DRAFT Methodology:

\*\* Fix Data join code to remove the x & y occurences of “incident type”, then update EDA report. Will be linking github repo to my submission so want this to be cleaned up first. Don’t need to do any encoding because that only happens after split.

Explain steps taken to join the data into complete. Should anything that was already done be pushed to later after split of test/train?

**Splitting data:**

~~Source:~~ [~~https://towardsdatascience.com/time-based-cross-validation-d259b13d42b8~~](https://towardsdatascience.com/time-based-cross-validation-d259b13d42b8)

~~“When dealing with time-related and dynamically changing environments, where the characteristics of the environment change throughout time, it is best to use time-based splitting to provide statistically robust model evaluation and best simulate real-life scenarios.”~~

~~Instead of doing an 80/20 data split between training and test data, I will split my dataset based on time windows, with all observations from 2010 – 2020 going into my training data and all observations from 2021 going into my test data for the purposes of using 10 years of data to predict incident occurrences in 2021.~~

~~This article indicates:~~

~~“We need to pay attention to 3 important aspects:~~

~~1.~~**~~Time-based train\test split~~**~~- in each split, test indices must be higher than before.~~

~~2. We would like to choose our~~**~~train\test set sizes~~**~~in order to mimic real world scenarios in which we will train a model over some period and then apply it on the upcoming period. For example- train the model over last month data and apply it to predict on the upcoming week data.~~

~~3.~~**~~Dates matter~~**~~. For our intention, the number of records in each set does not matter. What matters is the size of the windows in terms of days. We would like to split the data so that each window will consist data from X days.”~~

~~“Scikit-learn has [TimeSeriesSplit](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html" \t "_blank) method, but it has several drawbacks. Assuming our data is sorted by time, this method splits it into train\test sets in a “sliding window” approach, but it doesn’t allow us to choose the sets sizes, we can only choose how many splits we would like to have. Scikit-learn TimeSeriesSplit also assumes that there is one observation per date, and therefore does not address 2 and 3 above.”~~

\*\* This is really interesting and if I wanted to predict for a specific year (say 2021) I could probably do it this way, but for the purposes of this project, (and because prof. abdou mentioned using cross – validation for the split, I think I can just ensure that with the cross-validation, the test data is always after the train data in time. This can be done with:

Dif. article

<https://towardsdatascience.com/how-to-correctly-perform-cross-validation-for-time-series-b083b869e42c>

“For time series, we always predict into the future. However, in the above approach we will be **training on data that is further in time than the evaluation test data**. This is data leakage and should be avoided at all costs.

To overcome this quandary, we need to ensure the **test set always has a higher index (the index is usually time for time series data) than the training set.**This means our test is always in the future compared to the data our model is fitted on.”

We’ll use the TimeSeriesSplit sklearn function to split and cross-validate.

**What to do for unbalanced data sets?** Our dependent variable – incident type – has a lot of highway incident types. Does this get taken into consideration when splitting? **ANY CORRECTION OF IMBALANCE MUST BE DONE ON TRAINING SET ONLY.**

<https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html>

“Under-sampling balances the dataset by reducing the size of the abundant class. This method is used when quantity of data is sufficient. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for further modelling.”

<https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18>

# “The Problem with Imbalanced Classes:

Most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce error.”

“As we saw above, accuracy is **not** the best metric to use when evaluating imbalanced datasets as it can be very misleading. Metrics that can provide better insight include:

* **Confusion Matrix:** a table showing correct predictions and types of incorrect predictions.
* **Precision:**the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier’s exactness. Low precision indicates a high number of false positives.
* **Recall:** the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier’s completeness. Low recall indicates a high number of false negatives.
* **F1: Score:** the weighted average of precision and recall.”

<https://towardsdatascience.com/how-to-handle-multiclass-imbalanced-data-say-no-to-smote-e9a7f393c310>

<https://towardsdatascience.com/guide-to-classification-on-imbalanced-datasets-d6653aa5fa23>

re oversampling: “As you may have suspected, there are some downsides to adding false data points. Firstly, you risk overfitting, especially if one does this for points that are noise — you end up exacerbating this noise by adding reinforced measurements. In addition, adding these values randomly can also contribute additional noise to our model.”

**“SMOTE (Synthetic minority oversampling technique)**

How does SMOTE work? SMOTE generates new samples in between existing data points based on their local density and their borders with the other class. Not only does it perform oversampling, but can subsequently use cleaning techniques (undersampling, more on this shortly) to remove redundancy in the end. Below is an illustration for how SMOTE works when studying class data.”

“Using **random oversampling** (with replacement) of the minority class has the effect of making the decision region for the minority class very specific. In a decision tree, it would cause a new split and often lead to overfitting. SMOTE’s **informed oversampling** generalizes the decision region for the minority class. As a result, larger and less specific regions are learned, thus, paying attention to minority class samples without causing overfitting.”

**“Drawbacks of SMOTE**

**Overgeneralization.** SMOTE’s procedure can be dangerous since it blindly generalizes the minority area without regard to the majority class. This strategy is particularly problematic in the case of highly skewed class distributions since, in such cases, the minority class is very sparse with respect to the majority class, thus resulting in a greater chance of class mixture.

