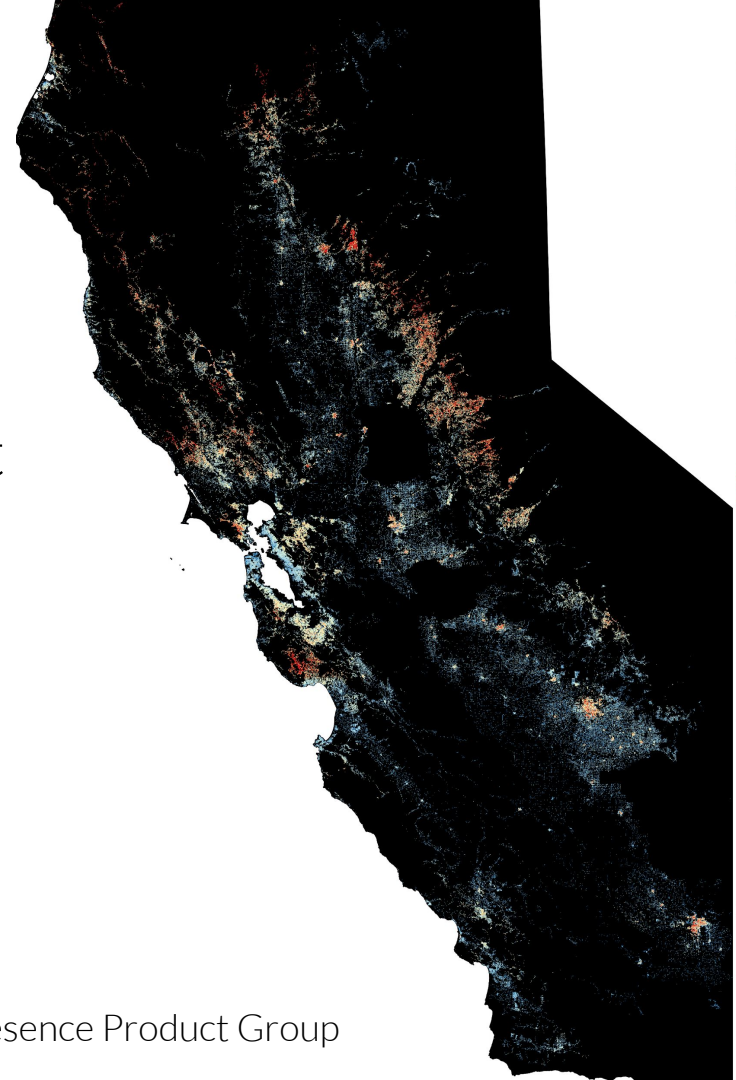
DxRisk Phase 1 results:
Predicting outages - where, when and what



May 15, 2020
For PG&E internal use

Convergence Data Analytics, Salo Sciences, Presence Product Group

Talk outline

Project background

Modeling objectives

Data sets and processing

Current capabilities and results

Phase 2 and beyond

Discussion and Q&A

Project background

Context – Why are we undertaking this initiative?

In support of the new EO Risk Paradigm, PG&E is developing a Distribution (Dx) Asset Risk Model (the Model), tuned for Wildfire Risk, which will:

- Provide situational awareness of the current wildfire risk on the Dx system
- Enable risk-informed decision making in the budget planning process
- Allow PG&E to report risk reduction to regulatory entities

Note: This project will be an input into and is proceeding in coordination with ongoing PRA risk modeling and will be validated by and is expected to be an input into EORM's process.

Phase 1 modeling complete

Milestone 1: proof of concept “backstop” spatial model (MaxEnt) built with public data

Milestone 2: PG&E data sets validated, cleaned and integrated into project “data pipeline”

Milestone 3: Modeling “toolkit” used to develop Where, When, and What Type models and all delivered as working code

Modeling objectives

Why make predictions?

Planning / budgeting / prioritization of mitigation

Risk estimation and management

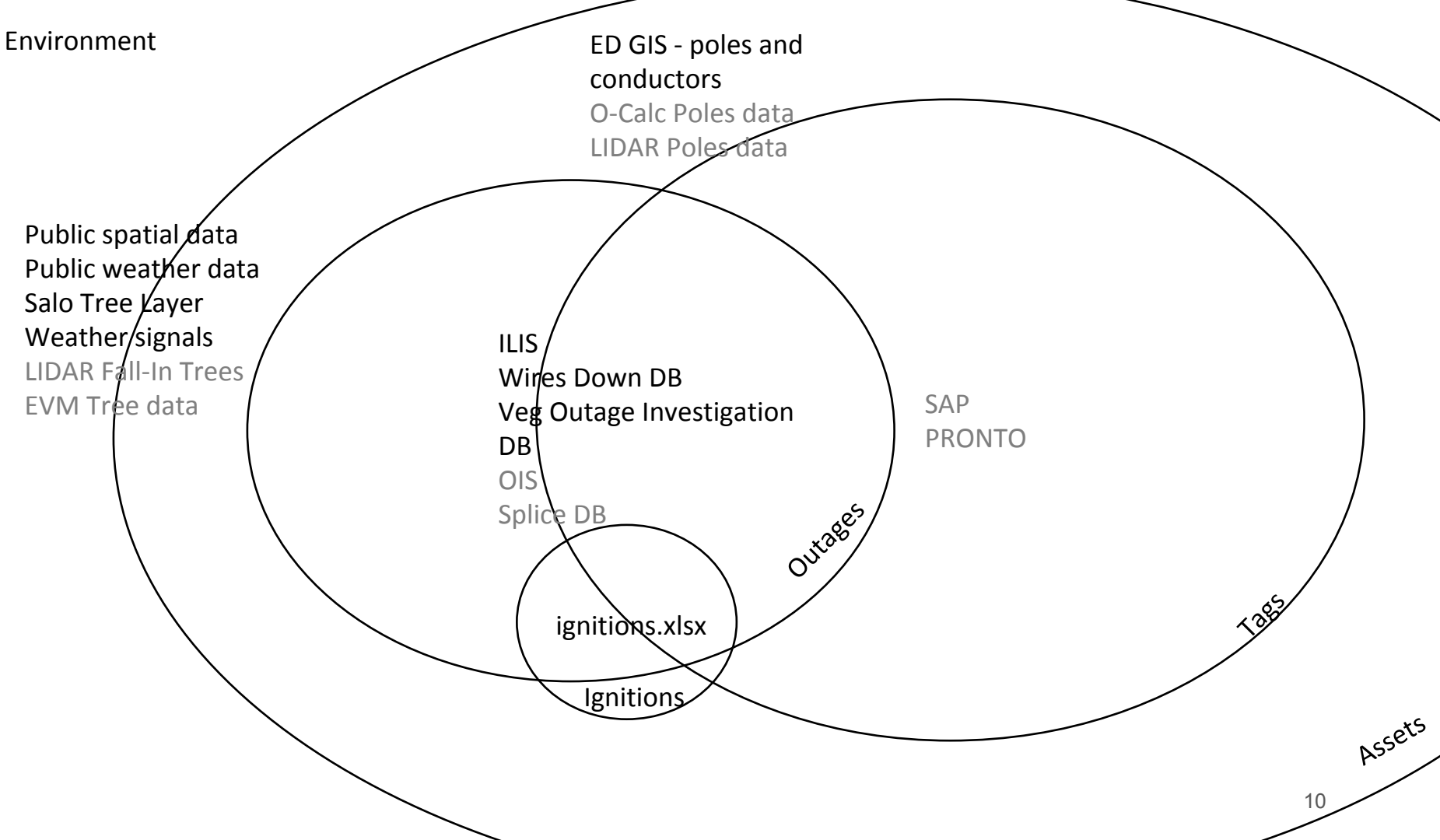
Operations / PSPS

Learning and discovery

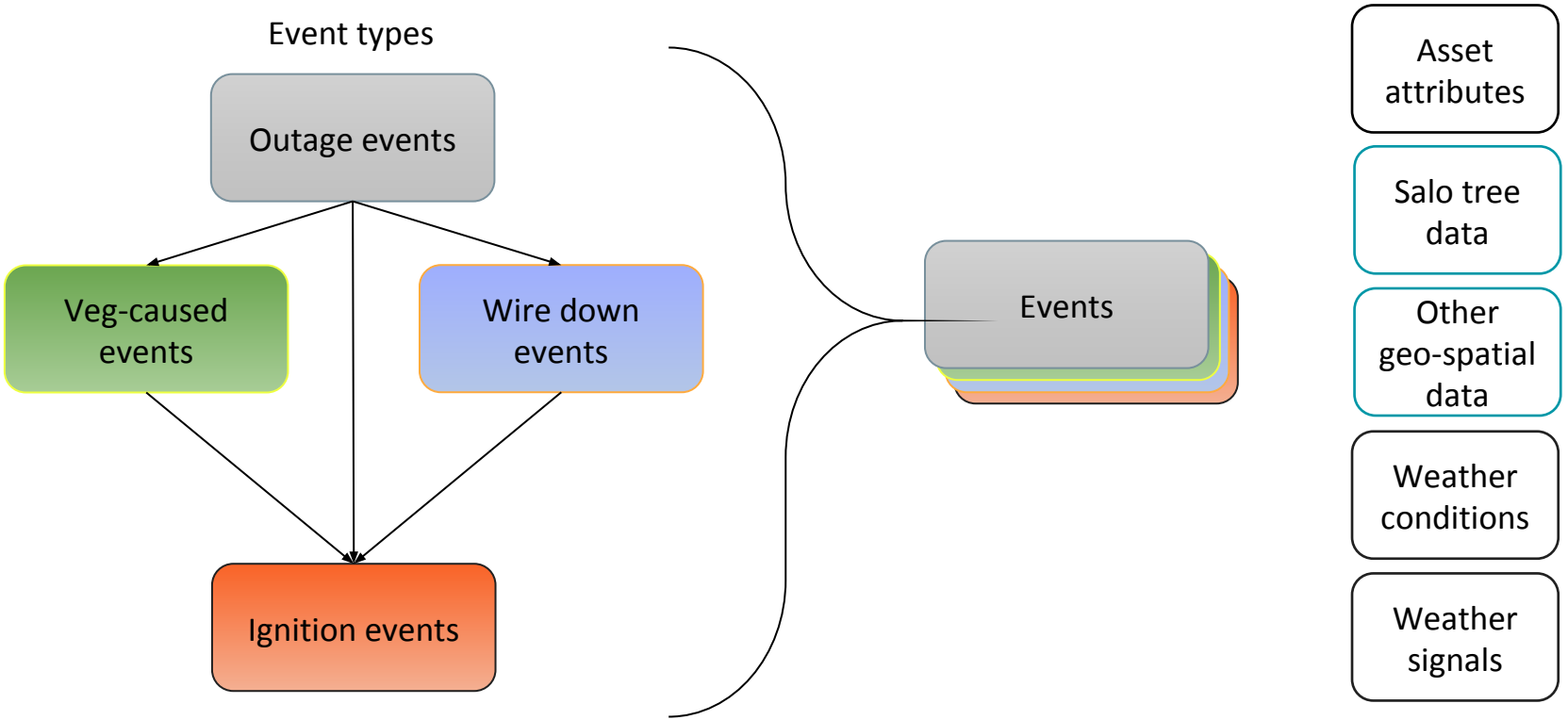
Categories of questions - each implies a different modeling strategy

