# Project Phoenix – Wildfire Prediction in Canada

Zakaria El Hjouji, Louis Lecluse
Dimitris Bertsimas, Lea Kapelevich, Carine Simon
Christian Klose, Nataliya Le Vine

New York, NY

## Problem

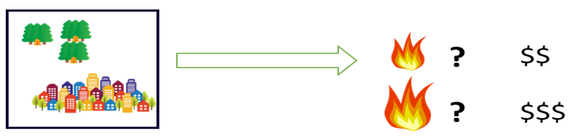
**Intro:** Wildfires are very rare and costly events. As of today, wildfires have cost the (re)insurance industry billions of dollars. For example, Fort McMurray's fire in 2016 is expected to cost more than $9 billion. While some people think that such events are one-off events, others believe that there are common atmospheric and geographic patterns that lead up to wildfires.

**Project Statement:** In this project, we hope to harness the power of Machine Learning and Artificial Intelligence to recognize those patterns. Our goal is to understand the risk of wildfires for any region in Canada in time and space through predictive modelling.

Our model can be broken down into two:

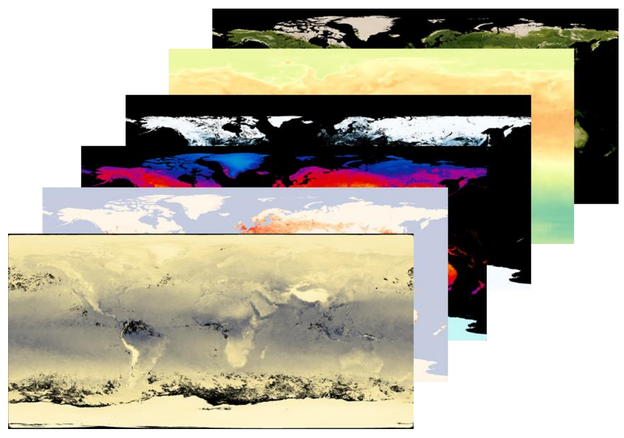
- **Fire Occurrence Model:** For each location *(x,y),* this model predicts whether such location will experience a fire in one month, two months, … , up to fifteen months.

- **Fire Severity Model** : For each location *(x,y),* this model predicts the size of the fire such location might have in one month, two months, … , up to fifteen months.




Our Data is Heterogeneous in the following ways:

- **Different Sources:** Our data comes from different sources such as NASA Earth Science, Swiss Re's proprietary data and other publicly available data.
- **Different Time & Space scale:** Our data comes in different scales. For instance, some features are at 0.1 degree scale *(10 km),* while others at 1 degree scale. Features also cover different timespans.
- **Different Forms:** Our data comes in both Structured and Unstructured Format (*i.e. Satellite Images).*