**Inflexibility.**The number of synthetic samples generated by SMOTE is fixed in advance, thus not allowing for any flexibility in the re-balancing rate.”

“Another potential issue is that SMOTE might introduce the artificial minority class examples too deeply in the majority class space. This drawback can be resolved by hybridization: combining SMOTE with undersampling algorithms. One of the most famous of these is [**Tomek Links**](https://github.com/ojtwist/TomekLink). Tomek Links are pairs of instances of opposite classes who are their own nearest neighbors. In other words, they are pairs of opposing instances that are very close together.

Tomek’s algorithm looks for such pairs and removes the majority instance of the pair. The idea is to clarify the border between the minority and majority classes, making the minority region(s) more distinct. Scikit-learn has no built-in modules for doing this, though there are some independent packages (e.g., [TomekLink](https://github.com/ojtwist/TomekLink), [imbalanced-learn](https://imbalanced-learn.readthedocs.io/en/stable/index.html)).

Thus, Tomek’s algorithm is an undersampling technique that acts as a data cleaning method for SMOTE to regulate against redundancy. As you may have suspected, there are many additional undersampling techniques that can be combined with SMOTE to perform the same function. A comprehensive list of these functions can be found in the functions section of the [imbalanced-learn documentation](https://imbalanced-learn.readthedocs.io/en/stable/api.html#module-imblearn.ensemble).”

Because this is a multi-class classification problem:

<https://machinelearningmastery.com/multi-class-imbalanced-classification/>

## “Cost-Sensitive Learning for Multi-Class Classification

Most machine learning algorithms assume that all classes have an equal number of examples.

This is not the case in multi-class imbalanced classification. Algorithms can be modified to change the way learning is performed to bias towards those classes that have fewer examples in the training dataset. This is generally called cost-sensitive learning.”

The [RandomForestClassifier class](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) in scikit-learn supports cost-sensitive learning via the “class\_weight” argument.

“We can specify the “class\_weight” argument to the value “balanced” that will automatically calculates a class weighting that will ensure each class gets an equal weighting during the training of the model.”

<https://towardsdatascience.com/guide-to-classification-on-imbalanced-datasets-d6653aa5fa23>,

“Some authors claim that cost-sensitive learning is slightly more effective than random or directed over- or under-sampling, although all approaches are helpful, and directed oversampling, is close to cost-sensitive learning in efficacy. Personally, when I am working on a machine learning problem I will use cost-sensitive learning because it is much simpler to implement and communicate to individuals.”

I found at least 3 articles all making claims that cost-sensitive learning is superior to using over or under sampling techniques: <https://towardsdatascience.com/guide-to-classification-on-imbalanced-datasets-d6653aa5fa23>, <https://machinelearningmastery.com/multi-class-imbalanced-classification/>, <https://towardsdatascience.com/how-to-handle-multiclass-imbalanced-data-say-no-to-smote-e9a7f393c310>,

**Performance metric:**

<https://medium.com/p/7304aedf9037>

# “Step 4: Decide the Performance Metric for Classification Model

Before building the model, we need to decide the performance metric we would like to optimize towards.

The most critical performance metric for the rare events modeling is usually the minority class recall or precision values. For example, in the context of fraud detection, we would like to maximize the true positive rate and capture as many fraud cases as possible, so recall for the minority class is the metric we would like to optimize.

While in the context of spam email classification, we would like to minimize the false positive rate and not misclassify any important email as spam, so the precision for the minority class is the metric we would like to optimize.

In this tutorial, we use fraud detection as an example and choose recall for the minority class as the metric to optimize.”

For my project, I want to maximize the true positive rate and capture as many incident types as possible, so I will use recall as my main performance metric (as opposed to precision which would aim to minimize the false positive rate and not misclassify any incident types

1. Preprocessing:
   * Encoding nominal categorical variables using One Hot Encoding with scikit-learn. Choosing this method for encoding nominal categorical variables to ensure that we aren’t losing any information by having some incident numbers (data observations) including a list of categories and therefore not being able to view all data with a certain category or predict for certain categories as efficiently. Looking online and reading <https://towardsdatascience.com/stop-using-pandas-get-dummies-for-feature-encoding-5d2cd07cb4fc> helped me choose One Hot Encoding instead of get\_dummies as a method because get\_dummies can have issues where your test data set has a different set of columns depending on whether or not some categories exist in the test set compared to the training set. At what point in time do I split the dataset into test and training? Should this be done before encoding or after? (i.e. is encoding done separately on training and test sets or done before the split)???.
   * Will apply one hot encoding to just “response type” and “Activity type” for ALL models (to deal with the lists), but not do encoding on other categorical variables.
     + Single decision trees in H2O can handle categorical variables but scikit-learn requires one hot encoding of categorical variables (see <https://www.kaggle.com/code/gabrielaltay/categorical-variables-in-decision-trees> and <https://notebook.community/roaminsight/roamresearch/BlogPosts/Categorical_variables_in_tree_models/categorical_variables_post>

In my research I came across the following quote and decided it would be worth it to do a comparison of my own between models that use one hot encoding on categorical variables and one that doesn’t. I also want to compare performance between a decision tree classifier and a random forest classifier. The 3 machine learning models I will be generating and comparing are listed after the quote below.

