Visualising Unstructured Data

New breed data visualisation for informing Machine Learning models

zegami

Introduction

Machine learning models are a complex network of *nodes* and *weights* and creating them is not a simple task. Even more complex is understanding the inner workings of these models and how they make decisions.

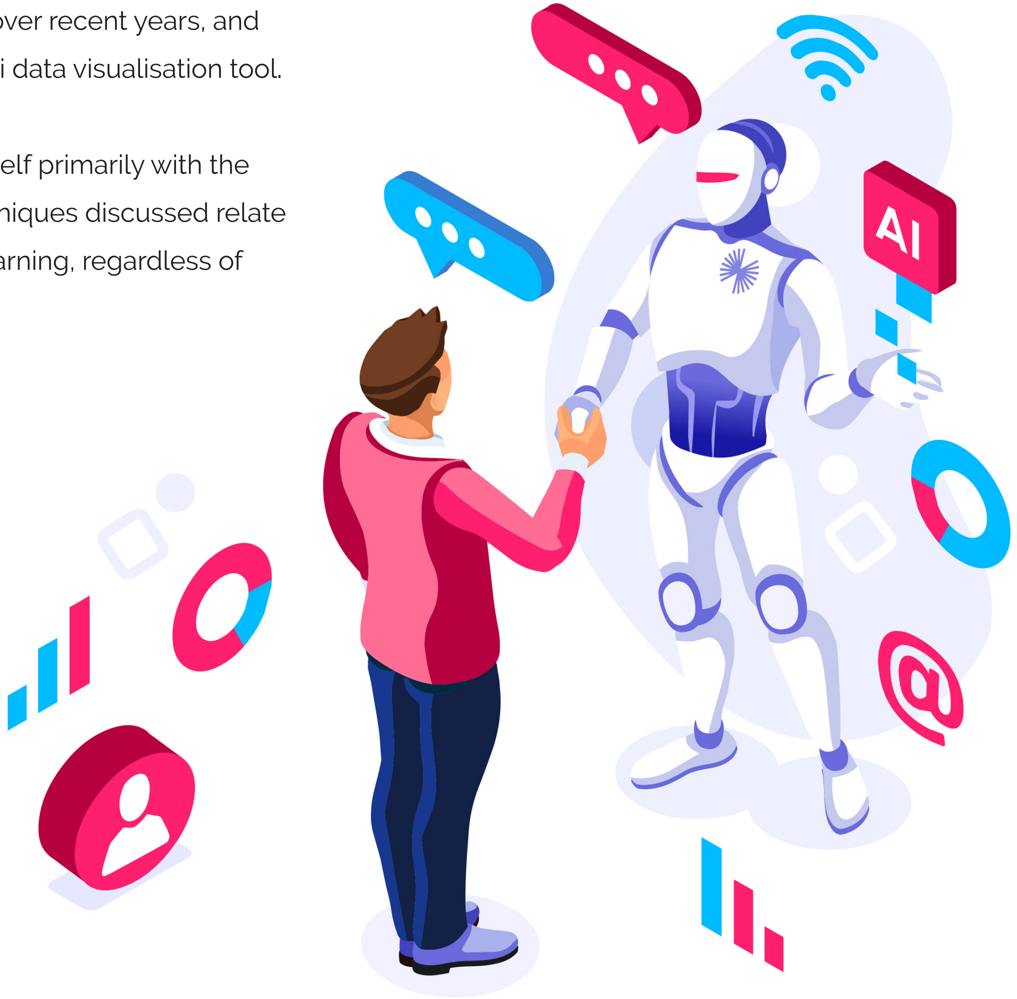
Because of this, data visualisation has a key role to play in the process of developing machine learning models.

Data visualisation not only helps us understand how a model works; it also enables us to create better training data and achieve models of greater accuracy.

This White Paper discusses techniques and practices developed at Zegami while building machine learning models in the field over recent years, and while developing the Zegami data visualisation tool.

While the paper concerns itself primarily with the handling of images, the techniques discussed relate more broadly to machine learning, regardless of data type.

> “While investment may be increasing exponentially there is still huge untapped potential in what will be achieved using AI.”
>
> Roger Noble, CTO, Zegami Ltd



Visualising Unstructured Data

New breed data visualisation for informing Machine Learning models

zegami

Introduction

Machine learning models are a complex network of *nodes* and *weights* and creating them is not a simple task. Even more complex is understanding the inner workings of these models and how they make decisions.

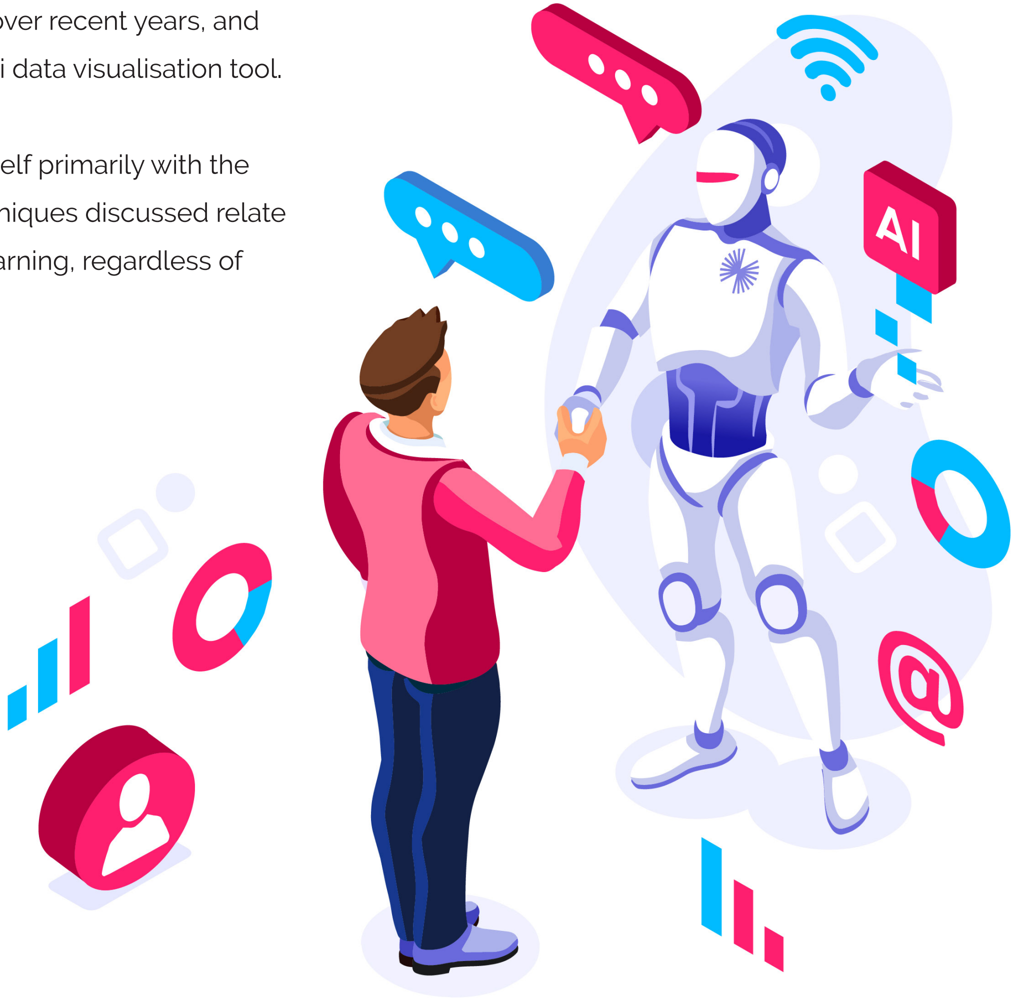
Because of this, data visualisation has a key role to play in the process of developing machine learning models.

Data visualisation not only helps us understand how a model works; it also enables us to create better training data and achieve models of greater accuracy.

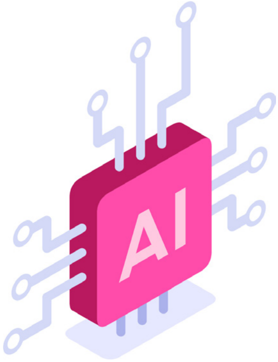
This White Paper discusses techniques and practices developed at Zegami while building machine learning models in the field over recent years, and while developing the Zegami data visualisation tool.

While the paper concerns itself primarily with the handling of images, the techniques discussed relate more broadly to machine learning, regardless of data type.

> “While investment may be increasing exponentially there is still huge untapped potential in what will be achieved using AI.”
>
> Roger Noble, CTO, Zegami Ltd



Machine Learning, Artificial Intelligence and the data that informs them.



The Machine Learning & AI boom

It will be news to no-one that Artificial Intelligence and Machine Learning have for some time now been receiving widespread interest and investment. Investment in Machine Learning projects is predicted to reach $57 billion by 2020, virtually a five-fold increase on the investment level in 2012. In parallel, the number of new greenfield projects is doubling every year. Increased activity and escalating investment make this an exciting time for all working in these fields.

The data we work with

While investment may be increasing exponentially there is still huge untapped potential in what will be achieved using AI.

Countless opportunities await to be unlocked in every area of research and enterprise by making better use of the data which exists and which continues to be generated and captured minute by minute.

Understanding types of data

Data may be broadly categorised into three types: structured data; semi-structured data; and unstructured data. For our purposes, we need compare only structured and unstructured data, leaving semi-structured data (data which does not sit conveniently in fixed fields or records, but does contain elements which can be used to organise and work with it) to one side.

Structured data is the material we commonly think of when the word data is mentioned : tables, spreadsheets, databases etc in which the data has been collated and packaged into a format which allows it to be worked with using formalised techniques.

Unstructured data, in contrast, describes things which can't be quantified or classified quite so simply.
With unstructured data, a set has data about it or even in it; however extracting that data will call for additional - and frequently complex - work.

Unstructured data might comprise images, video, emails, documents and tweets, inter alia. Working with such data can often be difficult due to its very nature: file sizes are frequently large, and so difficult to collect, store and move around.

Processing this kind of data requires a lot of compute and memory, frequently exceeding the capabilities of a single machine. It is only relatively recently, with the advance of cloud computing, that the capability to manage data of this kind effectively has become more accessible.

It is estimated that, when all three data types are considered, around 85% of all available data is unstructured. When looking at AI related projects, however, only 29% of these are actively built around unstructured data.

Why is unstructured data not used more in Machine Learning?

We have already made mention of some of the reasons that unstructured data can be difficult to work with. However, there are other factors which contribute to it not being utilised to its full potential.

