019: every home page captured since the beginning of the UKGWA, which was the dataset used to create Dataset 2; Shallow crawls of government blogs, although most blogs aren't crawled every month.

The websites these dataset were drawn from address all areas of life and activity touched by UK government work. They were selected to reflect a range of these activities identified through topic modelling as relating to the Olympics, healthy eating, regional development, climate change, and statistics and transparency.

The datasets were created by selecting snapshots taken of websites on the date closest to 1 January from 2006 onwards. In many cases this covers the entire lifespan of sites. The snapshots extracted through this were processed by a combination of shallow and deep web crawling, extracting text from the HTML. Each page of each site is stored in an individual text file, organised by site and snapshot.

All files in the datasets are plain text *.txt files, with the HTML tags removed. By far the largest domain is www.gov.uk which amounts to 1.6 million files across 8 snapshots. The site was only a Beta release in 2013 but the number of pages grew from 91000 at the start of 2014, to over 400,000 by the beginning of 2019. The estimated size of the combined datasets is 20GB in 2-3 million files. This represents approximately 10% of the sites on UKGWA.

The challenge with this data is not only the scale but the duplication: an individual page may appear in every snapshot over time, changing subtly, for example through added links or rephrased text, in ways that the process of selection cannot distinguish. The most commonly occurring links — for example to site navigation, accessibility help and contact details — have been removed to ensure the focus of the datasets remains on their primary contents.

Goals of the Data Study Group

The goal is to use these curated datasets of reference documents to build algorithms that are capable of identifying like documents across the corpus and inferring the likely topics they cover.

This work will contribute to generating an overview of the contents of the UKGWA which will be developed for inclusion in user-facing services, enhancing search and unlocking the potential of the UKGWA.

The main aim of the week is to give insight into (i) what approaches can be used to assist the understanding of the UKGWA "as data" and (ii) what are the most viable approaches to improve the resource discovery offer to users.

- What unsupervised machine learning can be used to generate document-level metadata of linguistic clusters within WA
- How can the resulting metadata be used to inform description of the nature of the information they contain and guide the interpretation of categories
- What methods can track the emergence and evolution of topics across time
- What approaches can be used to differentiate between the functions and aims of government departments as expressed in individual websites
- How do we best explain the data science methods used on the WA to its readers and users
- Thinking of machine learning algorithms that will be developed as part of the DS , what self-explaining workflows for algorithmic introspection can be developed to aid interpretation of their processes and encourage engagement with their strengths and limitations

What we are looking for

Obvious collaborators are those who are experienced in the fields of text analysis, topic modelling and resource discovery technologies, developing algorithms and methods that will address the goals outlined above. We would like to encourage new thinking and encourage those with cognate skills and experience to oin this challenge, from cutting-edge deep learning techniques to general methods for clustering and classifying heterogeneous data. t is expected that these novel approaches in con unction with other, existing techniques will provide a successful outcome for the challenge.

WWF

Smart monitoring for conservation areas

Overview

WWF monitors over 250,000 protected areas (e.g. National Parks and Nature Reserves), thousands of other sites and critical habitats (e.g. coral reefs and mangroves). These sites are the foundation of our global natural assets and are central to the preservation of biodiversity and human well-being. Unfortunately, they face increasing pressures from human development. With mines, oil and gas operations destroying and degrading habitats; dams altering river flows; agriculture causing habitat loss; road and rail expansion fragmenting and opening areas to further degradation.

With a vast array of threats emerging on a daily basis across 300,000+ sites, a growing challenge for WWF and the wider conservation movement has been to consistently and timely identify;

1. emerging or proposed human developments within key sites.
2. the stakeholders involved.

The timely provision of this actionable information is vital to enable the wider machinery of WWF and the conservation community to engage with governments, companies, shareholders, insurers, and others to help halt the degradation or destruction of key habitats. Earlier engagement is often critical in gaining a positive outcome for the environment before projects are well established and significant investment has already occurred.

About the Challenge Owner

WWF is one of the world's largest independent conservation organisations, active in nearly 100 countries. Our supporters – more than five million of them – are helping us to restore nature and to tackle the main causes of nature's decline, particularly the food system and climate change.