Quote from notebook article: “But one-hot encoding also presents two problems that are more particular to tree-based models:

1. The resulting sparsity virtually ensures that continuous variables are assigned higher feature importance.
2. A single level of a categorical variable must meet a very high bar in order to be selected for splitting early in the tree building. This can degrade predictive performance.

This post substantiates both of these points with a comparison between [scikit-learn](http://scikit-learn.org/), which presupposes one-hot encoding, and [H2O](http://www.h2o.ai/), which does not.”

* + - My 3 machine learning models to compare performance (what specific performance metric am I interested in?) on will be:
      * Decision tree classifier using scikit-learn – data will have one hot encoding of ALL categorical variables AND using sklearn.utils.class\_weight function to address the imbalanced classes among the dependent variables by balance the class weights (how: <https://towardsdatascience.com/guide-to-classification-on-imbalanced-datasets-d6653aa5fa23> and <https://machinelearningmastery.com/multi-class-imbalanced-classification/>).
      * Random Forest classifier using scikit-learn – data will have one hot encoding applied to ALL categorical variables. AND using sklearn.utils.class\_weight function to address the imbalanced classesamong the dependent variables by balance the class weights
      * Random Forest classifier algorithm using H20– data will maintain categorical variables (except for “response type” and “activity type” which need to be encoded because some observations in this attributes contain a list of categories). AND will use the balance\_classes option to address the imbalanced classes among dependent variable. (how: <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html>).

What attribute selection indicator will I use for these models? 🡪 information gain, gain ratio, and Gini index for each attribute.

Quote from notebook article: “But one-hot encoding also presents two problems that are more particular to tree-based models:

1. The resulting sparsity virtually ensures that continuous variables are assigned higher feature importance.
2. A single level of a categorical variable must meet a very high bar in order to be selected for splitting early in the tree building. This can degrade predictive performance.

This post substantiates both of these points with a comparison between [scikit-learn](http://scikit-learn.org/), which presupposes one-hot encoding, and [H2O](http://www.h2o.ai/), which does not.”

Encoding should only be done AFTER the split of train/test data to avoid data leakage (bias to the training set because the encoding could contain information about data contained only in test set).

* + Feature scaling/data normalization? What should be done for my categorical variables (issue with some being lists…) – referring to: <https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/>
    - In order to use kNNImputer, we have to encode categorical values into numerical values using mapping (<https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/#:~:text=For%20imputing%20missing%20values%20in,of%20categories%20to%20numeric%20variables>.).
  + How will I handle missing values and or special characters? Some columns can likely be deleted?
    - Variables with only a few (between 3 – 40) missing values: Incident Type, Species common name, Sum of Number of Animals, Activity Type, Latitude Public, Longitude Public, Within Park,
      * Impute this data using k-th Nearest Neighbour Imputation (kNNImputer) (referred to this source to decide: <https://arthurarchiproj.medium.com/classification-missing-data-imputation-2932166e1000>).
    - Variables with around 1500 missing values: Response Type,
      * 1500 missing values is still only about 2% data for this column so I will impute this data (using k-th nearest neighbour imputation) Note that the “Response Type” variable contains some values that are lists (from merging data for Incident ID’s prior to dataset join) which may need to be considered when imputing.
    - Variables with several missing values (over 30,000): Animal Health Status, Cause of Animal Health Status, Animal Behaviour, Reason for Animal Behaviour, Animal Attractant, Deterrents Used, Animal response to Deterrents.
      * There are too many missing values to consider any kind of imputing. But I don’t want to completely lose this data as it would be good to find patterns for when specific animal health status’ occur, for example when an animal is dead. I would like to subset the data (all rows that have complete values for these variables) variables and analyze it separately from the rest of the data (all rows but delete the specified columns with missing values)
  + Data reduction? – use dimensionalty reduction via random forest to select which variables are most important or least important for the dataset. Need to look into this more.
  + Feature selection:
    - Factor Analysis of Mixed Data (FAMD):
      * “When dealing with mixed data, FAMD is a recommended way to handle the unnecessary factor and reduce the dimensions of the data.” 🡪 <https://towardsdatascience.com/factor-analysis-of-mixed-data-5ad5ce98663c>

1. Exploratory analysis:
   * Univariate: Use frequency distribution tables/histograms to look at variance within each of the variables. Can look at some summary statistics for each variable (depending on data type, these will vary but some possible are : mean, mode, median, range, variance, maximum, minimum, quartiles, and standard deviation.)
   * Look for outliers in categorical data using the histogram (<https://analyticsindiamag.com/how-to-detect-and-treat-outliers-in-categorical-data/#:~:text=or%20box%20plot.-,Detecting%20outliers%20in%20the%20categorical%20data%20is%20something%20about%20the,the%20bar%20chart%20or%20histogram>.)
   * multivariate Use EDA report (generated using Pandas ProfileReport) to look at correlations between variables.