Cloud computing looks at first glance to open up the possibilities for unstructured data. Yet scalable, Cloud-based architectures require specialist skills from a range of individuals within an organisation if they are to be used effectively. Data Scientists, Data Engineers, Software Engineers, Developers, Database Administrators and other capabilities may all be required if a project is to leverage unstructured data through the Cloud.

This is not solely a human resource/skills issue, however.

There is also a significant dearth of software suited to working with unstructured data. Most data analytics and visualisation tools are built with structured data in mind, due largely to the fact that working with unstructured data requires a wholly different understanding and approach.

This lack of software impacts the whole process of developing Machine Learning models, causing a great deal of time to be spent on low value tasks. Pleasingly, in this area Machine Learning is the perfect tool to help us better understand this kind of data...

...a case of Machine Learning coming to the aid of Machine Learning!

Section 2

Challenges & Opportunities for improvement

The Machine Learning workflow

A useful starting point in understanding an operational challenge is to look at the steps that a project might go through from start to finish. By considering the situation as a workflow, it becomes easier to see where the potential bottlenecks are, and which areas can be improved upon.

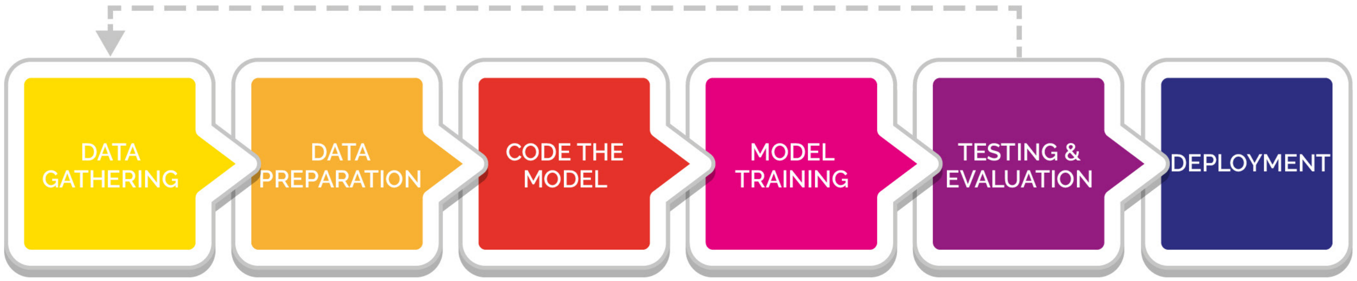
AI projects typically assume the following flow, starting with the collection and preparation of data, progressing through model training to eventual deployment to the end user.

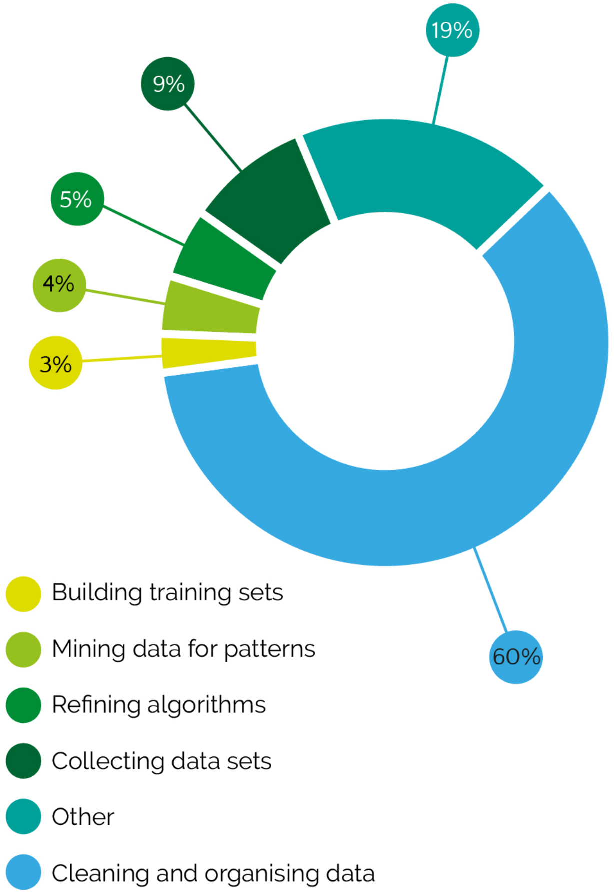
Data Scientists or Data Cleaners?

A survey carried out recently amongst data scientists looked at what they spent most of their time on relative to the workflow outlined below.

82% of the overall time spent by these highly skilled and valuable people went on sourcing and cleaning data, as opposed to being deployed on high-value activities like mining data or refining algorithms.

What's more, much of this time went into labelling and annotating data - manually adding descriptive metadata about an item so that it could then be used for training Machine Learning models.





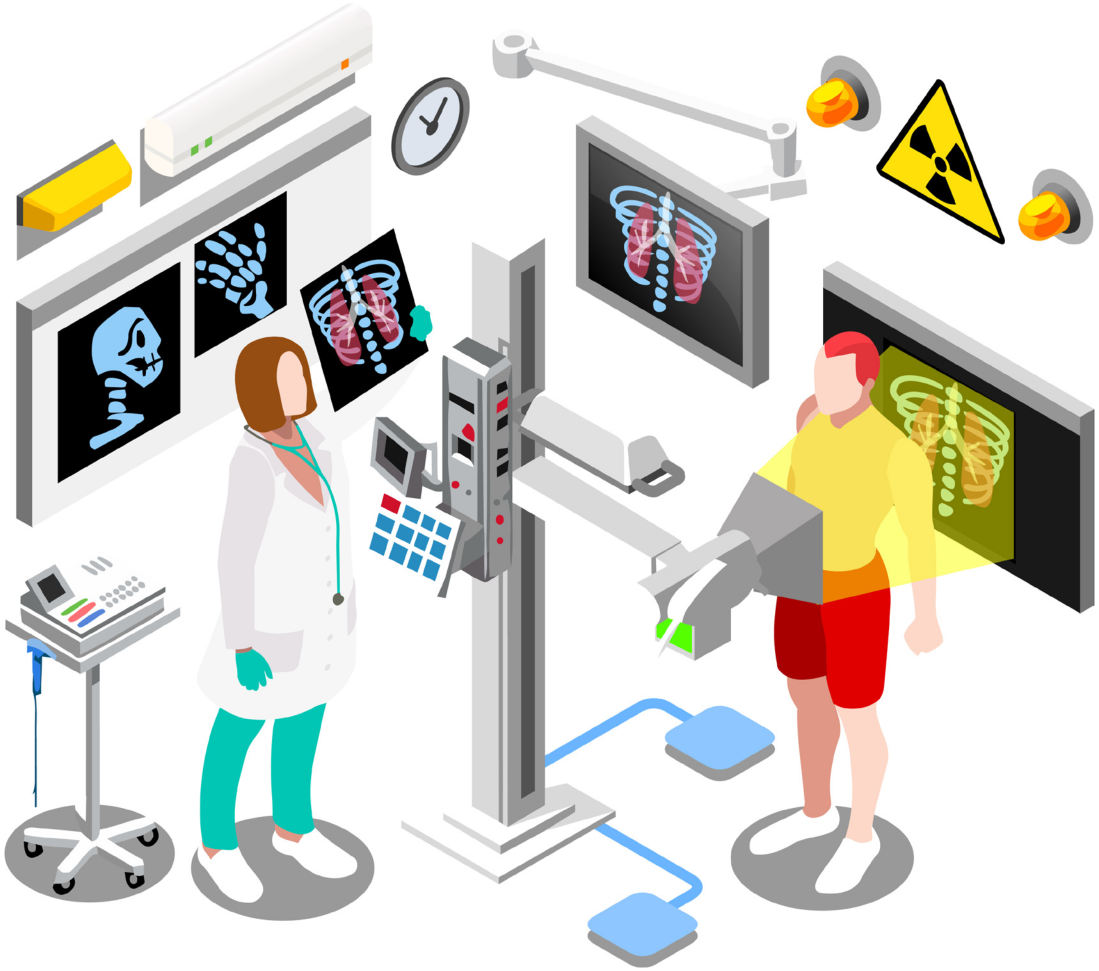
The problem and the Big Question

Imagine training a Machine Learning model to identify a tumour in an x-ray image.

In order for a Machine Learning algorithm to understand what the tumour looks like it requires thousands - even tens of thousands - of example images to train on. Each of these images needs to be hand labelled to highlight exactly where the tumour is. This, however, can only be done by an expert radiologist.

Experts of this kind generally do not have the time (and may well not have the inclination) to be labelling thousands of images. Yet because the expertise required is so specialised the work is virtually impossible to outsource and offshore.
All of which raises a key question, the answer to which has massive implications for the future of Machine Learning: What can be done to help make the process of developing Machine Learning models more productive?

Can we build tools that are capable of speeding up, and removing the tedium, from collecting, preparing and labelling data?



Areas of improvement

Based on our experience at Zegami in developing machine learning models, we have developed a number of techniques that have proven their ability to make a dramatic difference to the process of developing Machine Learning models.

Data visualisation

The first approach to have evidenced significant potential to improve the process of developing Machine Learning models from unstructured data, is that of data visualisation.

By visualising and communicating the structure of a training data set, users were enabled to quickly spot under-represented classes, inconsistencies and missing data which would not have been otherwise apparent.

While data visualisation is common practise when developing machine learning models, it is generally used only for structured data. Critically, the tools used to do this are unable to process unstructured data. What's needed are new types of data visualisation, better suited to working with unstructured data.

Subject identification and labelling

Machine learning requires a high volume of data in order to train a model to learn something new. We will show later how, when dealing with visual data such as images, it is necessary to hand label each item by drawing a 'bounding box' around the point of interest. By creating better tools which can not only help speed up the bounding box process, but also produce higher quality labels, much of the pain of preparing such data could be removed.

Transfer learning

Transfer learning is the process of taking an existing, pre-trained model and re-training it to learn something new. This means that training a new model requires relatively little data (other than that which was required to train the original model).
Cutting down the bulk data requirement in this way means less time needs be spent labelling, and so the training time is significantly reduced.

Technological Advances and the Zegami tool

Huskies vs. Wolves. The case for better data.

It is beneficial to consider a concrete example of a Machine Learning problem and look at how, by applying the techniques outlined in (6) we might be able to build better models.

The following example, a classic from academia, is based on a paper[2] in which researchers demonstrated that classification models can easily contain biases. They did this by training a model to identify the difference between different breeds of canine, including both domestic dogs and wild canines such as wolves.