To help achieve this, the Conservation Intelligence (CI) team at WWF pools vast volumes of spatial data, running GIS assessments and applying remote sensing to define the extent of threats and degradation to scale the issues, inform the public and guide WWF's intervention/s. Wherever possible, the CI team attempts to identify the actors involved, that is, governments, companies, shareholders, insurers and others as this provides an actionable means to effectively engage to help limit negative impacts.

As part of their work, the CI team developed WWF-SIGHT. SIGHT is a global mapping platform based on ESRI software. Uniquely, it integrates both commercial and non-commercial spatial data, satellite imagery and various functionality. As a result, this data portfolio provides WWF with unparalleled insights into who is operating where, and with basic remote sensing functionality to check the sites in near real-time. WWF-SIGHT will be shown and access provided on request throughout the event.

If successful, relevant news stories will be geolocated and tagged within WWF-SIGHT to allow WWF staff to click and quickly consider relevant news stories in their areas of operation. WWF will also openly provide the results to support other conservation actors interested in tracking news stories.

The Challenge

Can we detect and geolocate public news stories describing emerging threats to key protected areas to better inform conservation interventions, public transparency and accountability?

The earlier we are aware of a threat (e.g. a proposed new dam or road) in the project lifecycle, the more likely the WWF and the wider community is to be able to mitigate negative social and environmental impacts. To get better at early identification of proposed developments in highly sensitive sites, we aim to build a monitoring system based on web scraping news stories. We start with a pilot study on the flagship protected areas: natural World Heritage sites (244 sites, e.g. Serengeti, Great Barrier Reef).

We aim to build a system to differentiate relevant news stories from the total of 50,000 scraped articles. We then plan to text mine these news stories to highlight key terms, including involved stakeholders, dates, location, etc. We will then geolocate, tag and rank relevant news stories to sites within WWF global mapping systems to help highlight emerging issues.

If the pilot is successful, we plan to scale it to a wider group of conservation assets (250,000+) and provide open access to the relevant geolocated news stories. This will, hopefully, support the wider conservation community and provide greater transparency on the stakeholders driving biodiversity loss.

The Data

We will supply approx. 50,000 JSON describing news stories relevant to UNESCO World Heritage Sites (WHS) scraped using Google News API over the past 2 years (January 2018 - October 2019).

If required, the API, codes and the list of keywords used for each WHS will also be available during the challenge. In addition, a training dataset of over 200 news articles will be provided to facilitate the news' relevance assessment in 3 classes. The study will only be considering news in English.

To support but not essential to the challenge, we will also provide the details of assets within World Heritages Sites. These values and data points may be useful in developing lookup tables, identifying ownership and integration across datasets;

- Power Plants - All types
- Mining Assets and Concessions
- Oil and Gas Assets and Concessions
- Protected Areas
- Admin Regions 1 - 5
- Oil Palm Concessions
- Global Ports
- Global Airports
- EBSA
- Basins – Major Rivers, Lakes, Rivers and Wetlands
- Reservoirs
- Major Rivers
- Hydroelectric Dams
- Dams and Future Dams

Goals

The overall aim would be to identify current and potential human pressures across all 244 natural World Heritage Sites using Google News API and different text-data mining techniques. We intend to explore news scraping and the use of different text-data mining and geolocation techniques to:

1. Assess the relevance of each article according to the classes explained in the data section (we will be only working with articles in English).
2. Extract new information to refine and better target interventions, and identify and tag new information (i.e. activities and key holders involved, refine areas of interest, etc).

This information will then be scored and integrated into WWF's global GIS mapping platform.

What we're looking for

We are seeking data scientists with experience in text-data mining techniques and natural language processing (NLP) methods (e.g. toponym resolution, georeferencing, text classification, deep learning and topic modelling).

This challenge cannot be solved purely through a quantitative approach. We therefore highly encourage participants who may not be familiar with the specified quantitative tools to take part.

SenSat

Semantic and Instance Segmentation of 3D Point Clouds

Overview

Autonomous vehicles require digital maps to avoid possible collision and navigate safely, smart cities requires knowledge of urban features to be managed appropriately, and digital twins require physical assets to be recognised before they can simulate predictive models. All those applications require a detailed representation and understanding of the spatial environment. SenSat captures high resolution images via drones with a ground sampling distance of ~2.5cm. Those images are then transformed into 3D point clouds, using techniques such as Structure from Motion (SfM).