**~~Looking at trends by geographical area?~~**

* ~~Use data classifiers to determine what factors occur in what geographical areas.~~ 
  + ~~Clustering latitude & longitude and creating a heat map for each of 35 parks of incident frequency by location.~~
* ~~Which category in this particular area - data mining and association rules – can only be applied.~~

As part of methodology, look up to see who else used this dataset before and what techniques and tools did they use. Check the efficiency of their classifiers. Check conclusions/recommendations that Parks Canada may have developed based on this data.

Multinomial logistic regression (using scikit-learn in Python), can be used as an alternative prediction model (instead of decision tree) if desired. I don’t think this is needed as I already have 3 models to compare. Also, the article at: <https://www.datasklr.com/logistic-regression/multinomial-logistic-regression> describes “the method remains relatively unpopular because it is difficult to interpret and it tends to be inferior to other models when accuracy is the ultimate goal”

Add details about WHY I am choosing each method I choose.

**Literature Review – answer following questions:**

* What do you already know about the topic?
* What do you have to say critically about what is already known?
* Has anyone else ever done anything exactly the same?
* Has anyone else done anything that is related?
* Where does your work fit in with what has gone before?
* Why is your research worth doing in the light of what has already been done?

Similar research and what methodologies they used…

1.

Human–wildlife coexistence in a changing world: <https://www.zoology.ubc.ca/conservation/wp-content/uploads/2021/04/Konig-et-al-2020.pdf>

This one refers to HWC as human wildlife CONFLICTS and emphasizes the importance of looking at the negative interactions (which are considered conflicts). Could be useful for backing up my research question around looking specifically at most important incidents as being the most severe with human or animal injury.

**Quotes**:

* “Human–wildlife conflict (HWC) is a key topic in conservation and agricultural research. Decision makers need evidence-based information to design sustainable management plans and policy instruments”
* “Building on current literature and articles in this section, we developed a conceptual model to help frame HWC and coexistence dimensions. The framework can be used to determine damage prevention implementation levels and approaches to HWC resolution.”
* “Human–wildlife conflicts (HWCs) are common near agricultural and other production landscapes, such as urban and peri-urban areas or near protected areas (PAs). Human-wildlife conflicts is defined as interactions between wildlife humans with a negative outcome (Madden 2004)”
* “HWC is one of the most complex and urgent issues facing wildlife management and conservation (Frank et al. 2019), especially outside PAs (Woodroffe et al. 2005). Scholars are seeking ways to refocus policy-relevant conflict research on finding pathways toward human–wildlife coexistence (Marchini et al. 2019) and coadaptation (Carter & Linnell 2016).”
* “Coexistence is defined as a dynamic but sustainable state in which humans and wildlife coadapt to living in shared landscapes, where human interactions with wildlife are governed by effective institutions that ensure long-term wildlife population persistence, social legitimacy, and tolerable levels of risk (Carter & Linnell 2016).

Looking at similar research questions as me but at the next level – this is not a data analysis paper but rather a paper after the data analysis was done and working to develop a framework for how to deal with the analysis.

2.

# Frank B, Glikman JA, Marchini S. 2019. Human–wildlife interactions: turning conflict into coexistence. Cambridge University Press, Cambridge, United Kingdom

Book that I was able to access online via TMU online library at the following link:

<https://www-cambridge-org.ezproxy.lib.torontomu.ca/core/books/humanwildlife-interactions/planning-for-coexistence-in-a-complex-humandominated-world/6A6C2215C46898E3E9123E246D439E63>

Quotes below (from chapter 19 of the book specifically), talk about importance of this research. I see the data analysis phase as being step 1 for conservation planning referred to below as “situation assessment”.