The training data set included pictures of both domestic and wild dogs in their 'everyday' habitats, and was labelled using bounding boxes around each dog. When the data was classified, a problem quickly appeared in differentiating between huskies and wolves, with huskies in certain scenarios being incorrectly classed as wolves.

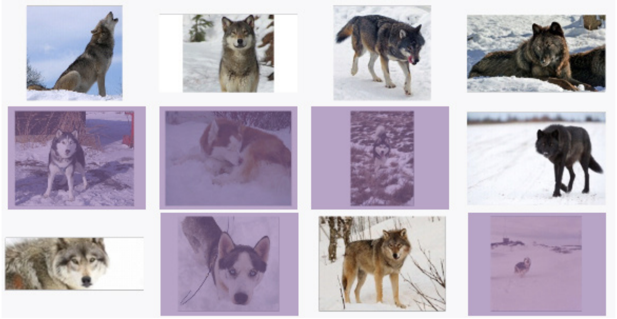
By visualising all of the misclassified images, however, a pattern emerged.

In the cases where huskies (domestic) were being erroneously identified as wolves (wild and usually photographed in snow), the common denominator was that the images contained a lot of snow in the background (ie the huskies had been photographed outdoors in a snow situation).

In effect what the model had created was a 'snow classifier' – it was making the judgement based on its successful classifications of wolves, that if there were snow in the background, the animal in the image must be a wolf.

In some ways, this was very smart. By picking up that snow was a common denominator in images that really did show wolves, it found the shortest path to identifying images which it believed had wolves in.

However, simply because a dog happens to be in snow does not necessarily make it a wolf.



This example may seem trivial. However, it turns out to be a common problem when collecting data to be used for training. Unconscious biases and blind spots can easily creep into data, causing a model to behave in unexpected ways. To compound matters, such problems are not always immediately apparent and will often be noticed only once the model has been deployed to end users.

[2] https://arxiv.org/abs/1602.04938

Data Visualisation

In addressing the shortcomings outlined previously, data visualisation can play a big role in helping us better understand a data set.

Traditionally data visualisation has only been applied to structured data. Tools such as Excel, Power BI and Tableau do an excellent job of summarising and visualising tabular data, for example.

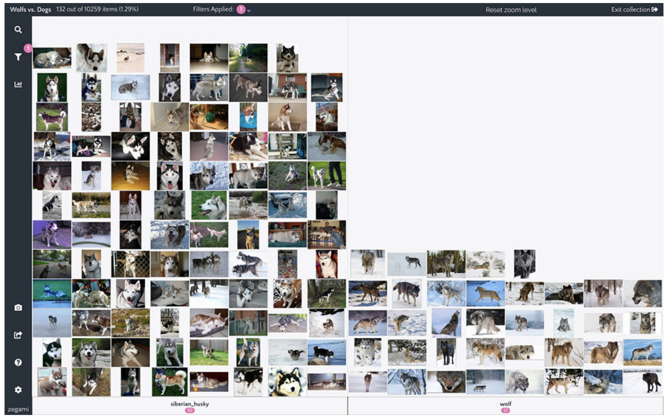
For unstructured data, a different approach is required. This is because 'structured data' visualisation techniques simply won't work. We cannot, for example, summarise a group of images in a pie chart. As a consequence, we need new and better tools, designed specifically to work with unstructured data types.

Data visualisation and Zegami

Zegami is a tool developed to specifically addresses the challenges of working with unstructured data.

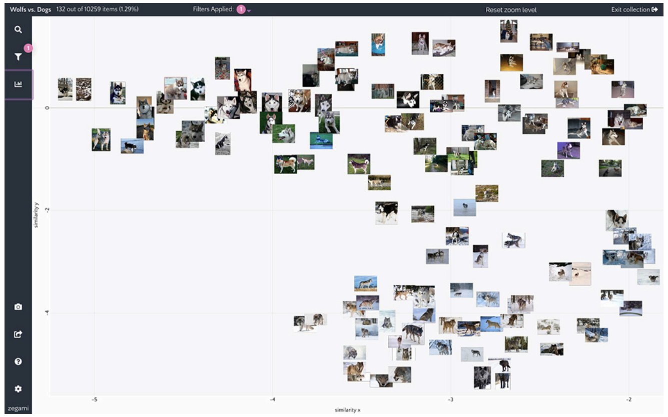
By treating each item as a data point and pairing it with its metadata, Zegami lets users consider the visual nature of images, video and documents while also using familiar tools for visualising structured data. So, in the example use case of Wolves v Huskies, all of the husky and wolf images can be loaded into Zegami, with the user able to see instantly how and where potential problems occur in the data.

By arranging the images so that the two classes sit side by side, it is immediately apparent that there are few Wolf data points compared to Husky data points. Equally, by visualising the Wolf and Husky data in this way, it becomes apparent that there is some similarity between the two classes of image that may perhaps cause issues.

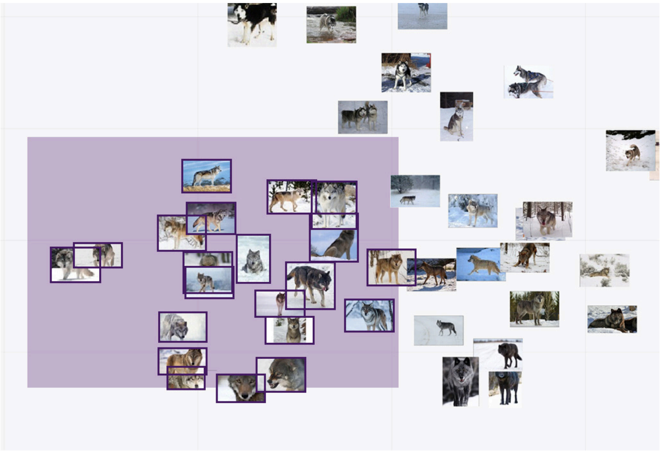


The next step might be to incorporate some unsupervised Machine Learning to cluster the images together based on visual similarity.

By doing this, we can see clearly that two distinct clusters form at the top (huskies) and bottom (wolves). In the middle of the two clusters, however, it is noticeable that there is something of a 'grey' area which contains a mix of huskies and wolves. This is the most interesting observation, as it highlights a potential problem: that there is a strong similarity between these images which could cause our Machine Learning model to misclassify a husky as a wolf or vice versa.



Once these problem images have been identified, however, Zegami makes it easy to select clusters of them in order to add tags or other additional metadata that can then be fed back into the training process. Tagged items can then be quickly filtered out, making it possible to focus only on 'problem' images which can then be 'fixed' with the addition of better quality labels.

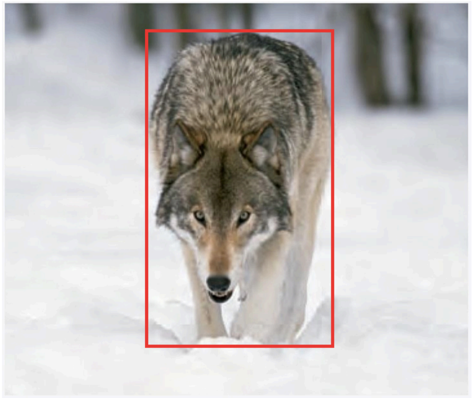


Data Annotation

The quality of data annotations is a further area requiring improvement.

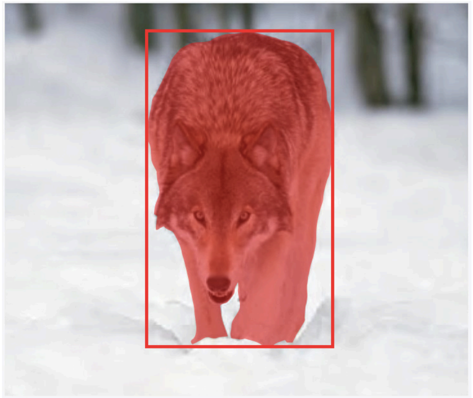
It is common in any kind of training for object detection to use bounding boxes to indicate the area of interest.

This involves drawing a box around the relevant area of the image to indicate the extent of the information to be fed into the training process.



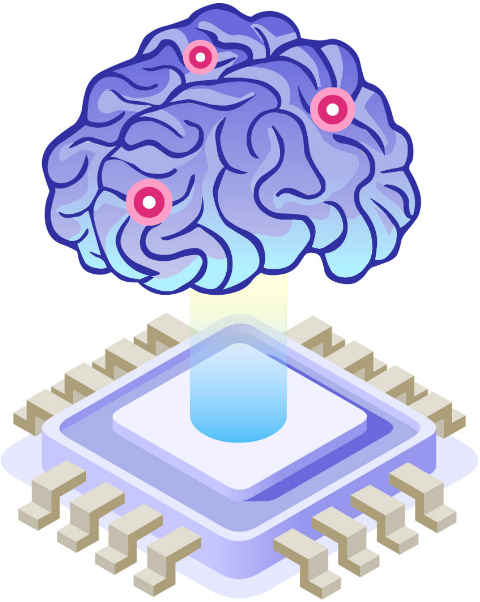
By their nature, bounding boxes include other pixels (in this example the 'non-wolf' part of the image - the snow) which are then also taken into consideration when training the model.

A preferable approach to object detection would be to create a mask over the object, so that only the relevant 'wolf' pixels are used to train the model.



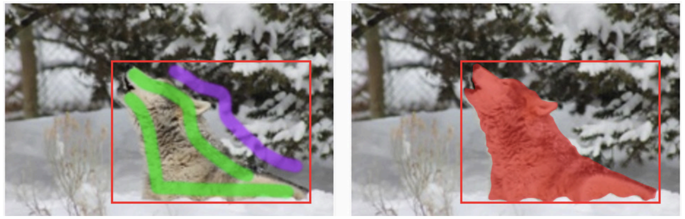
However, creating detailed masks like this can be extremely time consuming, particularly when dealing with thousands, or even millions, of images. While bounding boxes tend to produce less than ideal results, the time (and money) required to create them is significantly less.

Additionally, in non-trivial cases such as medical imaging, it can be impossible to outsource the annotation task, i.e. the drawing of the bounding box or creation of the mask, leaving the bulk of the annotation burden with time-poor experts. With the inherent restriction on their availability, a faster process such as bounding boxes will often win out over a more time consuming, albeit more accurate, process such as making masks.



Data annotation and Zegami

Amethyst is a companion tool to the Zegami product, capable of automatically generating segmentation masks using advanced computer vision techniques. The tool can create masks in almost the same amount of time normally required to create bounding boxes, while outputting significantly superior quality training data.



Using a bounding box as a starting point, Amethyst is able to differentiate foreground from background, even when the characteristics of the object make it look similar to the background.