- Where
  - What assets are at elevated risk of failure or ignition?
  - What locations experience similar conditions to locations where a certain type of asset has tended to fail in the past?
- When
  - Under what changing conditions is there an elevated risk of failures?
  - What is the expected count of outages over a month / year of operations?
- What type
  - Given an outage, what are the odds that it is associated with wires down or ignitions?
  - What factors affect the odds of ignitions and how are those odds altered through preventative work?

Data sets and processing

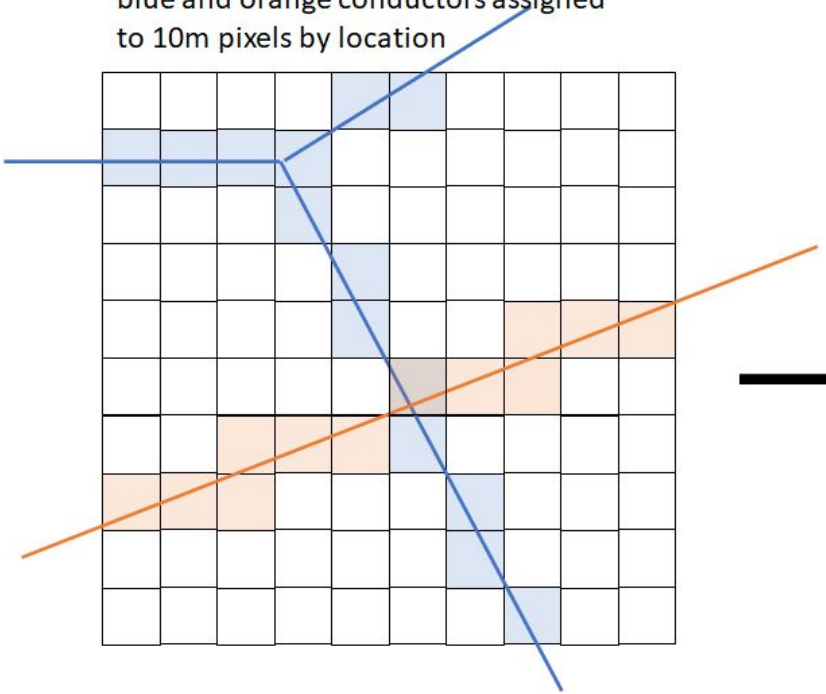

Environment
ED GIS - poles and conductors
O-Calc Poles data
LIDAR Poles data
Public spatial data
Public weather data
Salo Tree Layer
Weather signals
LIDAR Fall-In Trees
EVM Tree data
ILIS
Wires Down DB
Veg Outage Investigation DB
OIS
Splice DB
SAP
PRONTO
Outages
ignitions.xlsx
Ignitions
Tags
Assets

Project data

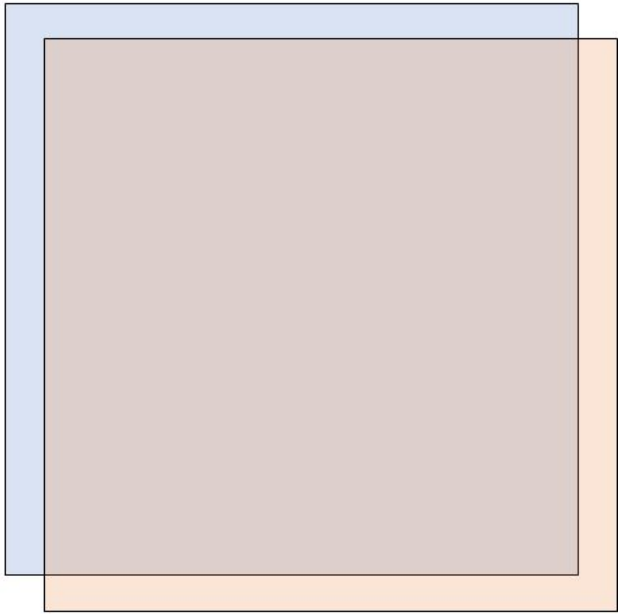


Feeder “pixels” and “roll-ups”

Conductor asset characteristics for blue and orange conductors assigned to 10m pixels by location

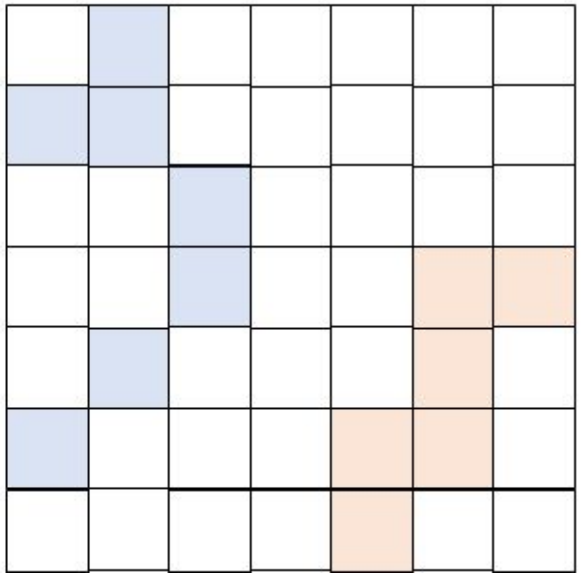


Two 100m pixels (at the same location) summarize aggregate data from underlying pixels: based on 14 for the blue line and 12 for the orange line, preserving conductor total length, number of phases, etc.



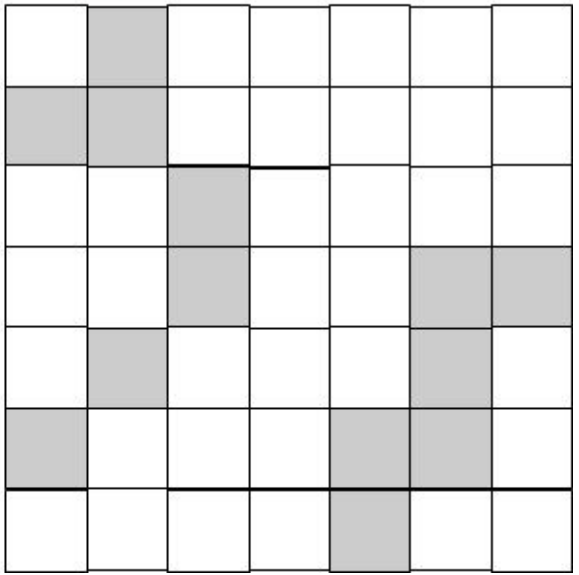
Pixel data augmentation

Geo-spatial data from the same locations as 100m pixels added to data fields tracked per-pixel.



+

Elevation, population density, eco-region, etc. at 100m scale.



Current capabilities

What makes this a hard problem?