## Data

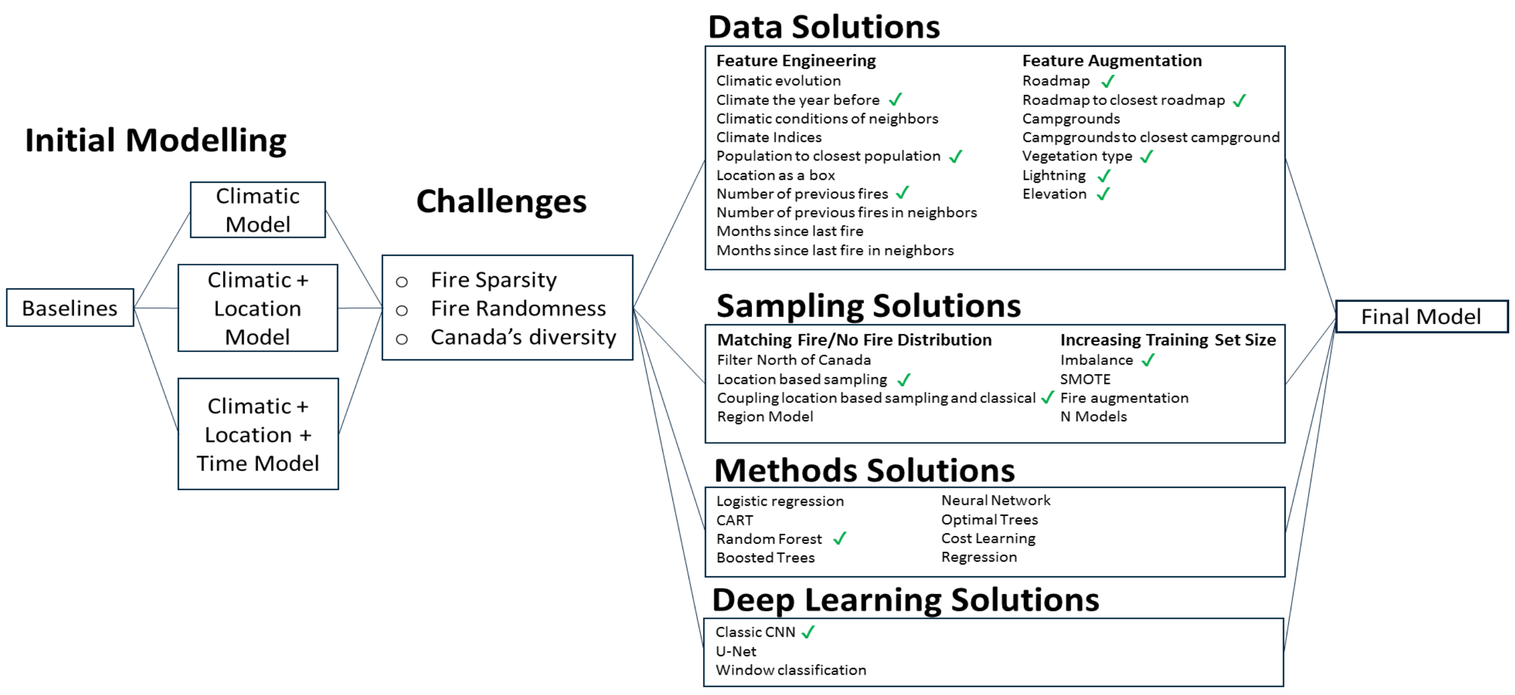
Our Data encompasses major wildfire predictors. They can be broken down into four different categories:

- **Climatic features:** Such features are important as they allow the model to capture climatic patterns under which wildfires occur. For example, wildfires occur frequently in dry areas with high Surface Temperature.
- **Geographical Features:** Wildfires occur under specific geographical settings. For instance, wildfires occur in places with high vegetation and low elevation.
- **Sources of Ignition:** These features help the model capture some of the randomness that triggers fires. For example, in June 2018, lightning sparked nearly 100 wildfires in British Columbia in 24 hours. Hence, taking into account the lightning activity in each region is key
- **Fire History:** Some areas might have high wildfire activity, however, our features are unable to set such regions apart. Using the history of fires as a feature allow the model to form a prior about this region's risk.

| Climatic | Geo | Sources of Ignition | Fire History |
|---|---|---|---|
| Temperature | Elevation | Lightning | Number of Past Fires |
| Wind speed and Direction | Vegetation Index | Campground | Month Since last Fire |
| Drought Index | Vegetation Type | Roadmaps | |
| Water Vapor | Snow Cover | | |
| Net Radiation | | | |

## Modelling

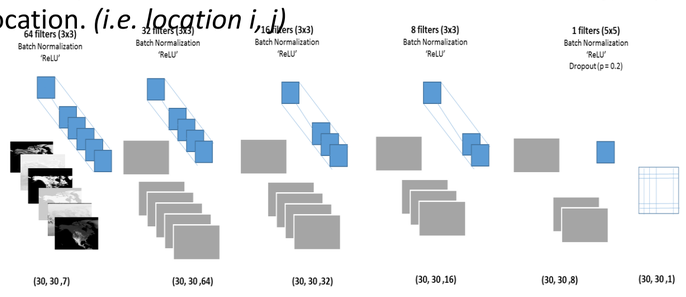
There are many challenges to our problem, chiefly: data imbalance (0.1% fires), wildfires can be random and skewness of fire sizes. To overcome those challenges, we explored different modelling approaches. We started with strong baselines and initial modelling attempts providing us with insights and performance references. We then increased our performance by closely exploring our features and varying our sampling methods and modelling techniques.