As certain kinds of images, generally containing complex objects, may still prove difficult for the tool to resolve, Amethyst allows users to draw additional positive and negative hints to help it decide what is foreground and what background.

In either case, the masks outputted by Amethyst are ready to be fed into an instance segmentation algorithm such as Mask R-CNN, and to go to work training new models.

Transfer learning

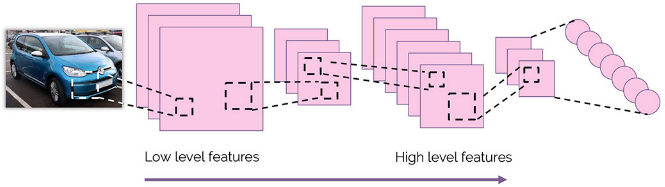
The technique known as Transfer Learning offers a further excellent opportunity for improvement in training Machine Learning models.

Transfer learning involves taking an existing, pre-trained model and repurposing this for a new task. In practise, this means re-training only parts of the model instead of starting from scratch, dramatically reducing the amount of training data and training time required.

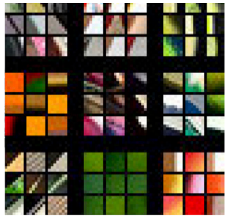
To better understand how this works, we can look at the internals of a Deep Neural Network.

The network is made up of multiple layers, with each layer connected to the next.

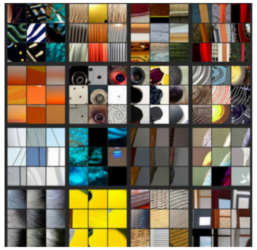
When reading from left to right, the first layers are concerned with extracting features, while the final layer is used to classify the input image by estimating a percentage likelihood that it's one of the things it knows about.

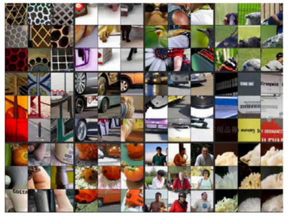


The first layers of the network are used to extract low level features or simple lines, edges and corners.



As we come to the middle layers, things like basic shapes and textures begin to form. And finally, the high-level features progressively combine these into more complex shapes and objects. Transfer Learning is then a way to retrain an existing model to learn something new. It keeps the lower layers, that have learnt basic features, but then tells the higher layers how those low-level features should be combined in order to identify our new objects.





Not only does this dramatically reduce the amount of time needed to train a new model but also, when combined with masking, means that very few examples are needed in order to produce a model with a good level of accuracy.

Case Study:
Natural disaster damage

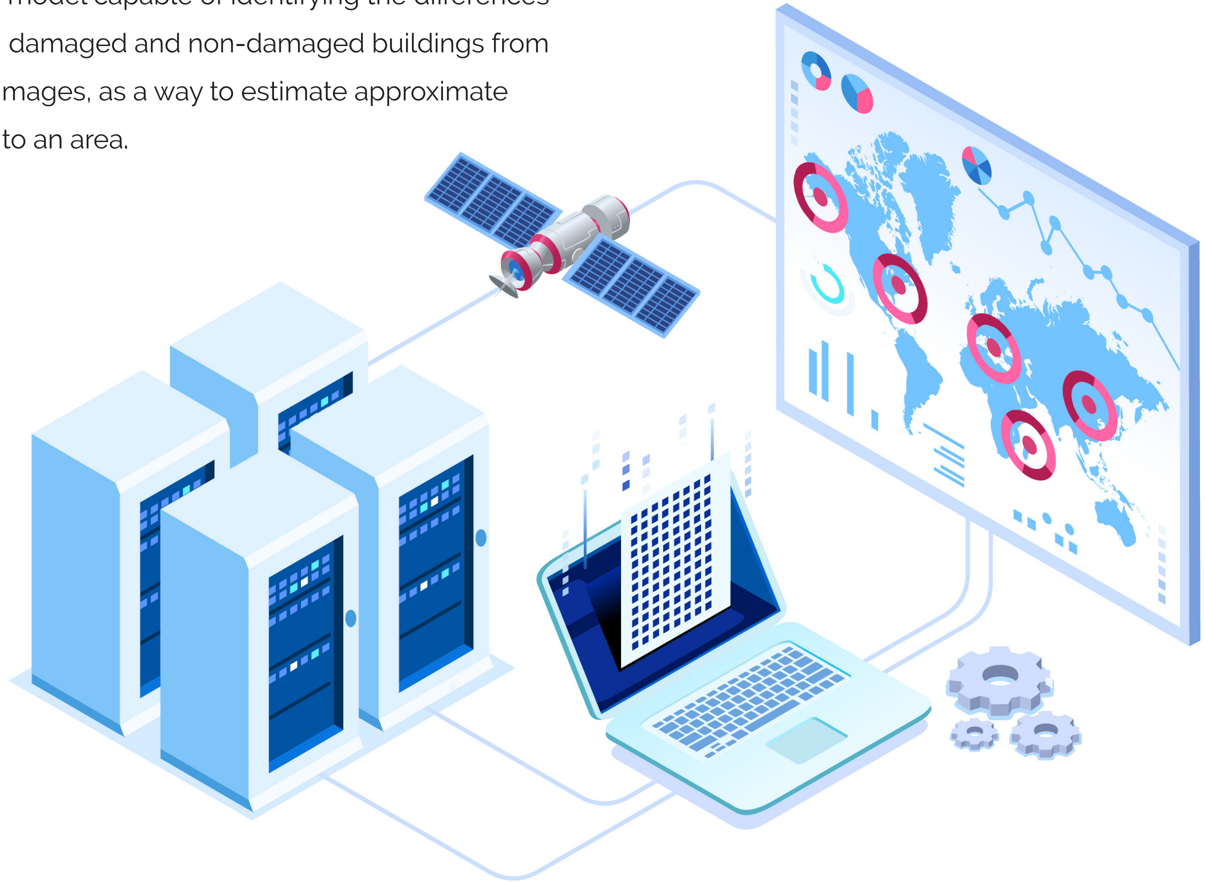
Background

The team at Zegami recently embarked on a project that utilised the techniques outlined above to develop a Machine Learning model which can be used to identify damage to buildings caused by natural disasters.

When a significant natural disaster occurs it can easily overwhelm first responders, as they do not know exactly where the damage is and which areas to prioritise. However, the first four days following such a disaster are the most important, time is of the essence and situational information is key.

During this timeframe, international awareness and interest are at their peaks and people are most willing to help. Satellite imagery can play an important role in helping to get a better understanding of the extent of the damage, as well as knowing where the resources and effort on offer will prove most effective.

To help with this, we decided to build a Machine Learning model capable of identifying the differences between damaged and non-damaged buildings from satellite images, as a way to estimate approximate damage to an area.



Process - utilisation of Zegami and Amethyst

The training data set was a single 100km² area over the coast of Dominica, hit by hurricane Irma in 2017. The satellite image of the area was divided up into hundreds of slices, called 'chips'.

On each chip, we annotated a few hundred buildings, split into two classes - 'damaged' and 'non-damaged', using our Amethyst annotation tool.

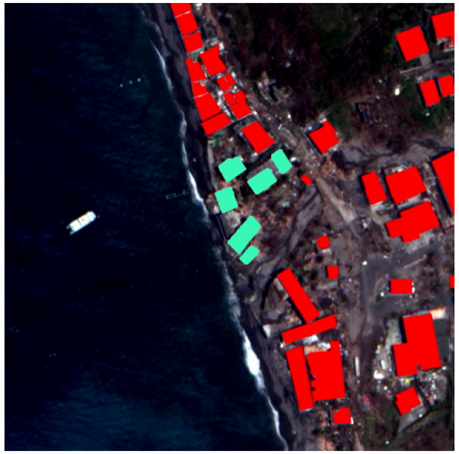
Once the labelling was complete, we used Mask R-CNN (a technique for instance segmentation) to train a new model based on a pre-trained model with ImageNet weights.

The training process took us around 20 minutes to do on a moderately powerful machine and GPU.

We then inferenced new images and cropped these out into a new Zegami collection, to help us better understand how well the model was performing on different building types.

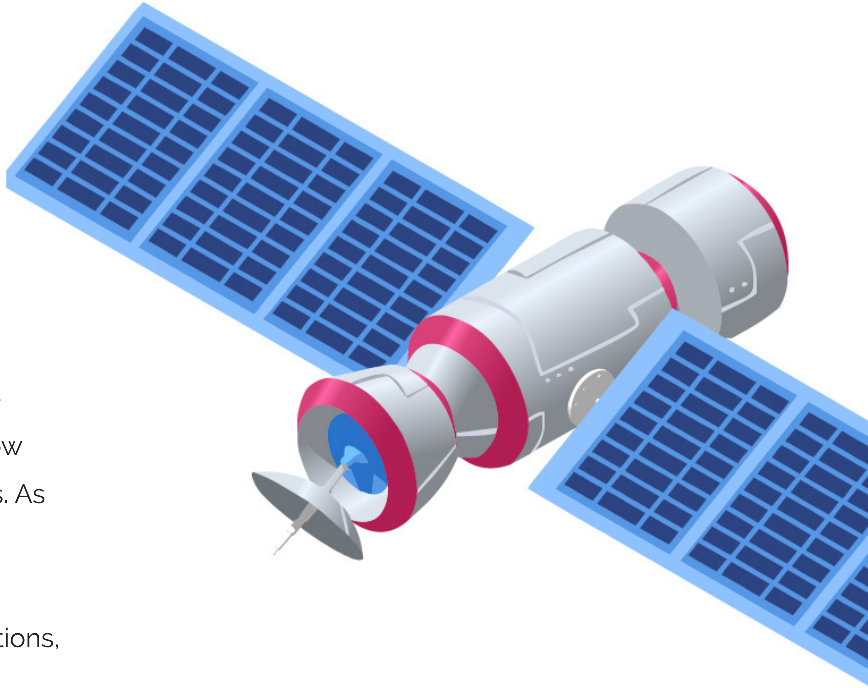
For this, we used a separate technique to extract features from each of the building instances, and reduced those vectors down to produce a 2D scatter plot of similarity.

As the image below shows, buildings are roughly grouped by colour and shape.



Using Zegami, we were then able to explore the similarities and get a better understanding of how well the model predicted each of the categories. As we have the confidence score of each of the buildings, we can add filters on low confidence instances and see if there are any mis-classifications, as well as looking at high confidence instances.
It is also useful in this situation to identify the ratio of damaged to undamaged buildings. This, when combined with the geospatial information, enables us to see where the greatest damage lies.