- Sparse data
  - Risk of overfitting or mis-interpreting “fit” that “accurately” predicts no events
- Zero inflation
  - In some cases, events like ignitions or vegetation caused outages are impossible (wet conditions; no trees)
  - Those locations/times incapable of experiencing events dilute the pool of outcomes actually determined by asset health, environmental conditions, etc.
- Probabilities, counts, and their uncertainties
  - What we are looking for are black swan / long tail events
  - Inherently large uncertainties
  - “Standard” assumptions about variable distributions (often made for mathematical convenience) can lead to under-estimates of event counts
- Inferring causality
  - To assign expectation values to different risk mitigation scenarios requires models that have learned coefficients related to the grid attributes altered under the scenarios.
  - Models optimized around prediction alone may not be capable of modeling scenarios

Temporal trees:
e.g. using weather signals

Logistic classifiers:
e.g. zero inflation or pre-classification

fine time (weekly/daily)

coarse spatial (feeder or division)

fine spatial (protection zone; asset)

coarse time (annual/total)

Composite model:
Assemblage of model components

Event classification

$$\log\left(\frac{P(C|A)}{1-P(C|A)}\right) = \beta_0 + \sum_{i=1}^{N} \beta_i x_i$$

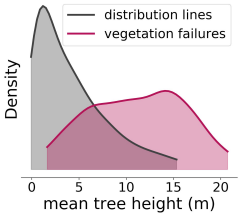
Synthetic controls?:
e.g. event day matching with non-events and non-event day at location of events

Arrival process
(Poisson/Neg. Binomial)

$$\Pr(X = k) = \binom{k+r-1}{k} p^r (1-p)^k$$

Spatial filtering:
e.g. filtering on locations with fall-in trees or logistic classification

Maxent

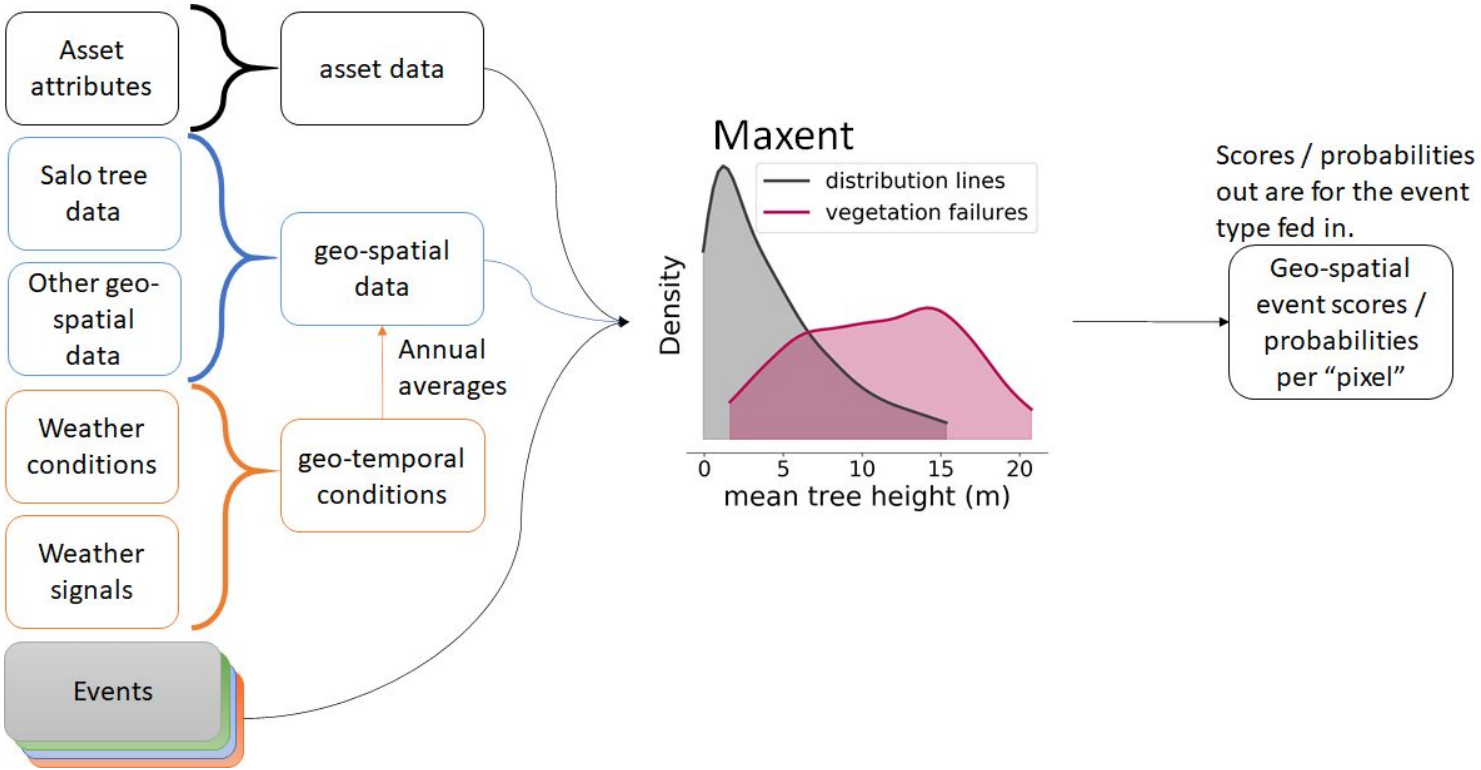


Standard regression tools:
Feature engineering and selection

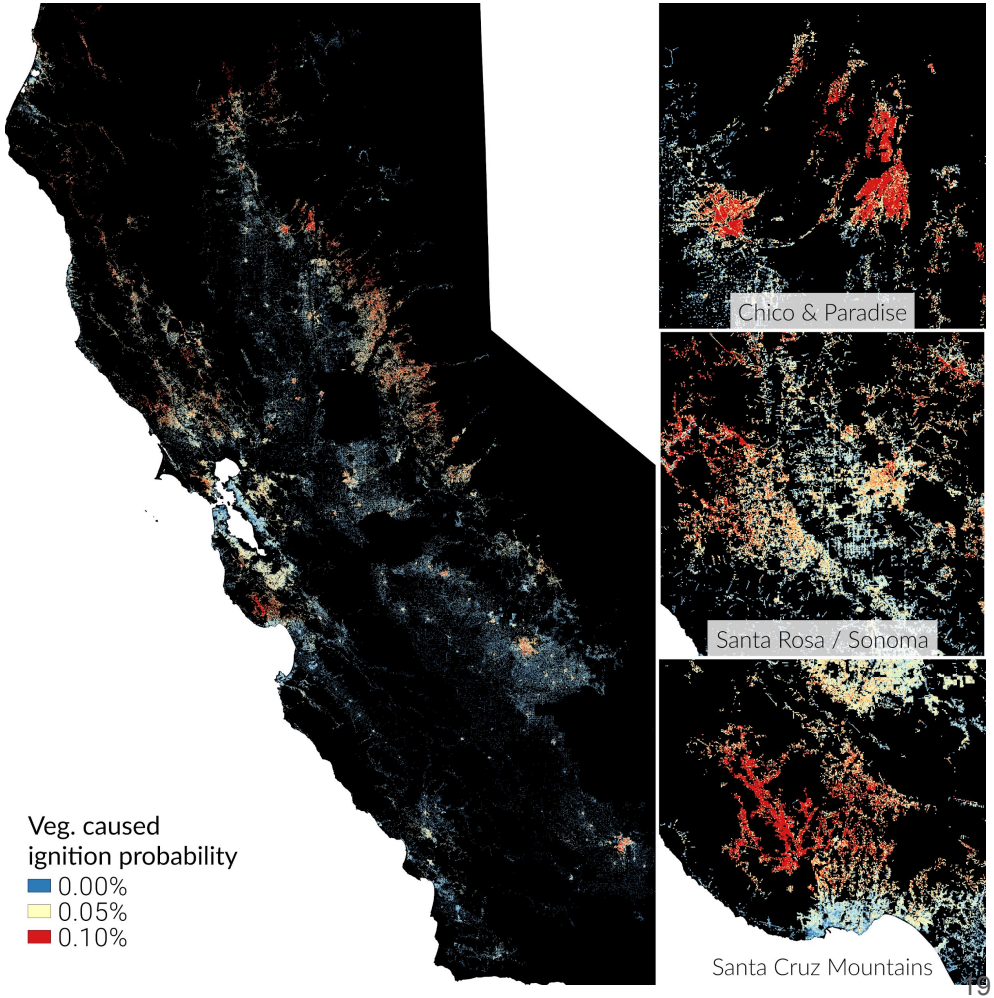
Asset grouping classifiers:
e.g. by type or pre-classification

Where?

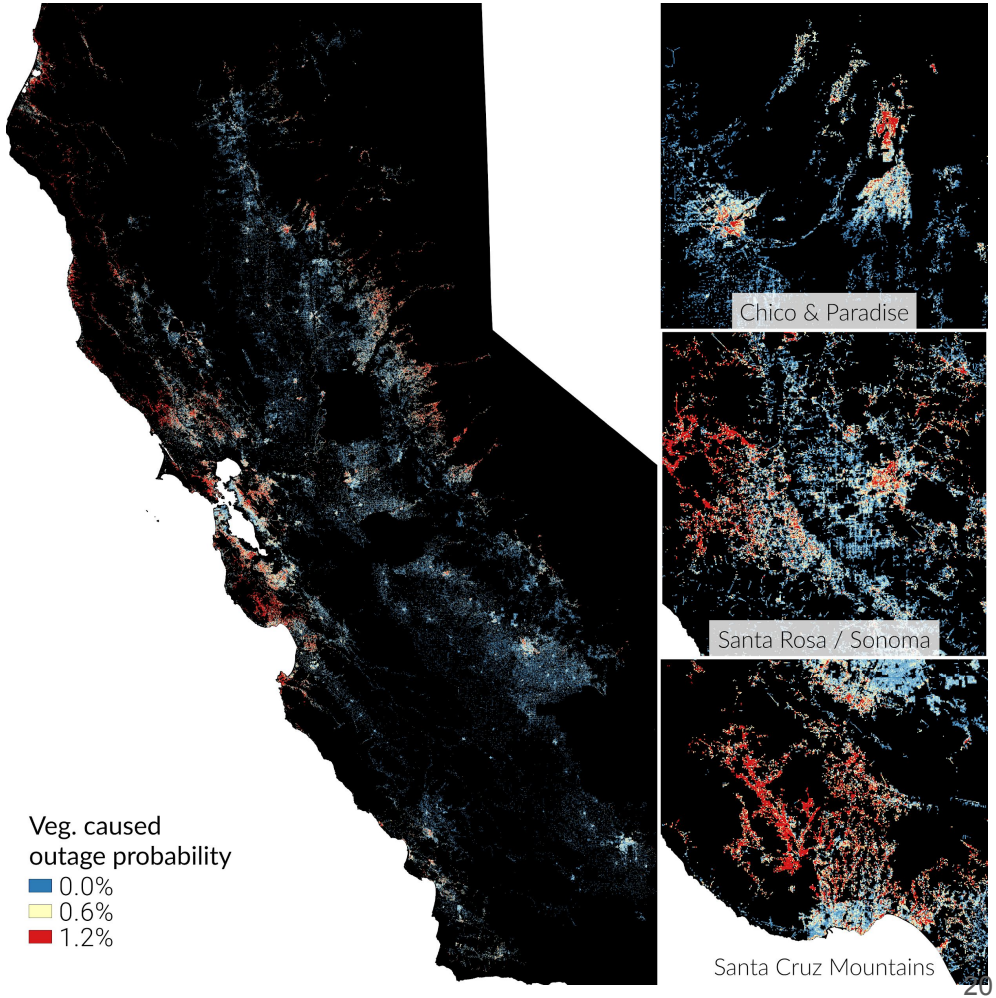
Where: via Maximum Entropy (MaxEnt) modeling