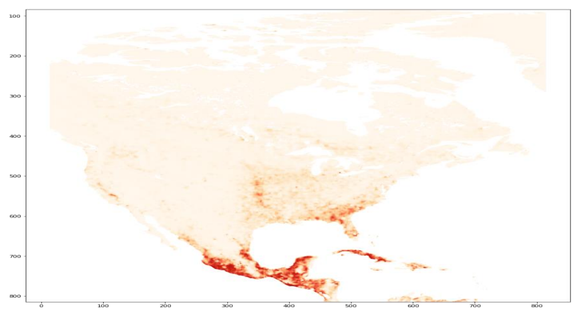



## Deep Learning

**Motivation:** Random Forest and Structured Data models are sometimes unable to capture complex patterns, mainly when it comes to spatial correlations. Also, given the nature of our data (i.e. satellite images) and the recent success of Deep Learning in computer vision, we believe that it is important to explore such models.

**Models:** We explored various models and architectures, spanning from Classical CNN to semantic segmentation architectures *(e.g. U-net, TernausNet, etc..*). The model that delivered the best out-of-sample performance is a CNN that takes as input a 3D-matrix (30x30) with 7 channels (each representing a different feature), and passes it through a series of convolutions with same padding and outputs a 2D-matrix such that each element *(i, j)* represents the conditional probability of having fire in the corresponding location. *(i.e. location i, j).*



**Performance**: This model was trained on North America data and delivered the best performance. The Average Precision score was 34% with a recall of 81%. Unlike the structured data model, this model can be easily scaled to the global scale.

## Impact

When an underwriter needs to understand the risk associated with wildfire for a particular region, they use Classic (probabilistic) models. However, such models are based primarily on wildfire history in the region, which becomes cumbersome when such data is not readily available. In addition to that, such models provide static risk scores and cover regions at a macro-level, which does not allow underwriters to build risk scores at a granular-level, or asset-level. Our model uses state-of-the art Machine Learning methods to help underwriters build a **forward-looking** view of the wildfire risk on a monthly basis and at a micro region *(10x10km)*.

**Loss Frequency Curve:** Loss Frequency curves depict the distribution of area burnt by wildfires on a particular region. Using our model with distribution fitting techniques, one can develop such curves at a pixel- and monthly-level. These can then be aggregated to cover larger regions and time periods.

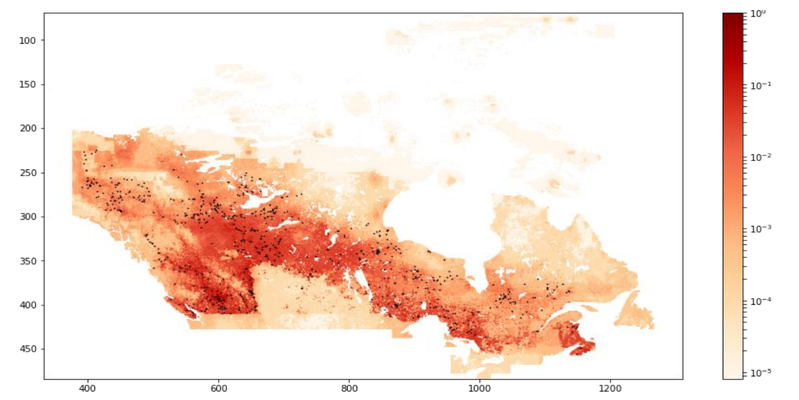
**Fort Mc-Murray:** Our model accurately detects the 2016 Fort McMurray event: it predicts a 2% increase in hazard for the May-June-July 2016 period with respects to 2015 levels.

## Best Occurrence Model

Through our modelling journey, we identified the key features, the model architecture (Random Forest) and sampling methods (Imbalance and location-based sampling) that yielded the best out-of-sample performance for the occurrence model. Below are the features selected.

**Performance:** This model has the best Average Performance Score: 10% (baseline: 3%) with a recall of 88%. Its performance remain strong as we predict further in the future. It is able to predict with good performance 15 months into the future.

| Climatic conditions | Vegetation Index | Lightning | Number of past fires |
|---|---|---|---|
| **Climatic conditions year before** | **Snow cover** | Roadmap | Number of past fires in neighbors |
| Climatic conditions evolution | **Vegetation type** | **Closest roadmap** | Month since last fire |
| Climatic conditions of neighbors | **Elevation** | Population | Month since last fire in neighbors |
| Climate indices | | **Closest population** | |



Feature map

## Best Severity Model

To obtain a thorough evaluation of wildfire risk, it is necessary to estimate the size of the wildfire event. The severity model builds on the occurrence model to predict size of wildfires. We broke down the size of wildfires into two: Small and Large.

**Performance:** This model has strong performance, it is able to catch more than 50% of the potentially costly wildfires with a relatively high precision.