By doing this, we were able to rapidly iterate on the model by having a visual feedback loop into how it was performing.



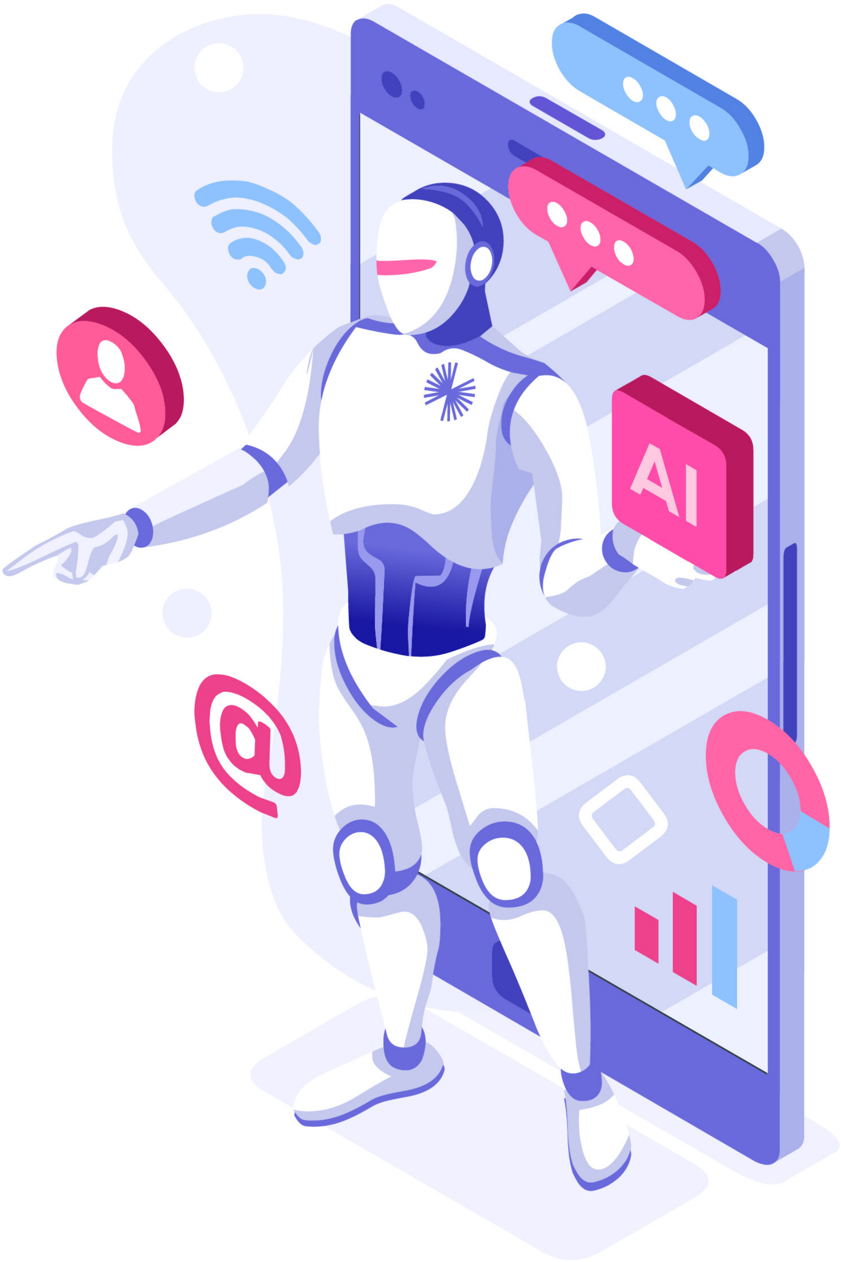
Summary

In summary, creating Machine Learning models is a non-trivial task. This is especially true when working with images, video or any other kind of unstructured data.

As in most situations, a model can only ever be as good as the data it has been trained on, placing a large burden on ensuring that the training set is well representative of real-world data.

To make things more difficult, biases are easy to introduce into a data set but notoriously difficult to spot. Data visualisation can play a key role in identifying biases, but traditional data visualisation tools struggle to function with unstructured data.

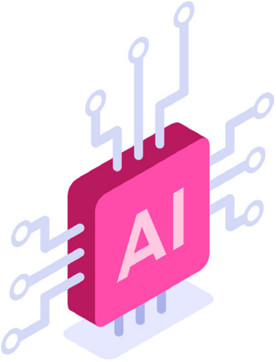
For this reason, a new breed of tools specifically designed to address this problem is needed. Zegami sets a marker in the development of such tools.



Learn more about Zegami at:

www.zegami.com

Machine Learning, Artificial Intelligence and the data that informs them.



The Machine Learning & AI boom

It will be news to no-one that Artificial Intelligence and Machine Learning have for some time now been receiving widespread interest and investment. Investment in Machine Learning projects is predicted to reach $57 billion by 2020, virtually a five-fold increase on the investment level in 2012. In parallel, the number of new greenfield projects is doubling every year. Increased activity and escalating investment make this an exciting time for all working in these fields.

The data we work with

While investment may be increasing exponentially there is still huge untapped potential in what will be achieved using AI.

Countless opportunities await to be unlocked in every area of research and enterprise by making better use of the data which exists and which continues to be generated and captured minute by minute.

Understanding types of data

Data may be broadly categorised into three types: structured data; semi-structured data; and unstructured data. For our purposes, we need compare only structured and unstructured data, leaving semi-structured data (data which does not sit conveniently in fixed fields or records, but does contain elements which can be used to organise and work with it) to one side.

Structured data is the material we commonly think of when the word data is mentioned : tables, spreadsheets, databases etc in which the data has been collated and packaged into a format which allows it to be worked with using formalised techniques.

Unstructured data, in contrast, describes things which can't be quantified or classified quite so simply.
With unstructured data, a set has data about it or even in it; however extracting that data will call for additional - and frequently complex - work.

Unstructured data might comprise images, video, emails, documents and tweets, inter alia. Working with such data can often be difficult due to its very nature: file sizes are frequently large, and so difficult to collect, store and move around.

Processing this kind of data requires a lot of compute and memory, frequently exceeding the capabilities of a single machine. It is only relatively recently, with the advance of cloud computing, that the capability to manage data of this kind effectively has become more accessible.

It is estimated that, when all three data types are considered, around 85% of all available data is unstructured. When looking at AI related projects, however, only 29% of these are actively built around unstructured data.

Why is unstructured data not used more in Machine Learning?

We have already made mention of some of the reasons that unstructured data can be difficult to work with. However, there are other factors which contribute to it not being utilised to its full potential.

Cloud computing looks at first glance to open up the possibilities for unstructured data. Yet scalable, Cloud-based architectures require specialist skills from a range of individuals within an organisation if they are to be used effectively. Data Scientists, Data Engineers, Software Engineers, Developers, Database Administrators and other capabilities may all be required if a project is to leverage unstructured data through the Cloud.

This is not solely a human resource/skills issue, however.

There is also a significant dearth of software suited to working with unstructured data. Most data analytics and visualisation tools are built with structured data in mind, due largely to the fact that working with unstructured data requires a wholly different understanding and approach.

This lack of software impacts the whole process of developing Machine Learning models, causing a great deal of time to be spent on low value tasks. Pleasingly, in this area Machine Learning is the perfect tool to help us better understand this kind of data...

...a case of Machine Learning coming to the aid of Machine Learning!

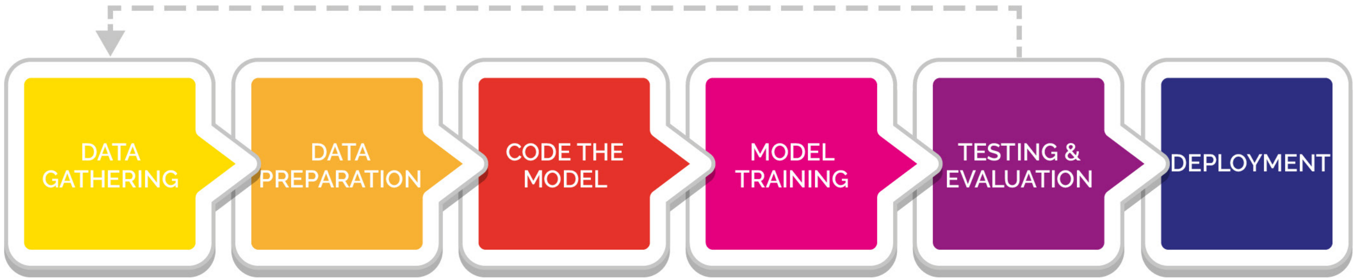
Section 2

Challenges & Opportunities for improvement

The Machine Learning workflow

A useful starting point in understanding an operational challenge is to look at the steps that a project might go through from start to finish. By considering the situation as a workflow, it becomes easier to see where the potential bottlenecks are, and which areas can be improved upon.

AI projects typically assume the following flow, starting with the collection and preparation of data, progressing through model training to eventual deployment to the end user.

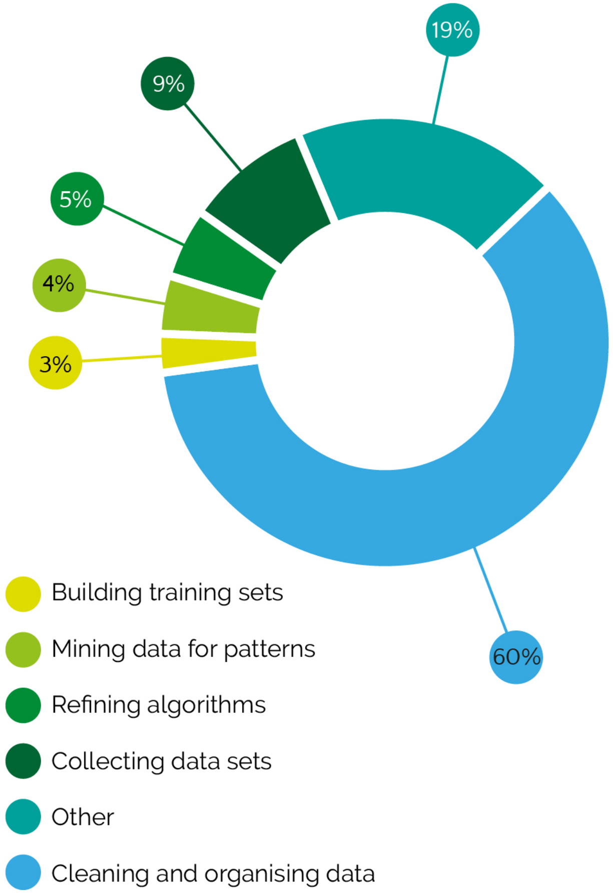


Data Scientists or Data Cleaners?