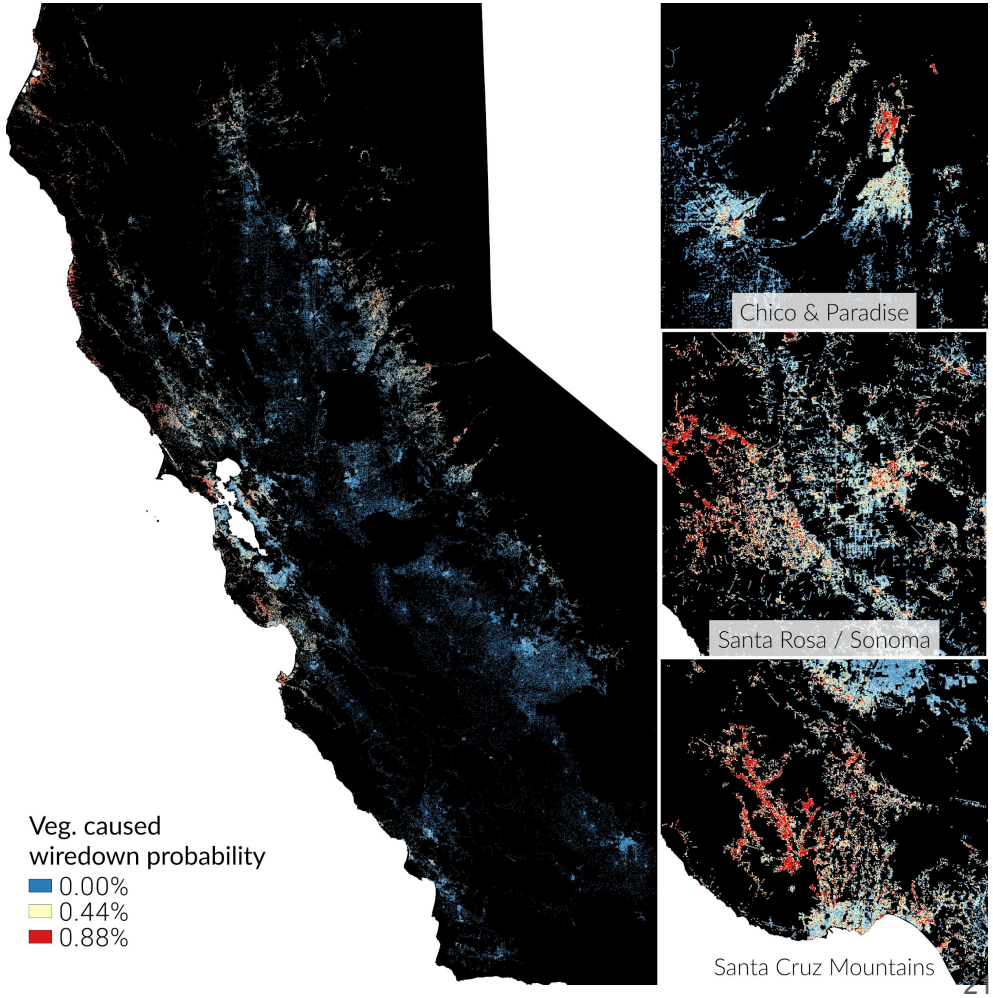
Vegetation caused ignitions



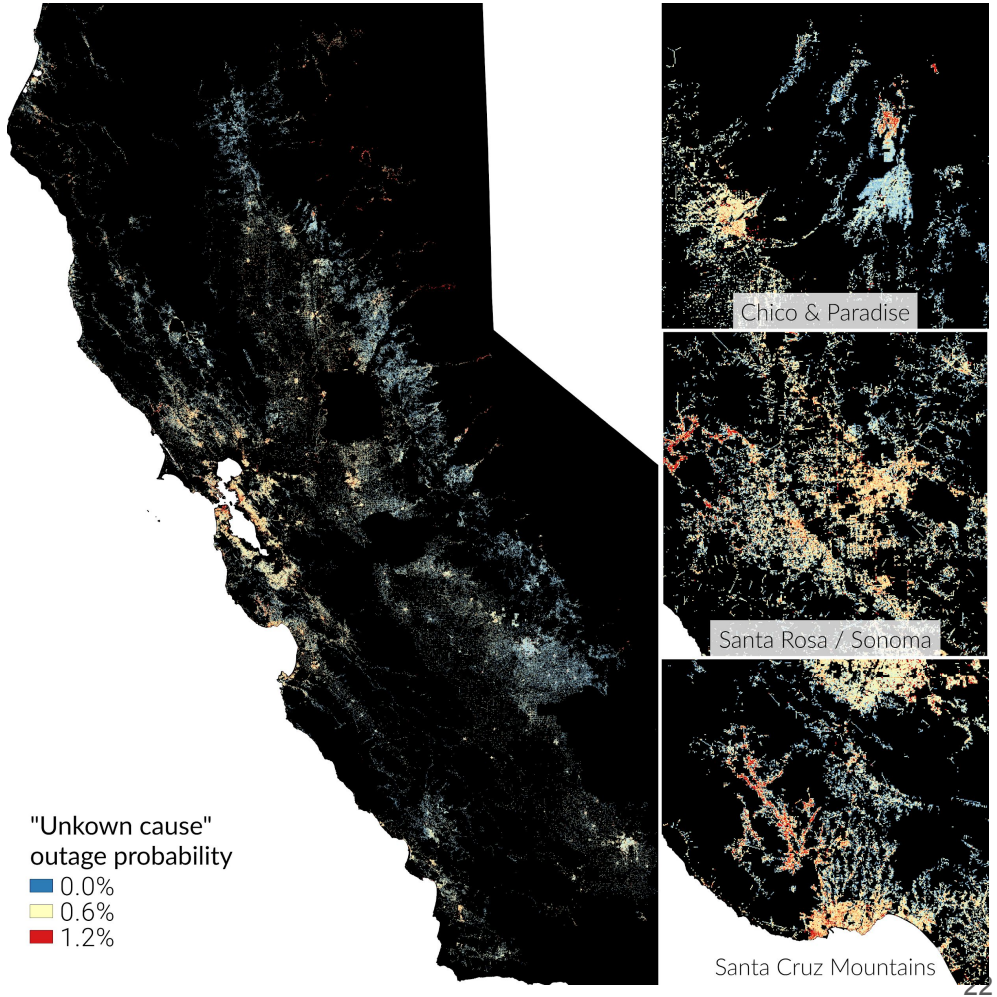
Vegetation caused outages



Vegetation caused wires down



Unknown cause outages



Deliverables and status

Based on this modeling approach, we've provided Vegetation Management

1. A geo-tif raster file of vegetation-caused ignition probabilities
2. A csv "roll-up" of expected count of annual ignitions by feeder