**19 - Planning for Coexistence in a Complex Human-Dominated World**

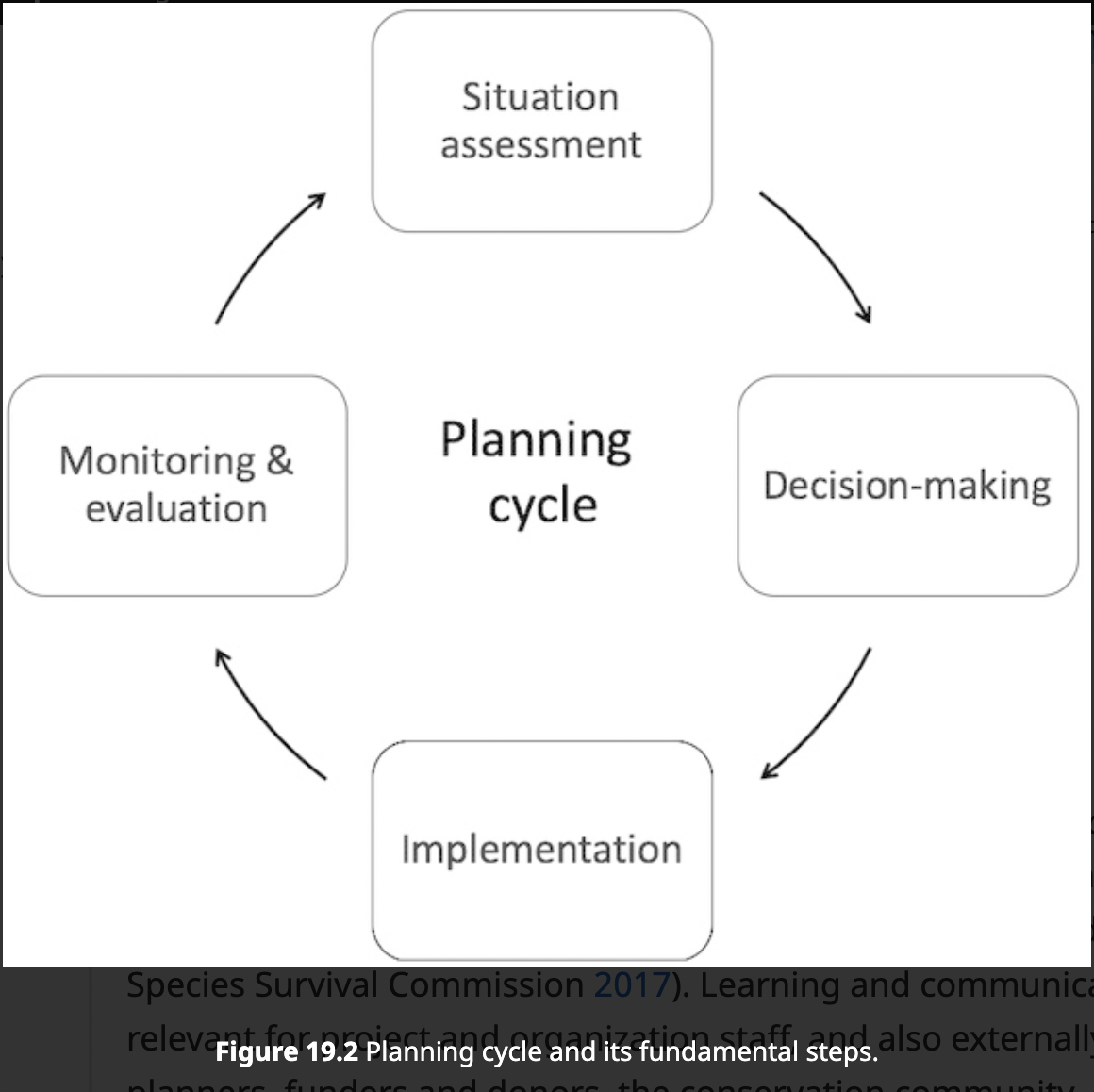
Quotes:

* “We argue that the bridge between HWC research and the implementation of large-scale human–wildlife coexistence is good planning. In this chapter, we address the potential application of strategic planning, combined with the growing fields of scientific modelling and data science, to inform decisions regarding the conflict-to-coexistence continuum (Frank [Reference Frank2016](https://www-cambridge-org.ezproxy.lib.torontomu.ca/core/books/humanwildlife-interactions/planning-for-coexistence-in-a-complex-humandominated-world/6A6C2215C46898E3E9123E246D439E63#REFe-r-1138)) and propose a framework for integrating data and stakeholders – planners, researchers, modellers, policy-makers, managers and citizens – in the process of planning for coexistence.”

### 19.1 What Is Planning and Why We Need it

### Planning is the process of identifying a course of action in a systematic manner to achieve objectives by utilizing the available resources competently in a cost-effective way (Mintzberg & Quinn [Reference Mintzberg and Quinn1996](https://www-cambridge-org.ezproxy.lib.torontomu.ca/core/books/humanwildlife-interactions/planning-for-coexistence-in-a-complex-humandominated-world/6A6C2215C46898E3E9123E246D439E63#REFe-r-1174)). Planning is often depicted as an iterative, cyclic process comprised of the following fundamental steps ([Figure 19.2](https://www-cambridge-org.ezproxy.lib.torontomu.ca/core/books/humanwildlife-interactions/planning-for-coexistence-in-a-complex-humandominated-world/6A6C2215C46898E3E9123E246D439E63#FIG-fig-47)):

* Situation assessment, where a description and understanding of the current situation is developed by addressing the questions where are we and why are we here?
* Decision-making, which involves establishing what the project will aim to achieve, defining the agreed vision and short-term activities that must be completed to ensure that longer-term goals are met, and determining what needs to be done to achieve the desired results, including how results will be monitored and evaluated. The guiding questions are where do we want to get and how do we get there?
* Implementation, which involves putting into practice the previous planning work through the development and implementation of specific work plans while ensuring sufficient resources, capacity and partners.
* Monitoring and evaluation, where data collected during and after implementation are analysed in order to measure success, usually at different levels (e.g. output, outcome and impact). The guiding questions are are we getting there? and have we got there? At this step, revisiting the decision-making and implementation steps should be considered, thus closing the cycle.