A survey carried out recently amongst data scientists looked at what they spent most of their time on relative to the workflow outlined below.

82% of the overall time spent by these highly skilled and valuable people went on sourcing and cleaning data, as opposed to being deployed on high-value activities like mining data or refining algorithms.

What's more, much of this time went into labelling and annotating data - manually adding descriptive metadata about an item so that it could then be used for training Machine Learning models.



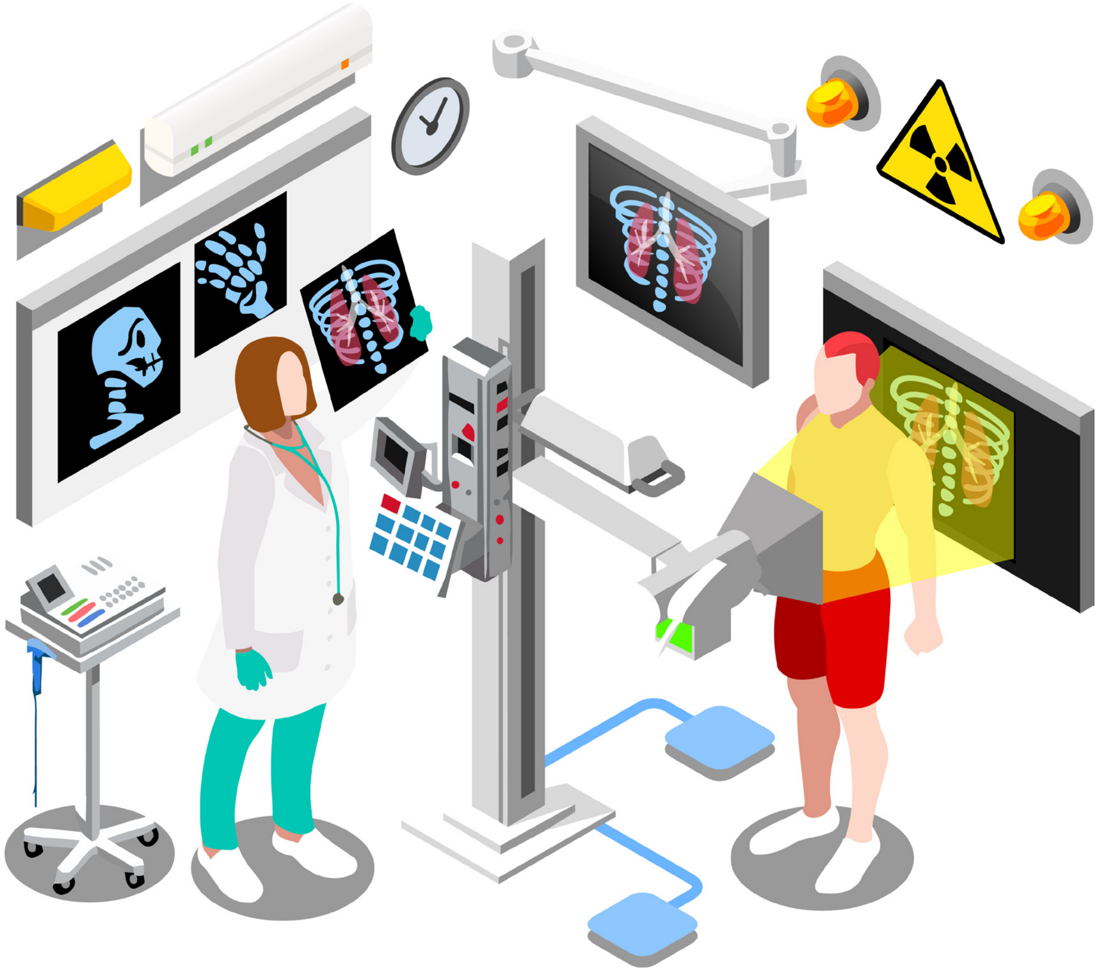
The problem and the Big Question

Imagine training a Machine Learning model to identify a tumour in an x-ray image.

In order for a Machine Learning algorithm to understand what the tumour looks like it requires thousands - even tens of thousands - of example images to train on. Each of these images needs to be hand labelled to highlight exactly where the tumour is. This, however, can only be done by an expert radiologist.

Experts of this kind generally do not have the time (and may well not have the inclination) to be labelling thousands of images. Yet because the expertise required is so specialised the work is virtually impossible to outsource and offshore. All of which raises a key question, the answer to which has massive implications for the future of Machine Learning: What can be done to help make the process of developing Machine Learning models more productive?

Can we build tools that are capable of speeding up, and removing the tedium, from collecting, preparing and labelling data?



Areas of improvement

Based on our experience at Zegami in developing machine learning models, we have developed a number of techniques that have proven their ability to make a dramatic difference to the process of developing Machine Learning models.

Data visualisation

The first approach to have evidenced significant potential to improve the process of developing Machine Learning models from unstructured data, is that of data visualisation.

By visualising and communicating the structure of a training data set, users were enabled to quickly spot under-represented classes, inconsistencies and missing data which would not have been otherwise apparent.

While data visualisation is common practise when developing machine learning models, it is generally used only for structured data. Critically, the tools used to do this are unable to process unstructured data. What's needed are new types of data visualisation, better suited to working with unstructured data.

Subject identification and labelling

Machine learning requires a high volume of data in order to train a model to learn something new. We will show later how, when dealing with visual data such as images, it is necessary to hand label each item by drawing a 'bounding box' around the point of interest. By creating better tools which can not only help speed up the bounding box process, but also produce higher quality labels, much of the pain of preparing such data could be removed.

Transfer learning

Transfer learning is the process of taking an existing, pre-trained model and re-training it to learn something new. This means that training a new model requires relatively little data (other than that which was required to train the original model).
Cutting down the bulk data requirement in this way means less time needs be spent labelling, and so the training time is significantly reduced.

Technological Advances and the Zegami tool

Huskies vs. Wolves. The case for better data.

It is beneficial to consider a concrete example of a Machine Learning problem and look at how, by applying the techniques outlined in (6) we might be able to build better models.

The following example, a classic from academia, is based on a paper[2] in which researchers demonstrated that classification models can easily contain biases. They did this by training a model to identify the difference between different breeds of canine, including both domestic dogs and wild canines such as wolves.

The training data set included pictures of both domestic and wild dogs in their 'everyday' habitats, and was labelled using bounding boxes around each dog. When the data was classified, a problem quickly appeared in differentiating between huskies and wolves, with huskies in certain scenarios being incorrectly classed as wolves.

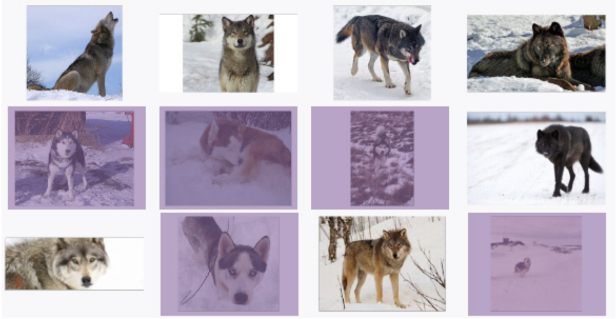
By visualising all of the misclassified images, however, a pattern emerged.

In the cases where huskies (domestic) were being erroneously identified as wolves (wild and usually photographed in snow), the common denominator was that the images contained a lot of snow in the background (ie the huskies had been photographed outdoors in a snow situation).

In effect what the model had created was a 'snow classifier' – it was making the judgement based on its successful classifications of wolves, that if there were snow in the background, the animal in the image must be a wolf.

In some ways, this was very smart. By picking up that snow was a common denominator in images that really did show wolves, it found the shortest path to identifying images which it believed had wolves in.

However, simply because a dog happens to be in snow does not necessarily make it a wolf.



This example may seem trivial. However, it turns out to be a common problem when collecting data to be used for training. Unconscious biases and blind spots can easily creep into data, causing a model to behave in unexpected ways. To compound matters, such problems are not always immediately apparent and will often be noticed only once the model has been deployed to end users.

[2] https://arxiv.org/abs/1602.04938

Data Visualisation

In addressing the shortcomings outlined previously, data visualisation can play a big role in helping us better understand a data set.

Traditionally data visualisation has only been applied to structured data. Tools such as Excel, Power BI and Tableau do an excellent job of summarising and visualising tabular data, for example.

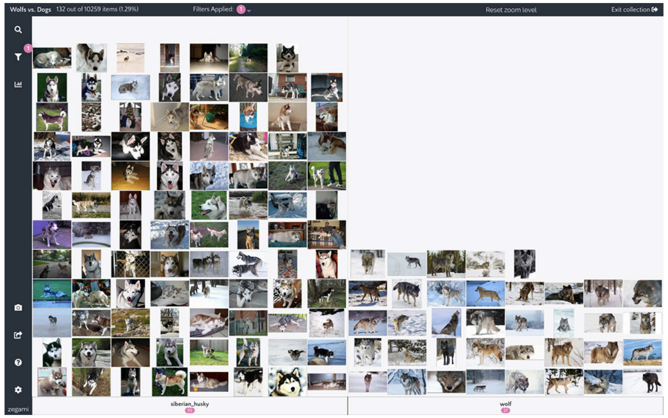
For unstructured data, a different approach is required. This is because 'structured data' visualisation techniques simply won't work. We cannot, for example, summarise a group of images in a pie chart. As a consequence, we need new and better tools, designed specifically to work with unstructured data types.

Data visualisation and Zegami

Zegami is a tool developed to specifically addresses the challenges of working with unstructured data.

By treating each item as a data point and pairing it with its metadata, Zegami lets users consider the visual nature of images, video and documents while also using familiar tools for visualising structured data. So, in the example use case of Wolves v Huskies, all of the husky and wolf images can be loaded into Zegami, with the user able to see instantly how and where potential problems occur in the data.

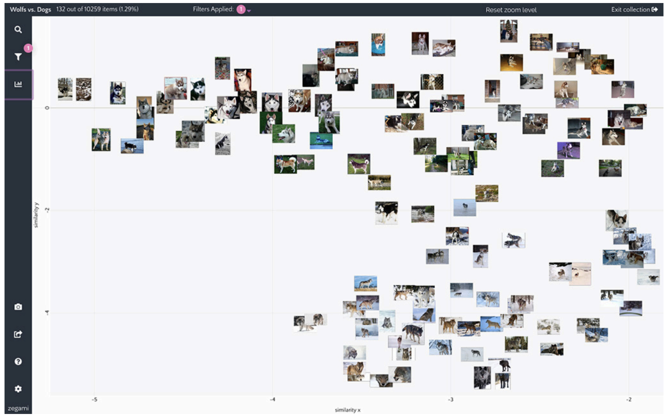
By arranging the images so that the two classes sit side by side, it is immediately apparent that there are few Wolf data points compared to Husky data points. Equally, by visualising the Wolf and Husky data in this way, it becomes apparent that there is some similarity between the two classes of image that may perhaps cause issues.