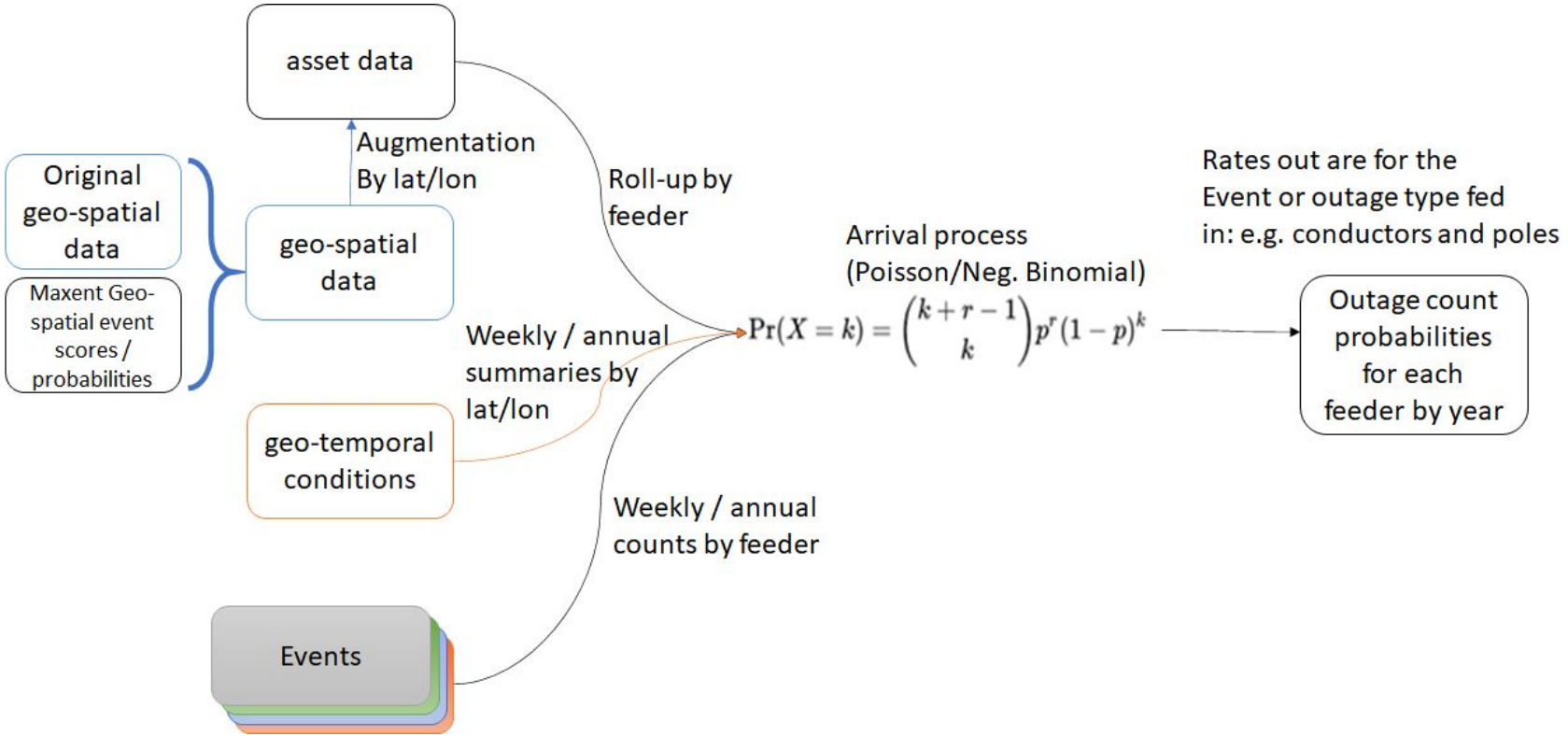
An earlier version of feeder roll-ups was also provided as input to the circuit prioritization effort.

At the request of VM, we are working to incorporate recently provided data on tree species into future modeling runs.

When?

When: Arrival process modeling:

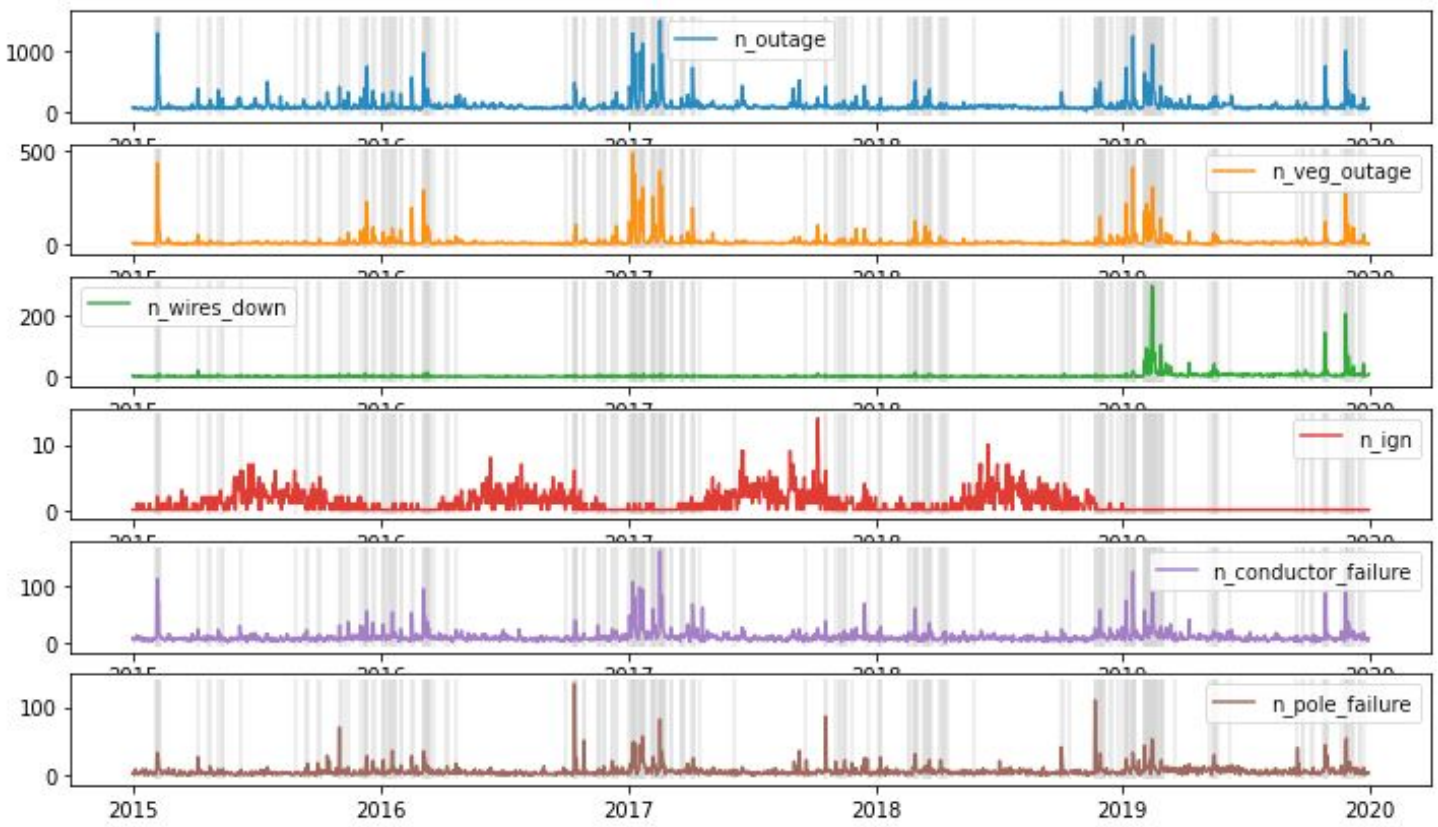
Probability of observing N events over a given time interval/conditions



Feeder outage rates model: inputs

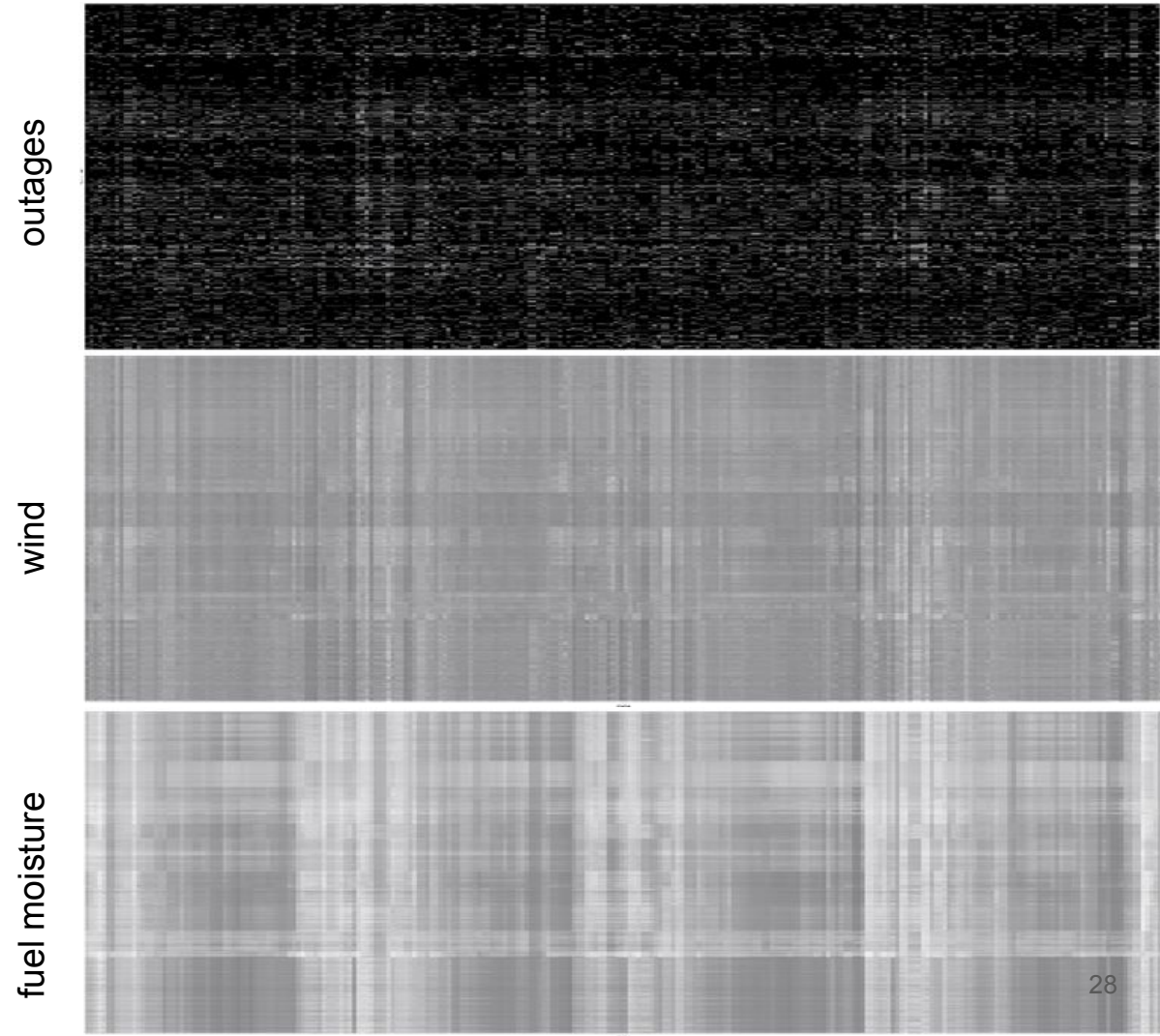
Description	Value
Outage data source	ILIS outages
Feeder data source	ED-GIS primary OH conductors
Weather signals source	Meteorology team data set
Count of GIS conductors entries	1.4M
Length of conductors studied	131,050 km
Timespan	2007-2019
Total outages	504,252
Unique feeder count	3,103