### “Species conservation planning is intended to generate a blueprint for saving a species or group of species, across all or part of the species’ range.”

3.

Human-Wildlife Coexistence: Recommendations for Improving Human-Wildlife Coexistence in the Bow Valley: <https://banff.ca/DocumentCenter/View/5520/Human-Wildlife-Coexistence-Bow-Valley-Report>

This report discusses current issues/trends with human and wildlife coexistence and plans to address it. The area of Bow Valley includes Banff National Park and Parks Canada technical experts were involved in the working group who put together the report. That said, there is no indication the that same dataset was used in the creation of this report and this report does not discuss the data source directly or how it was analyzed – only information regarding data is: “Data resources were compiled from the following agencies: • Parks Canada; • Government of Alberta; • Town of Banff; • Town of Canmore; and • WildSmart.”

This paper could be useful to refer to when making recommendations post analysis.

4.

Wildlife-Human Interactions in National Parks in Canada and the USA, Dr. Alistair J. Bath, Memorial University of Newfoundland Dr. Jody W. Enck, Cornell University, SOCIAL SCIENCE RESEARCH REVIEW: <http://npshistory.com/publications/wildlife/ssrr-v4n1.pdf>

This paper could be useful to refer to when making recommendations post analysis.

Quote:

* “The chance to view wildlife draws millions of visitors each year to the national parks of North America. The combination of a large number of people and abundant wildlife leads to a variety of wildlife-human interactions. In this paper we explore the nature of those wildlife-human interactions, theoretical frameworks social scientists are using to understand those interactions, and approaches used by national parks across North America to manage those interactions.”

5.

# Parks Canada working to reduce growing human-wildlife conflicts, Cathy Ellis, Rocky Mountain Outlook: <https://www.rmoutlook.com/banff/parks-canada-working-to-reduce-growing-human-wildlife-conflicts-5920752>

# Article about “Parks Canada’s new management plan for Banff National Park commits to reducing the number of human-wildlife conflicts within the park in two years.”

#### **6.**

**Email from David Gummer, Wildlife Management Specialist, Natural Resource Management Branch  
Parks Canada / Government of Canada:**

“On your second question, the main way that these data have been reported and currently affect policies, programs and operations is that we have built a live internal dashboard using PowerBI so that staff and management of the sites can refer to, summarize, analyze and export the (dynamic) incident data as needed to support day to day decisions for their operations (delivery of human-wildlife coexistence operational programs – teams that respond to HWC incidents and take management actions to avoid/reduce future conflict incidents). We are interested and pursuing the possibility of providing a similar interactive data tool that could be accessed externally or even promoted as a public tool, but have not gotten there yet—so in the meantime our effort was to at least publish the same underlying data for others to use until we have a better way to share a more efficient/functional interactive. I don’t know how to share a functional version of our dashboard with you currently, but if you would like more info on this, please let me know. Perhaps we could come-up with a good way to share some views with you.

These data and our ongoing analyses are also helping to inform new national policy/guidance that we are working on and many more that we propose for future, however there is not yet already one good stand alone example that I can send your way. This is why that sentence you quoted is written more generally and in the present tense. In addition, individual parks and sites may use their subsets of the data in their own local reports, policies or protocols/procedures. I do not have a ready example at hand for those either, but if there is a certain park, region or topic you are interested in, that could help me investigate if there are relevant examples or who might be a good contact I could refer you to.

***NOTE****:* ***Citing PERSONAL COMMUNICATION*** *Any communication that cannot be directly retrieved by a reader is considered “personal communication.” Emails, phone conversations, text messages, and social media messages are all examples of personal communication. You do not include personal communication in your reference list; instead, parenthetically cite the communicator's name, the phrase "personal communication," and the date of the communication in your main text only.* [*https://owl.purdue.edu/owl/research\_and\_citation/apa\_style/apa\_formatting\_and\_style\_guide/reference\_list\_other\_non\_print\_sources.html#:~:text=You%20do%20not%20include%20personal,%2C%20January%204%2C%202019*](https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/reference_list_other_non_print_sources.html#:~:text=You%20do%20not%20include%20personal,%2C%20January%204%2C%202019)*).*

# 

7.

**Madden F. 2004. Creating coexistence between humans and wildlife: global perspectives on local efforts to address human–wildlife conflict. Human Dimensions of Wildlife** (An International Journal)**, Volume 9, pages 247–257.**

<https://www-tandfonline-com.ezproxy.lib.torontomu.ca/doi/full/10.1080/10871200490505675>

I’ve seen this article referenced in a few of the other sources I list above (at least 2), and it has a good framing of the “HWC” problem.

## “The Problem of Human–Wildlife Conflict

“Human-wildlife conflict occurs when the needs and behavior of wildlife impact negatively on the goals of humans or when the goals of humans negatively impact the needs of wildlife. These conflicts may result when wildlife damage crops, injure or kill domestic animals, threaten or kill people” (WPC Recommendation, this issue). Such conflict may occur because a lion has attacked someone’s livestock or a gorilla has raided a person’s crops. The conflict also occurs when a person or community seeks to kill the lion or gorilla, or when a person retaliates against the authorities that are in charge of conserving wildlife and its habitat.