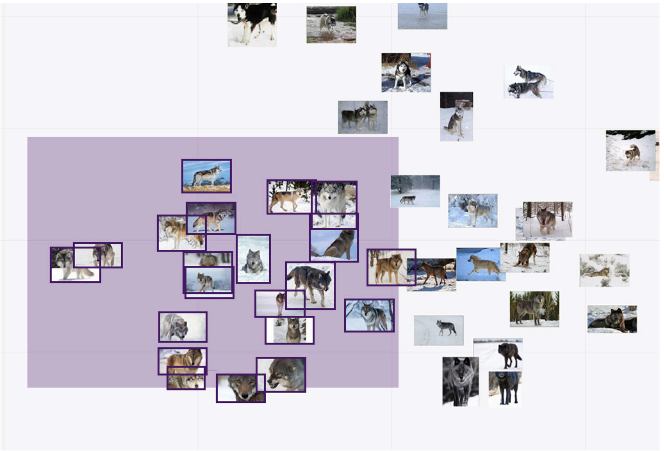


The next step might be to incorporate some unsupervised Machine Learning to cluster the images together based on visual similarity.

By doing this, we can see clearly that two distinct clusters form at the top (huskies) and bottom (wolves). In the middle of the two clusters, however, it is noticeable that there is something of a 'grey' area which contains a mix of huskies and wolves. This is the most interesting observation, as it highlights a potential problem: that there is a strong similarity between these images which could cause our Machine Learning model to misclassify a husky as a wolf or vice versa.

Once these problem images have been identified, however, Zegami makes it easy to select clusters of them in order to add tags or other additional metadata that can then be fed back into the training process. Tagged items can then be quickly filtered out, making it possible to focus only on 'problem' images which can then be 'fixed' with the addition of better quality labels.



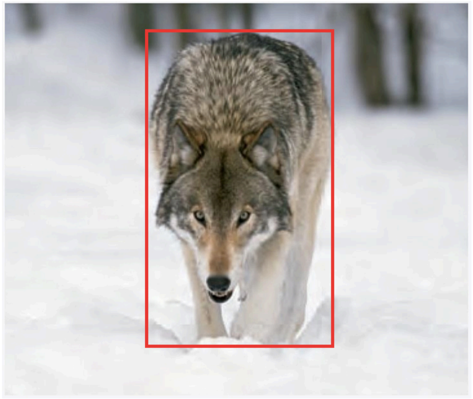


Data Annotation

The quality of data annotations is a further area requiring improvement.

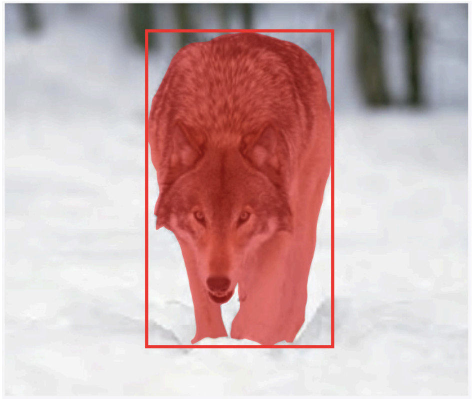
It is common in any kind of training for object detection to use bounding boxes to indicate the area of interest.

This involves drawing a box around the relevant area of the image to indicate the extent of the information to be fed into the training process.



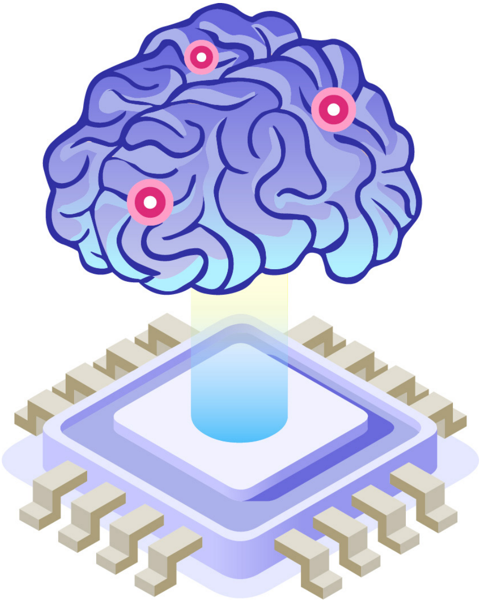
By their nature, bounding boxes include other pixels (in this example the 'non-wolf' part of the image - the snow) which are then also taken into consideration when training the model.

A preferable approach to object detection would be to create a mask over the object, so that only the relevant 'wolf' pixels are used to train the model.



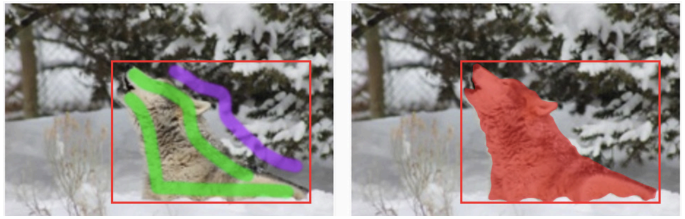
However, creating detailed masks like this can be extremely time consuming, particularly when dealing with thousands, or even millions, of images. While bounding boxes tend to produce less than ideal results, the time (and money) required to create them is significantly less.

Additionally, in non-trivial cases such as medical imaging, it can be impossible to outsource the annotation task, i.e. the drawing of the bounding box or creation of the mask, leaving the bulk of the annotation burden with time-poor experts. With the inherent restriction on their availability, a faster process such as bounding boxes will often win out over a more time consuming, albeit more accurate, process such as making masks.



Data annotation and Zegami

Amethyst is a companion tool to the Zegami product, capable of automatically generating segmentation masks using advanced computer vision techniques. The tool can create masks in almost the same amount of time normally required to create bounding boxes, while outputting significantly superior quality training data.



Using a bounding box as a starting point, Amethyst is able to differentiate foreground from background, even when the characteristics of the object make it look similar to the background.

As certain kinds of images, generally containing complex objects, may still prove difficult for the tool to resolve, Amethyst allows users to draw additional positive and negative hints to help it decide what is foreground and what background.

In either case, the masks outputted by Amethyst are ready to be fed into an instance segmentation algorithm such as Mask R-CNN, and to go to work training new models.

Transfer learning

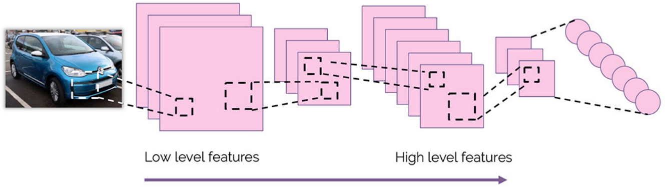
The technique known as Transfer Learning offers a further excellent opportunity for improvement in training Machine Learning models.

Transfer learning involves taking an existing, pre-trained model and repurposing this for a new task. In practise, this means re-training only parts of the model instead of starting from scratch, dramatically reducing the amount of training data and training time required.

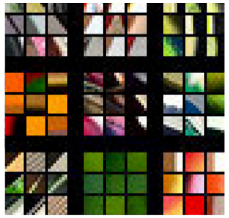
To better understand how this works, we can look at the internals of a Deep Neural Network.

The network is made up of multiple layers, with each layer connected to the next.

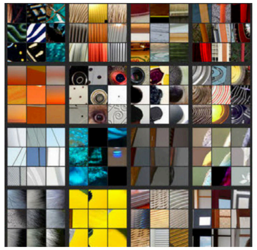
When reading from left to right, the first layers are concerned with extracting features, while the final layer is used to classify the input image by estimating a percentage likelihood that it's one of the things it knows about.

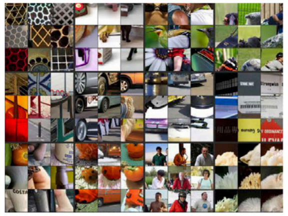


The first layers of the network are used to extract low level features or simple lines, edges and corners.



As we come to the middle layers, things like basic shapes and textures begin to form. And finally, the high-level features progressively combine these into more complex shapes and objects. Transfer Learning is then a way to retrain an existing model to learn something new. It keeps the lower layers, that have learnt basic features, but then tells the higher layers how those low-level features should be combined in order to identify our new objects.





Not only does this dramatically reduce the amount of time needed to train a new model but also, when combined with masking, means that very few examples are needed in order to produce a model with a good level of accuracy.