System wide weekly event counts: 2015-2020

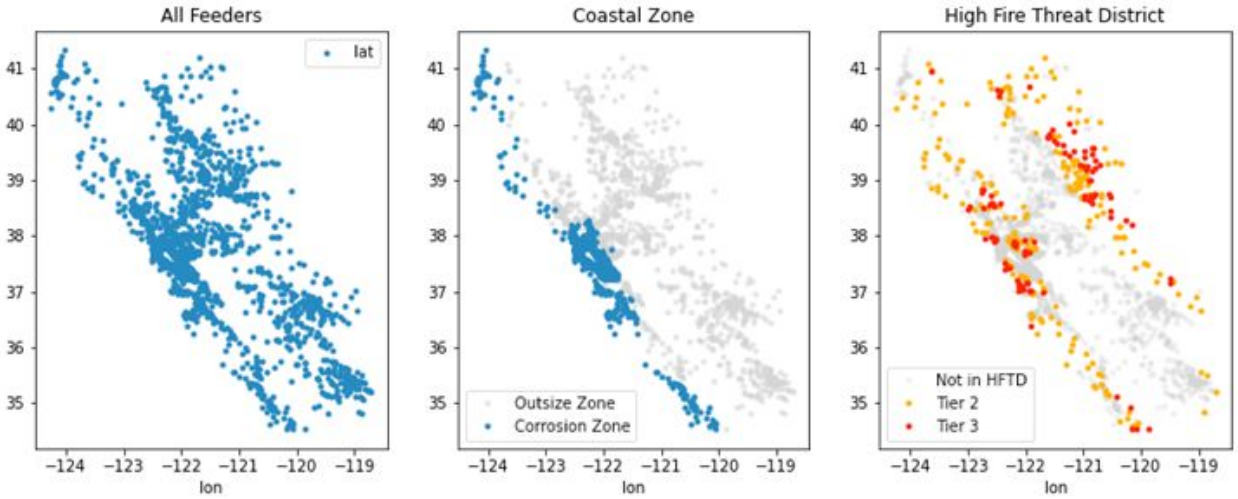


Visualizing outages and covariates

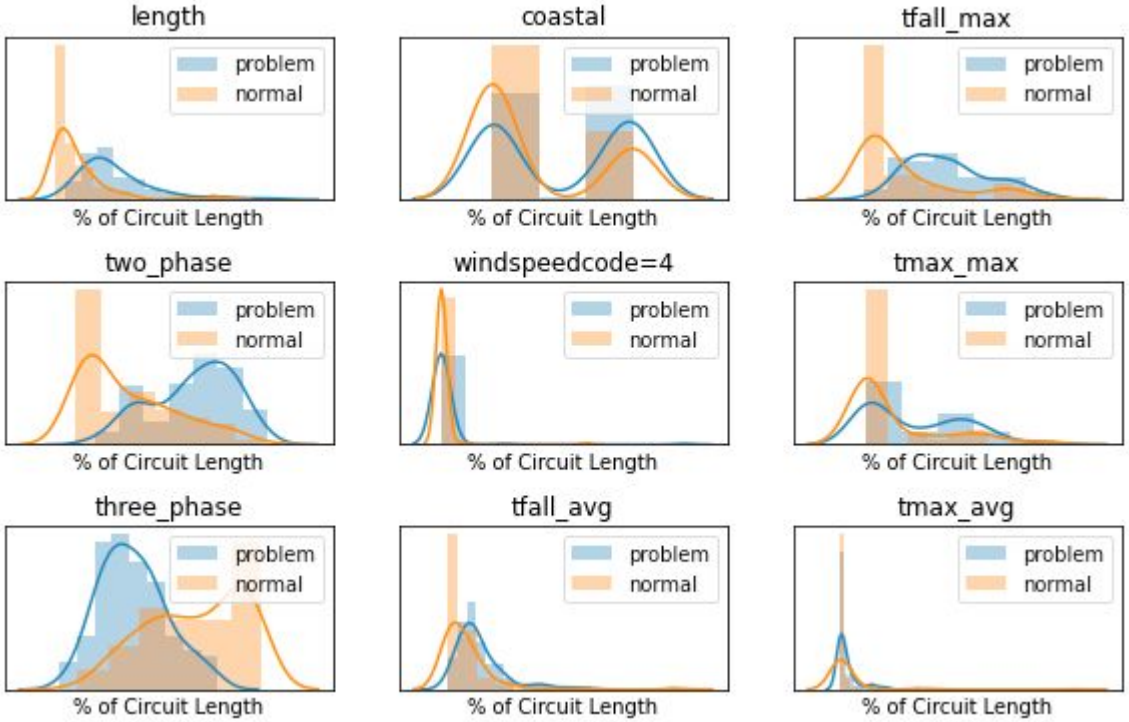
1 row per feeder (3300 in total)
columns are weeks (2016-2019)



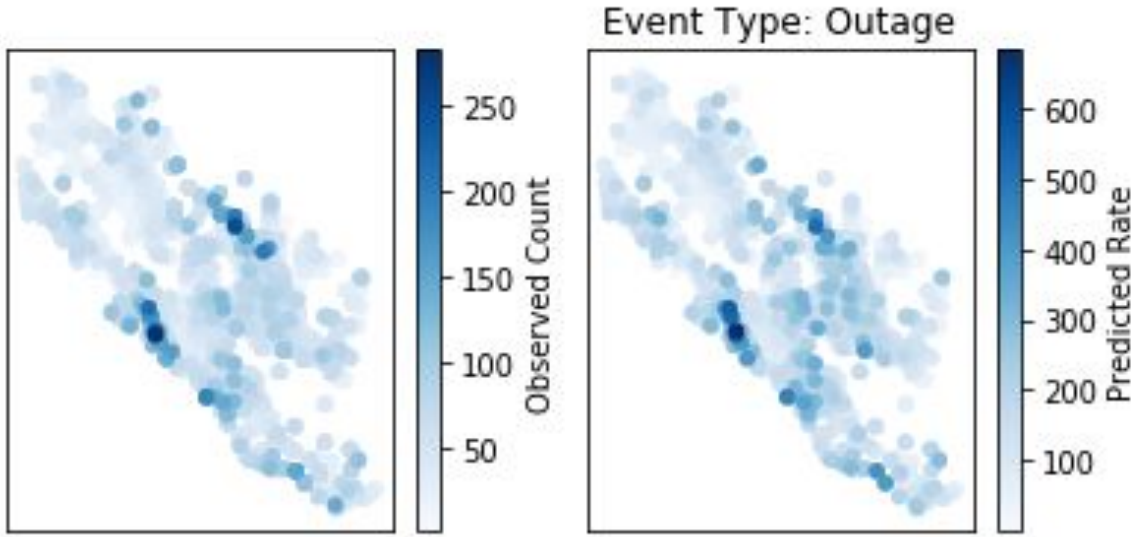
Feeder attributes



Feeder attributes of “normal” vs. top 100 outages



Observed vs. predicted annual feeder outage count



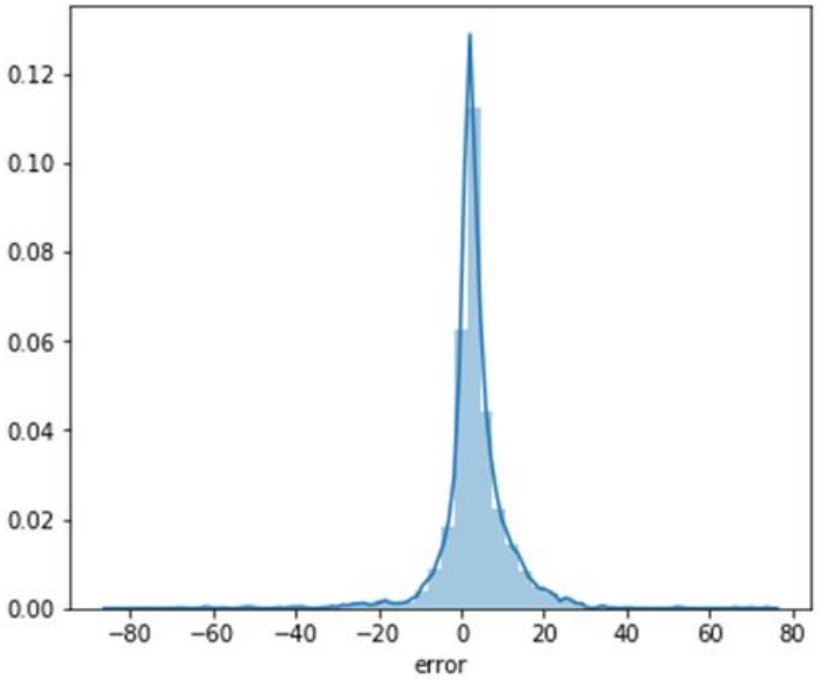
Annual feeder count errors

Error in outages per-feeder-year:

The mean is 5.7

The median is 3.4

85th percentile is 10



Deliverables and status

Flexible framing

- Model can run with feeder-specific counts, modified by weather signals
- Model can also run on feeder attributes, modified by weather signals
- Inputs can be outages or outage subsets by cause or equipment type