HWC escalates when local people feel that the needs or values of wildlife are given priority over their own needs, or when local institutions and people are inadequately empowered to deal with the conflict. If protected area authorities fail to address the needs of the local people or to work with them to address such conflict adequately, the conflict intensifies, becoming not only conflict between humans and wildlife, but also *between* humans *about* wildlife. Frequently, wildlife conservation initiatives suffer, the economic and social well-being of local people is impaired, local support for conservation declines, and conservation and development efforts meant to offset more general “costs” of living near a protected area may be impeded.

Human–wildlife conflict is increasing in both frequency and severity worldwide and will likely continue to escalate. Protected areas are increasingly becoming islands of habitat surrounded by seas of cultivation and development. Wildlife and humans increasingly compete for space, resources, and places to call home. Although ecosystem-based approaches (including the development of corridors between protected areas) offer improved long-term protection for many species from a biological perspective, they also involve extensive regional opportunities for interaction and conflict between local people and wildlife. Without properly addressing HWC in the effort to conserve wildlife and their habitat, conservation efforts will lose stability and progress, as well as the support of local communities.

Human–wildlife conflict, as we understand it today, is not always inevitable and has not been the norm in all cultures and communities. In some communities and cultures, evidence of human–wildlife co-evolution and cultural tolerance to wildlife may offer clues as to how coexistence can be achieved elsewhere.”

8.

Source describing actual data analytic process:

# Characterization and management of human-wildlife conflicts in mid-hills outside protected areas of Gandaki province, Nepal

* Kedar Baral ,
* Hari P. Sharma,
* Bhagawat Rimal,
* Khum Thapa-Magar,
* Rameshwar Bhattarai,
* Ripu M. Kunwar ,
* Achyut Aryal,
* Weihong Ji

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260307>

**What?:** “With the intent to better management human wildlife conflict (HWC) and wildlife conservation in mid-hills outside protected areas of Gandaki province, Nepal, we analyzed the patterns and drivers of HWC.”

**Data Analysis Methods**:

“We conducted multivariate logistic regression with the entire independent variables [total of 11 variables: temporal (year, month, season, time), socioeconomic (gender, occupation and age group of the victims), and spatial (district, elevation, distance between point of attack and forest), and land use (forest, agricultural land, road and settlement)] in the model to understand the relationship between predictive and explanatory variable (human death and injury due to wildlife attack).”

“We conducted two different statistical approaches following Naha et al. [[52](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260307#pone.0260307.ref052)]. Chi-square test of independence was used to understand the association between temporal (year, month, season and time), and socio-demographic variables (gender, age and occupation) with the wildlife attacks. We classified victims into four age groups, < 20, 21–40, 41–60 and > 61 years. The association between socio-demographic and temporal variables and the attacks were analyzed using Pearson chi-square test ([Table 1](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260307#pone-0260307-t001)). For the spatial dataset, we conducted a generalized linear model with binomial distribution to predict the effect of variables on the wildlife attacks following Acharya et al. [[6](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260307#pone.0260307.ref006)]. We used a priori candidate model and ranked them based on Akaike Information Criterion (AIC) values. Those models with lowest AIC values were considered the appropriate for explaining the wildlife attacks ([Table 2](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260307#pone-0260307-t002)). A model with a ΔAIC (the difference between the two AIC values being compared) of ≤ 2 is considered significantly better than the model it is being compared to. The number of attacks per year was summarized in terms of mean (M) and number of events with regard to different spatial, temporal and socio economic variables were analyzed by using percentage. The variability was recorded in terms of SE and at confidence interval of 95%. Statistical analyses were undertaken in R statistical software (R Core Team 2016; Version 1.0.44).”

9.

Source describing actual data analytic process:

Landscape predictors of human–leopard conficts within multi‑use areas of the Himalayan region Dipanjan Naha1 , Suraj Kumar Dash1 , Abhisek Chettri1 , Pooja Chaudhary1 , Gaurav Sonker1 , Marco Heurich2 , Gopal Singh Rawat1 & Sambandam Sathyakumar1\*

<https://link.springer.com/content/pdf/10.1038/s41598-020-67980-w.pdf>

“Conflict with humans is a signifcant source of mortality for large carnivores globally.With rapid loss of forest cover and anthropogenic impacts on their habitats, large carnivores are forced to occupy multiuse landscapes outside protected areas.We investigated 857 attacks on livestock in eastern Himalaya and 375 attacks in western Himalaya by leopards between 2015 and 2018. Multivariate analyses were conducted to identify the landscape features which increased the probability of livestock depredation by leopards.The risk of a leopard killing livestock increased within a heterogeneous landscape matrix comprising of both closed and open habitats (very dense forests, moderate dense forests, open forests, scrubland and non-forests).We used the results to map potential human–leopard confict hotspots across parts of the Indian Himalayan region. Our spatial risk maps indicate pockets in the eastern, central and western part of eastern Himalaya and the central, northern part of western Himalaya as hotspots of human–leopard conficts. Most of the attacks occurred when livestock were grazing freely within multi-use areas without supervision of a herder. Our results suggest that awareness about high risk areas, supervised grazing, and removing vegetation cover around human settlements should be initiated to reduce predation by leopards.”