Case Study: Natural disaster damage

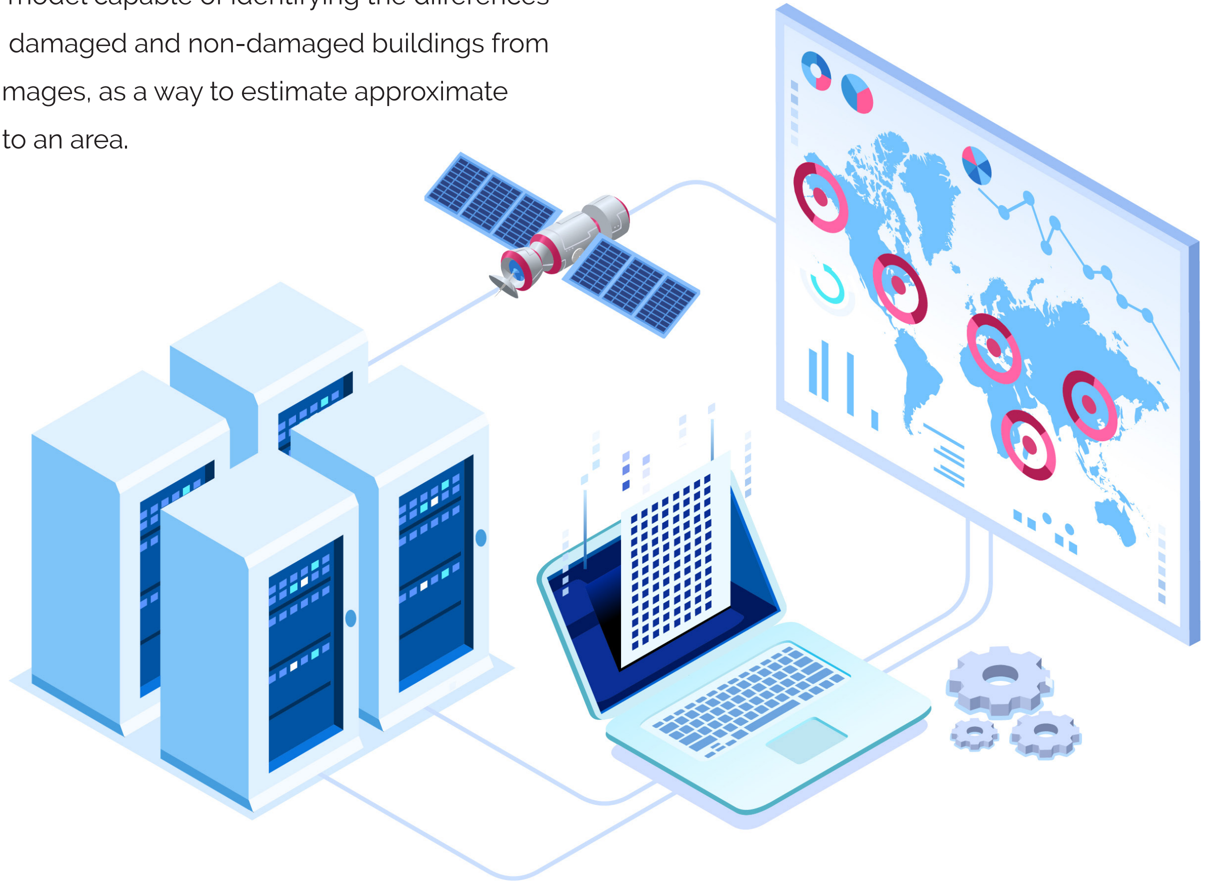
Background

The team at Zegami recently embarked on a project that utilised the techniques outlined above to develop a Machine Learning model which can be used to identify damage to buildings caused by natural disasters.

When a significant natural disaster occurs it can easily overwhelm first responders, as they do not know exactly where the damage is and which areas to prioritise. However, the first four days following such a disaster are the most important, time is of the essence and situational information is key.

During this timeframe, international awareness and interest are at their peaks and people are most willing to help. Satellite imagery can play an important role in helping to get a better understanding of the extent of the damage, as well as knowing where the resources and effort on offer will prove most effective.

To help with this, we decided to build a Machine Learning model capable of identifying the differences between damaged and non-damaged buildings from satellite images, as a way to estimate approximate damage to an area.



Process - utilisation of Zegami and Amethyst

The training data set was a single 100km² area over the coast of Dominica, hit by hurricane Irma in 2017. The satellite image of the area was divided up into hundreds of slices, called 'chips'.

On each chip, we annotated a few hundred buildings, split into two classes - 'damaged' and 'non-damaged', using our Amethyst annotation tool.

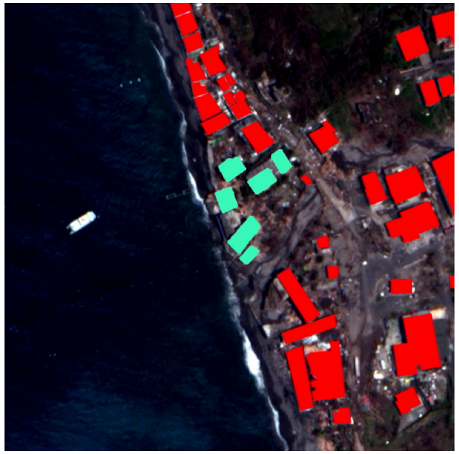
Once the labelling was complete, we used Mask R-CNN (a technique for instance segmentation) to train a new model based on a pre-trained model with ImageNet weights.

The training process took us around 20 minutes to do on a moderately powerful machine and GPU.

We then inferenced new images and cropped these out into a new Zegami collection, to help us better understand how well the model was performing on different building types.

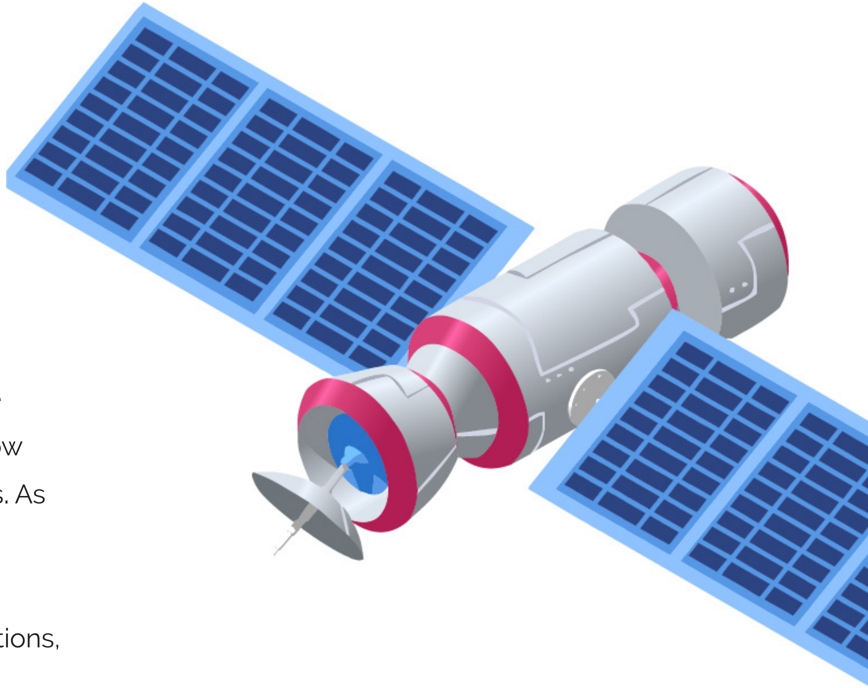
For this, we used a separate technique to extract features from each of the building instances, and reduced those vectors down to produce a 2D scatter plot of similarity.

As the image below shows, buildings are roughly grouped by colour and shape.



Using Zegami, we were then able to explore the similarities and get a better understanding of how well the model predicted each of the categories. As we have the confidence score of each of the buildings, we can add filters on low confidence instances and see if there are any mis-classifications, as well as looking at high confidence instances.
It is also useful in this situation to identify the ratio of damaged to undamaged buildings. This, when combined with the geospatial information, enables us to see where the greatest damage lies.

By doing this, we were able to rapidly iterate on the model by having a visual feedback loop into how it was performing.



Section 5

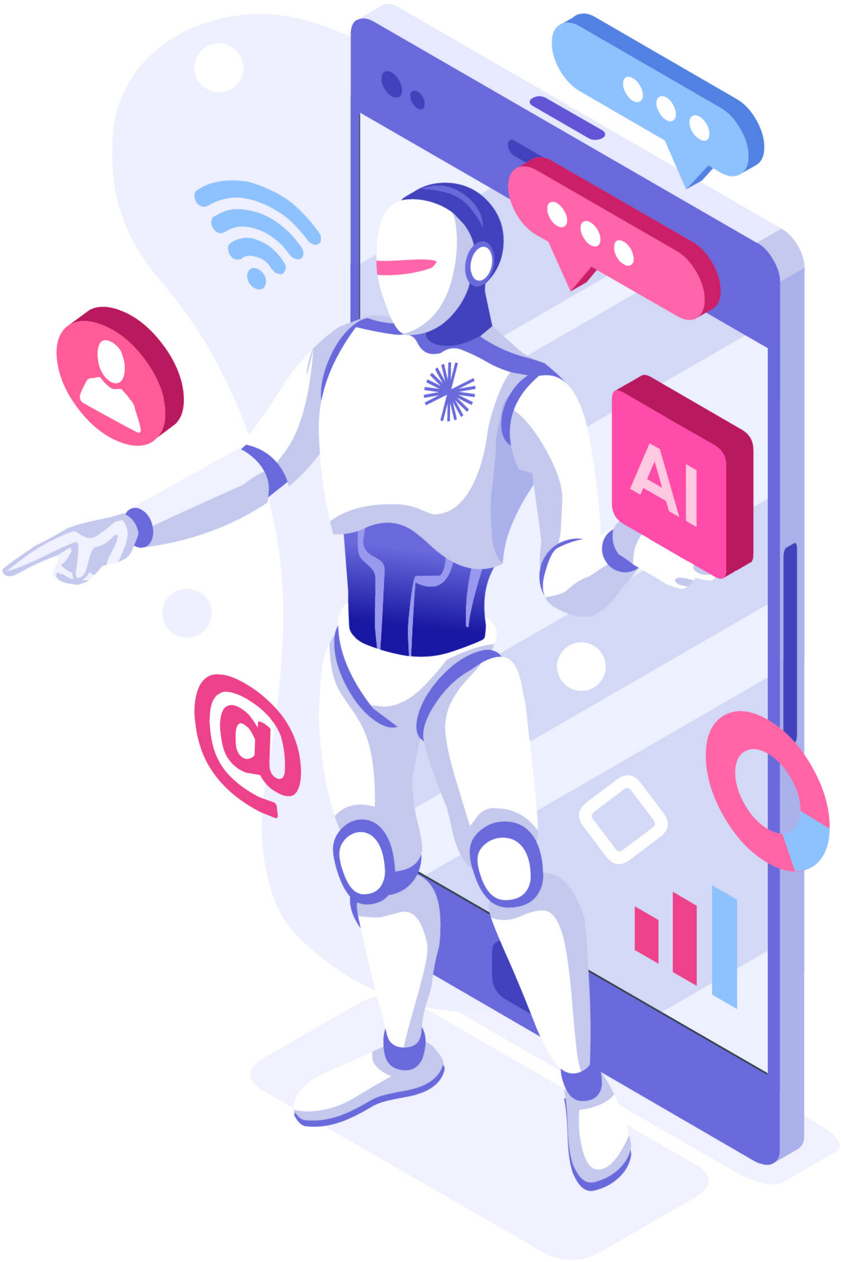
Summary

In summary, creating Machine Learning models is a non-trivial task. This is especially true when working with images, video or any other kind of unstructured data.

As in most situations, a model can only ever be as good as the data it has been trained on, placing a large burden on ensuring that the training set is well representative of real-world data.

To make things more difficult, biases are easy to introduce into a data set but notoriously difficult to spot. Data visualisation can play a key role in identifying biases, but traditional data visualisation tools struggle to function with unstructured data.

For this reason, a new breed of tools specifically designed to address this problem is needed. Zegami sets a marker in the development of such tools.



Learn more about Zegami at:

www.zegami.com