Currently working on

- Better normalizing long tailed covariates like length and fall-in tree count
- Feature engineering and regularization to take better advantage of asset attributes
- Optimized runs for each equipment type
- Characterizing the degradation of performance at finer timescales

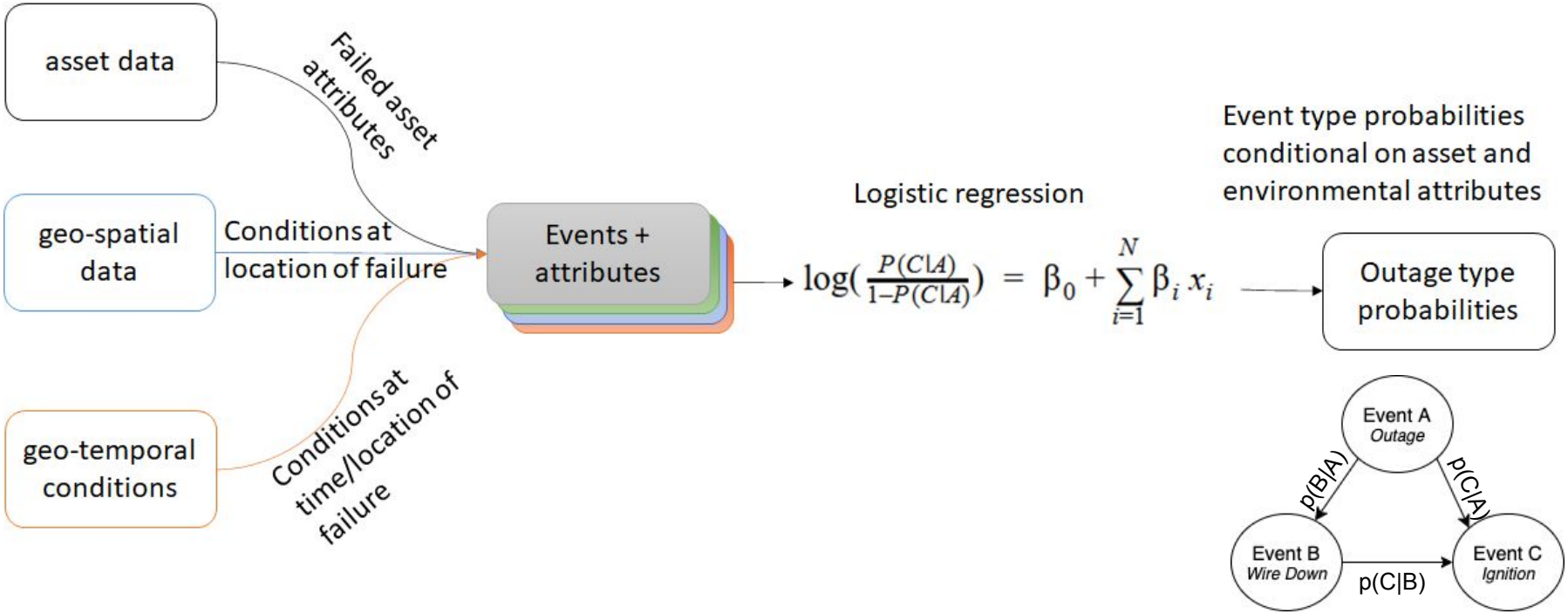
What type?

Unconditional wire down and ignition probabilities

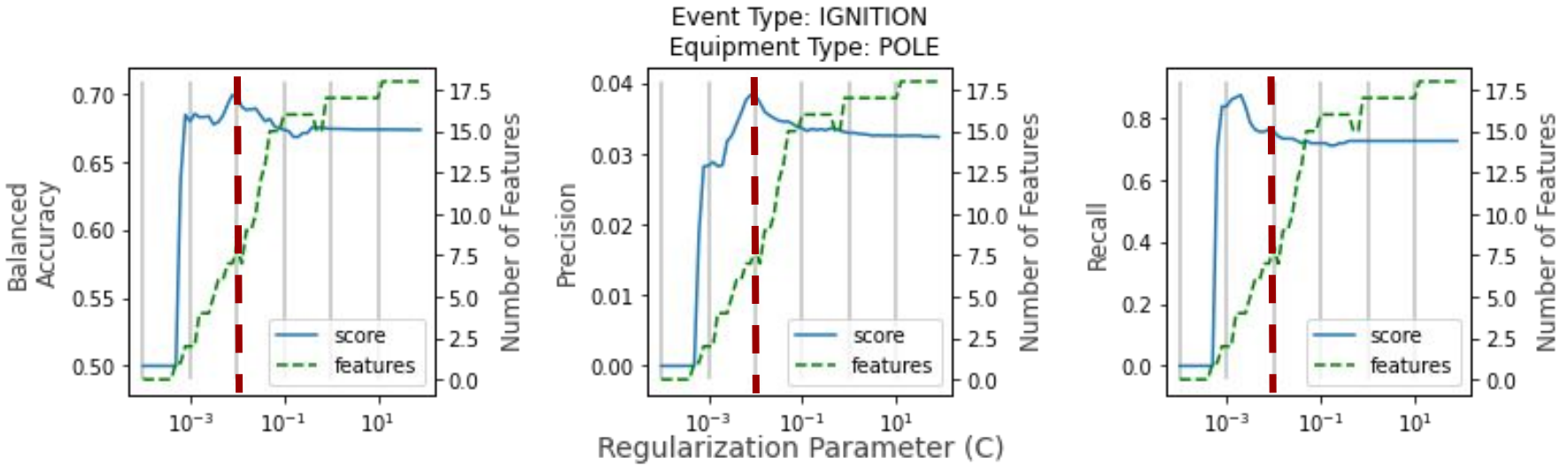
E.x.: Vegetation caused outages 2015-2019

	No Ignition			Ignition		
	Count	% of total	% of row	Count	% of total	%of row
Wire Down	21,557	10.3%	96.7%	734	0.35%	3.29%
No Wire Down	186,429	89.0%	99.6%	694	0.33%	0.37%

What type: Event classification

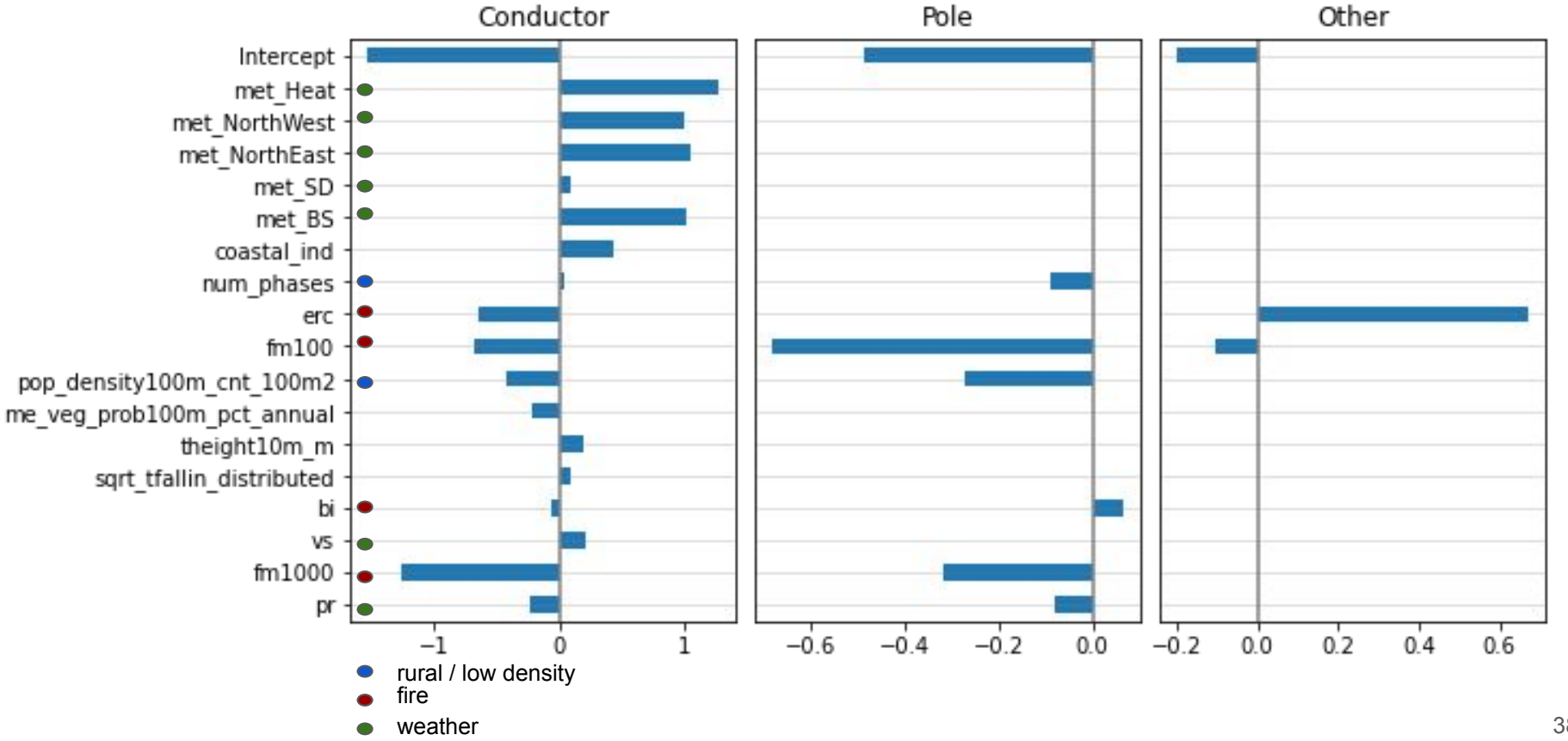


Regularization: fewer parameters; better fit



Fewer parameters improve precision until this point, so we select this value of C as for our model