Data analysis:

“We were interested in examining broader seasonal patterns of depredation (summer, monsoon and winter) and not just for individual months, hence each year was divided into 3 seasons of 4 months each (winter—November–February, summer i.e. February–June, monsoon i.e. June–November). We examined seasonal and temporal variation of attacks and diference in habitat types within the vicinity of predation sites using the chi-square test in R version 3.4.0. Statistical signifcance was p≤0.05 for all analyses. All spatial analyses were performed with Arc GIS 10.3.3 and R 3.4.0.”

“To model the spatial spread and extent of livestock depredation we prepared both binary (presence-1 cells with at least one or multiple attacks, absence-0-no attacks) and count data (presence—exact number of attacks recorded and absence 0-no attacks).”

“Data preparation for spatial risk analysis. We **identified a total of 5 major landscape features** (Habitat, Water, Human presence and infrastructure, Distance to Protected Reserves and Altitude) for North Bengal and Pauri Garhwal (Table 3) based on their ecological importance to model predation risk.”

Should I group latitude/longitude values into X bins of areas within each park 🡪 group based on closeness so all latitude/longitude values within (for example) 5 km of each other (of centroid) are grouped together and then can be classified as high risk, med risk, or low risk areas based on number of incidents or specific types of incidents. I think this is outside the scope of my project 🡪 I would likely need to add more data so that I can capture the entire parks area and not just space where incidents happened. This project is more about predicting causes of certain incidents and or which incidents occur most in which parks, etc. I don’t need to do everything here.

“We used 4 analytical approaches to model probability of livestock depredation by leopard. In the 1st step we evaluated spatial autocorrelation among livestock kills within the cells using function moran.test (Moran’s I) in package (spdep)75 in R 3.4.0. In the 2nd and 3rd approach we used generalized linear models (GLMs) with binomial and poisson structures to quantify the efect of landscape features (area of habitat types, availability of water, human presence and infrastructure, distance to protected areas and altitude) on livestock predation. All the predictor variables considered for the analysis were continuous in nature. We used a priori candidate models and ranked them based on AIC, AICc values76. Models with the lowest AICc values were considered the best or dominant model and the output (coefcients and estimates) explained the probability of livestock predation by leopards within IHR. Based on the results of the dominant model or the model averaged coefcients, we generated confict hotspots for both the study sites. We used coefcients of the best model (binomial structure) or averaged all candidate models (GLM with binomial structure) to estimate probability of livestock depredation for each cell (25 km2 ) using the equation p (x)=exp (z)/ (1+exp (z)) and generated human–leopard confict hotspots in Arc GIS 10.377. We generated ROC curve and AUC values to predict reliability of the dominant models using package ROCR78 in R 3.4.0. Since predictor variables between the two study sites were not normally distributed, we compared the identical landscape features using nonparametric Wilcoxon Signed-Rank Test in R 3.4.0. In the 4th step, we used the predictor variables of the dominant models to calculate conditional inference (CTREEs), as prescribed in the R-package “partykit”79. Tis method was adopted to obtain threshold values for the signifcant variables for confict mitigation recommendations. Trees based on maximally selected rank statistics were ftted using the Bonferroni correction for multiple testing and a minimum sum of weights. In addition, univariate trees were ftted for variables with a signifcant split in the multivariate tree. Te results of our two analytical approaches (regression and ctree) are similar and provide an overall understanding of landscape features prone to livestock predation in accordance with the behavior of common leopards. Both analytical methods are based on a maximum likelihood approach and when interpreted together provide meaningful results. Te GLM models computes probabilities of an event based on a logistic regression framework while the CTREE uses a machine learning classifcation approach and assigns values to predefned categories. Te decision tree approach is a non-parametric approach which helps simplify complex relationships between dependent and predictor variables.”

10.

I’m not super interested in this one – they focus a lot on the discrepancy between conflicts being reported for deprecation of livestock based on an animal when in fact the animal was not involved. Not super relevant to my project.

# A Framework for Estimating Human-Wildlife Conflict Probabilities Conditional on Species Occupancy

<https://www.frontiersin.org/articles/10.3389/fcosc.2021.679028/full>

“Managing human-wildlife conflicts (HWCs) is an important conservation objective for the many terrestrial landscapes dominated by humans. Forecasting where future conflicts are likely to occur and assessing risks to lives and livelihoods posed by wildlife are central to informing HWC management strategies.”

“We present a Bayesian hierarchical modeling framework that integrates conflict reporting data and species distribution data, thus allowing the estimation of the probability that conflicts with a species are reported from a site, conditional on the species being present. In doing so, our model corrects for both false-positive and false-negative conflict reporting errors. ”