MaxEnt in a nutshell:

Distributions are smoothed out histograms whose total area is equal to 1 used in statistics. The goal of a MaxEnt model is to compare the statistical distribution of environmental conditions and asset characteristics associated with ignition locations (i.e. the distributions of tree heights, conductor sizes, rainfall averages, fuel dryness, etc.) to the distribution associated with all locations along the distribution grid, also called the background. To the extent those distributions have statistically significant differences, the model can use values associated with any given location to assign a probability that that location would be a place where ignitions will occur in the future. For example if the fraction of vegetation caused ignitions near trees taller than 15 meters is 10x larger than the fraction of randomly selected distribution grid locations near trees that tall, then, all else being equal, locations with trees taller than 15 meters will be predicted to be roughly 10x more likely to experience ignitions.

The key step to differentiating "ignition occurrence conditions" from "background conditions" is to calculate the ratio of the ignition location distribution and the background distribution. This ratio is also known as the relative occurrence rate or the MaxEnt "raw output". To do this, the distribution of ignition occurrence conditions must be estimated based on the limited number of values at the known ignition locations. Those values will be consistent with many different potential distributions so the question is which one to use.

"Maximum entropy" in this context refers to the goal of maximizing the relative information entropy of the estimated ignition occurrence distribution compared to the background distribution, while still properly characterizing the rate of occurrence of values observed at ignition locations. The higher the relative entropy, the more similar the two distributions will be.

So in layman's terms, the MaxEnt model estimates the distribution of environmental conditions associated with ignitions in a manner that requires it to be as similar to the conditions found elsewhere on the grid as possible while still accurately characterizing the rate of occurrence of values observed at ignition locations. The similarity is quantified through a value know as relative information entropy, which gives this method its name.

Simpler:

We are interested in which environmental conditions and asset attributes (collectively called the model covariates) are more common among ignition locations than they are among all distribution grid locations. For example, tall trees are more common among vegetation caused ignition locations than they are among typical Dx grid locations. Metrics of dryness, HFTD tier assignments, conductor materials and size, and others, can all be checked for such patterns. The ratio of covariate value prevalence at ignition locations to their prevalence across all grid locations is called the relative occurrence rate.

MaxEnt provides a way of estimating the relative occurrence rate given a fairly modest number of ignition locations. The way it does this is to fit a statistical distribution of covariate values for ignition locations that is consistent with the values at known ignition locations, but otherwise as similar as possible to the distribution of values found everywhere else along the Dx grid. The similarity criteria is enforced using a metric called the *relative information entropy* between the ignition locations and the

Dx grid locations, where the larger that metric is, the more similar the two distributions are. For this reason, the overall approach is referred to as a maximum entropy or MaxEnt estimation of the relative occurrence rate. When multiplied by the fraction of all grid locations that experience ignitions annually, the relative occurrence rate is normalized into an estimate of the annual probability an ignition will occur for all values of the covariates. This can be used to look up (aka predict) annual ignition probabilities based on the covariate values found at each Dx grid location.

From our MaxEnt methods document section 3, Introduction

To answer the question of *where* ignition events are likely to occur, we have estimated fire season ignition probabilities using maximum entropy models (MaxEnt) pioneered in the modeling of ecological ranges of species. These models are trained on ignition (or outage) locations and gridded spatial (raster) environmental and asset attribute data. The data can draw from a specific time period, but the model itself is dedicated to spatial, not temporal, patterns. The Maxent model provides relative scores or, if properly calibrated, probabilities for fire-season ignitions per "pixel" of input data.

(See section 4 of that document for a more in-depth treatment of the mechanics and most of the important math….)

From our Phase 1, Milestone 1 analysis documentation (there is a pdf in ESFT):

Intuitive/physical basis for model fit: difference in distributions of environmental conditions for the whole Dx grid, vs ignition locations:

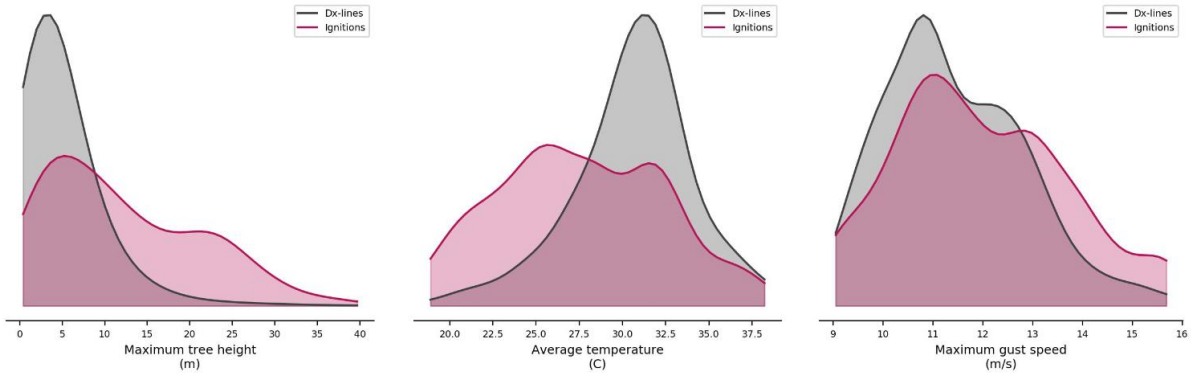

Figure 2. Normalized density distribution plots comparing the environmental conditions between all distribution lines (grey) and all ignition locations (red) for maximum tree height, average temperature, and maximum gust speed (from left to right). Ignition locations were more likely to occur in areas with taller trees nearby and in areas with lower average temperatures compared to the full population of distribution lines. Maximum gust speeds at ignition locations were similar to gust speeds at all distribution lines, with some sites experiencing disproportionately high maximum gust speeds.

The rest of this document provides a clear, but detailed explanation of the mechanics and application of MaxEnt to wildfire.

Presentation with slug example

See slides 16-31 of "Lunch and learn presentation.pptx" (just uploaded to ESFT) for a worked example and a visual explanation of how MaxEnt works and is typically used. (good figures)

Model choices

Our Phase 1, Milestone 3 documentation (DxRisk Phase 1 Milestone 3_ Overview and Model Specifications.pdf in ESFT) describes work on several model specifications. Those were models designed to answer:

- **Where**: high spatial resolution estimates of ignition probabilities over one or more years – models with spatial priority. This is the role that we cast MaxEnt for.
- **When**: determining what time varying conditions tend to lead to failures, limited by training data to fairly coarse timesteps and spatial resolution. This is the role that we cast an Arrival Process model for – specifically Poisson / Negative Binomial GLMs estimating ignition count probabilities under different circumstances.
- **What type**: determining the statistical relationships between events of different types. For example, what are the odds that an outage is associated with an ignition, conditional on environment, location, weather, outage characteristics. This is the role that we cast a regularized logistic regression classification model for.

Here is a matrix of model options with pros and cons for the Where problem:

	Spatial models			Asset models	
		SVM, kernel machines	Logistic regression, random forest, etc.	Logistic regression, random forest, etc.	Arrival process or survival models
Model requirements	**MaxEnt**	**Pixel**	**Pixel classification**	**Asset classification**	**Asset-based**
handles sparse data	x (regularized fit robust to over-fitting)	x (empirical question)	x (limits number of parameters; risk of over-fitting)	x (limits number of parameters; risk of over-fitting)	Not well
handles zero inflation	x (filter to grid/HFTD pixels)	x (filter to grid/HFTD pixels)	x (filter to grid/HFTD pixels)	? (trickier than pixels)	? (trickier than pixels)
handles class imbalance	x (ratio of distributions is presence-only model)	x (bootstrapping or class-specific weights - tougher than presence only)	x (bootstrapping or class-specific weights - tougher than presence only)	? (imbalance is worse for assets)	? (imbalance is worse for assets)
models assets or fine locations	x	x	x	x	NO
models time step varying conditions	NO	NO	NO	NO	x
good with environmental causes	x	x	x	x (envornment near asset can be regressors)	x
good with asset attribute causes	OK	OK	OK	x	x
robust to uncertainty in locations	x (prevailing conditions define distributions)	x	x (prevailing conditions dominate model fits)	NO	NO
robust to uncertainty in equipment involved	x	x	x	NO	NO
provides probability output	x (through tau adjustment)	NO	x	x	x (probability of failure by count)

Note that:

- data issues (uncertainties with locations and equipment involved) strongly disadvantage asset-based modeling approaches.
- The weakness of SVM and other kernel machines with providing probabilities render them not fit for purpose, given our probabilistic risk definition.
- The remaining spatial approaches are relatively weak on time-varying conditions and both operate at the pixel-level. MaxEnt's presence-only model form more directly side-steps class imbalance while remaining parsimonious.
- Pixel-based regression-family of models are expected to perform similarly to MaxEnt, given sufficient time to develop the bootstrapping machinery necessary to address class imbalance and regularization machinery necessary to counteract over-fitting.
- The 2021 model was based on strongly environmentally interacting ignitions (veg and conductor). Equipment-caused failure modes being added this year are expected to be more sensitive to asset

attributes and improved asset data and will benefit from improved outage-asset linkages and asset attribute data sources to be modeled using asset models.

- Data sets that track changes in asset attributes over time may be available in the future but will not be available for the next round of modeling.