The ability to create intelligent 3D models of the real world is a critical enabler for the reduction in cost and programme of major design activities. Highways England is one of the pioneers in developing parametric design solutions for complex national infrastructure. The understanding of point clouds and 3D shapes is critical for supporting the growing market of parametric design.

About the Challenge Owner

SenSat's vision is to make computer contextually understand the 3D digital worlds by simulating realities. The goal is to create digital twins, or digital representations of real world locations, then infuse real time data sets from a variety of sources. The long term vision is to build the third platform – an intelligent eco-system that translates the real world into a version understandable to AI. On this platform, all things and places will be machine-readable, subject to the power of algorithms to allow us to analyse and better understand the world in which we live.

The Challenge

Point clouds are unstructured and unordered data representing the real word with XYZ and RGB values. Although visually rich, these point clouds have limited spatial context associated for algorithms to extract meaningful information. In order to extract the spatial context, techniques such as point cloud segmentation and classification is commonly explored. This allows computers to recognize the composition of the 3D scene. However, the lack of effective semantic and instance segmentation techniques is currently acting as a blocker in this industry.

This DSG invites participants to explore point cloud segmentation techniques, both semantic and instance, in order to recognise objects such as roads, buildings, cars, trees, and ground in a large 3D urban environment. This will enable safer autonomous vehicles on the road, automated asset management in urban planning, and accurate digital twin simulations.

3D point cloud segmentation and classification has long been a popular research topic. The earlier methods generally follow the steps of ground filtering, clustering, feature extraction, and supervised classification. Features can be extracted from individual points directly or from a cluster of points, and then fed into a standard classifier. However, recent few years have witnessed exponential growth of 3D neural networks for point cloud understanding, such as PointNet, followed by PointNet++, PointCNN, etc. At the meantime, many benchmark data, both indoor and outdoor, became available to the research community, e.g., S3DIS, Scannet, Sematic3D, and recent Paris-Lille-3D.

A quick test of the PointNet++ on an airborne lidar point cloud classification has achieved around 90% overall accuracy. Regarding the given challenge of landscape scale point cloud semantic and instance segmentation, a comprehensive analyse of state-of-the-art 3D neural networks, especially those work efficiently on large scale outdoor point clouds, such as SPGraph and KPConv, is of interest. Since RGB information are available for point clouds generated by Structure-from-Motion, by incorporating RGB channels into the new networks, better results are expected.

The Data

There are both training data and test data. Point cloud file format: las, csv
Data Structure: X,Y,Z,class,R,G,B

Dataset 1: Training Data
This is a dataset representing a relatively small (200m x 200m) and computationally digestible suburb of Birmingham, UK. The dataset is a labelled point cloud.

The classes included in the dataset are houses, out-buildings, roads, trees, wooded areas, vegetation, cars and street furniture.
The exact classes to be used for training and test can be further determined.

Dataset 2: Test Data
This is a superset of the training data representing Perry Barr, Birmingham. This point cloud (las) cover a 1.19km2 region. This data has a much wider presence of different classes.

Dataset 3: Test Data
This point cloud (las) dataset represent a 3.2km2 area of central Cambridge with a diverse range of urban objects.

Goals

The overall goal is to segment the 3D point clouds and recognise the objects. There are few open source libraries available, such as pointnet++, SPGraph and KPConv. The participants are encouraged to explore the data on a specific aspect of interest in line with the overall goal, including:

1. Generation of semantic and instance labelled point cloud for supervised deep learning
2. Comparative analysis of 3D point cloud semantic segmentation approaches
3. Comparative analysis of 3D point cloud instance segmentation approaches

What we're looking for

We look for ideas from participants with difference disciplines or backgrounds, such as math, computer vision, machine learning, to push current techniques to their limit. Together, we would like to have a better understanding of the limitations of existing approaches and applicability of 2D methods to 3D, and to lay the foundation for an efficient and effective method to deal with large-scale 3D point cloud data.