Best fit models for Conductors, Poles, and Other equipment types



Deliverables and status

Most recently developed model

- Separate results by equipment type or event cause
- Expected applications in scenario-based risk scoring

Currently working on

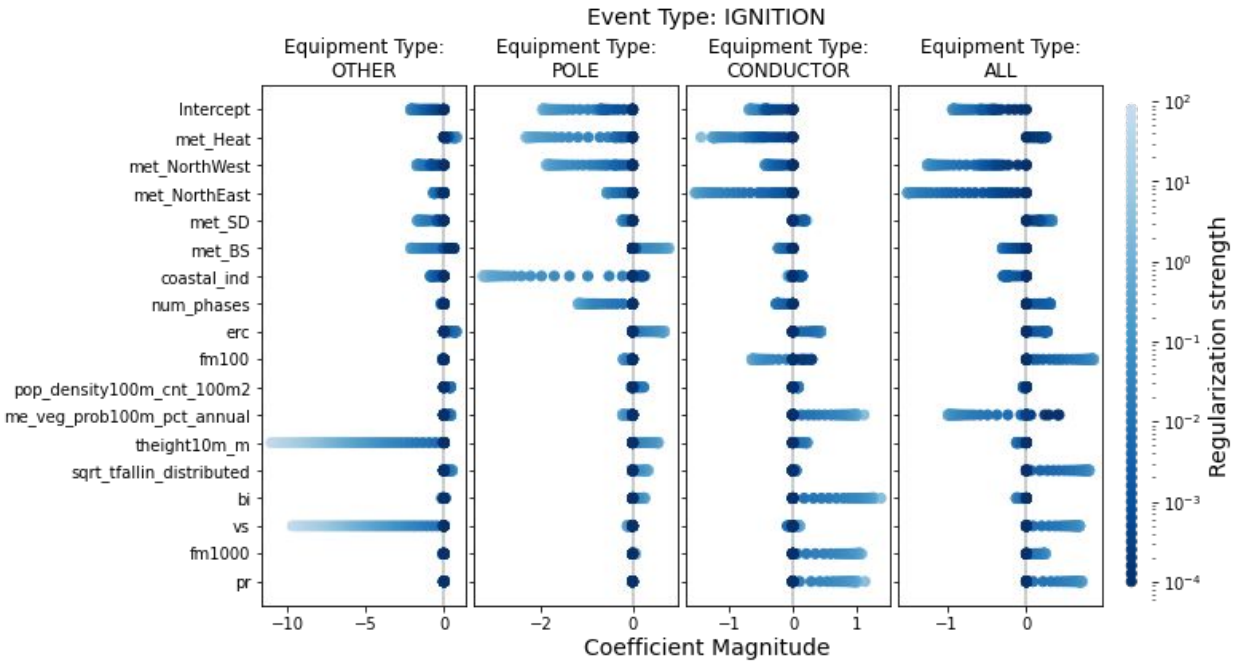
- Incorporating more/new asset attribute data (i.e. on poles and conductors) into event metadata
- Looking at Bayesian approach to estimating similar conditional probabilities

Phase 2 and beyond

Phase 2 May through July

- Modeling tools and data pipeline deployed as packaged codebase into PG&E compute environment
- PG&E data science team knows the tools and is successfully deploying improvements and new models
- Joint CDA/Salo/Presence/PG&E modeling work continues, with new data sets, and new and/or improved models
- Model results delivering value in support of PG&E planning, risk mitigation, and regulatory submissions

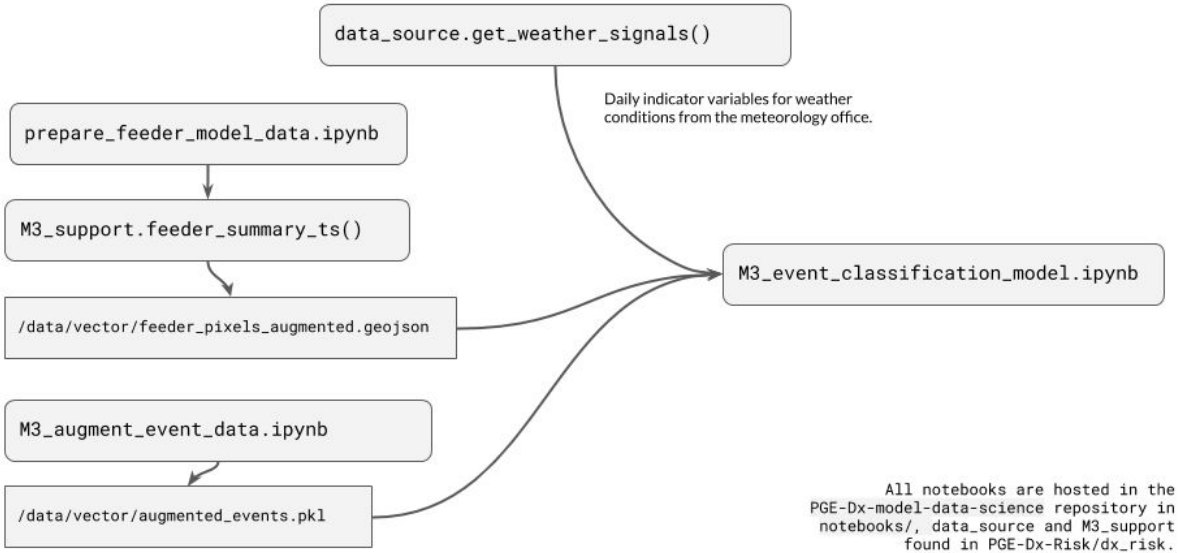
Q&A

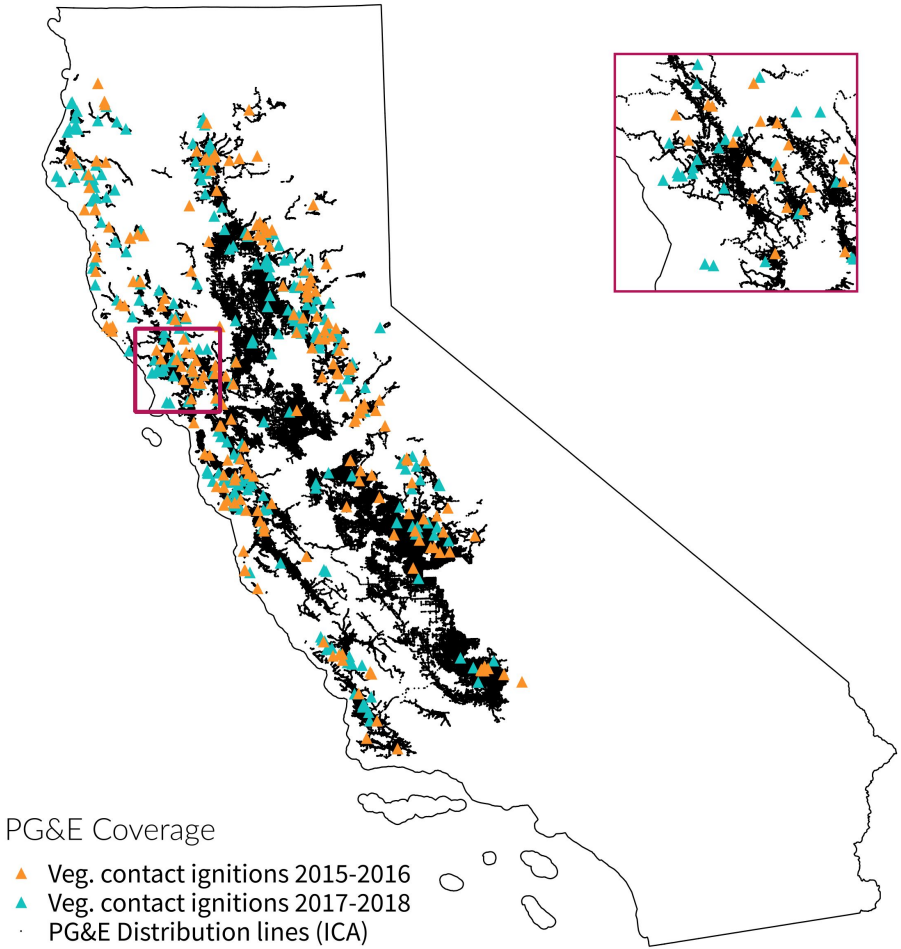

Event Type: IGNITION
Equipment Type: OTHER
Equipment Type: POLE
Equipment Type: CONDUCTOR
Equipment Type: ALL
Intercept
met_Heat
met_NorthWest
met_NorthEast
met_SD
met_BS
coastal_ind
num_phases
erc
fm100
pop_density100m_cnt_100m2
me_veg_prob100m_pct_annual
theight10m_m
sqrt_tfallin_distributed
bi
vs
fm1000
pr
−10
−5
0
−2
0
−1
0
1
−1
0
Coefficient Magnitude
Regularization strength
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}

Feeder attributes

Attribute	Definition
feeder_id	Also called circuit id - 9 "digit" numerical identifier with leading 0 padding.
feeder_name	Substation name and circuit number
district	Grid district
division	Grid division
region	Grid region
lat	feeder centroid latitude
lon	feeder centroid longitude
coastal_ind	1 if the feeder is flagged as coastal; 0 otherwise
num_phases	average number of phases of the feeder
length	total length of the feeder's conductors
n	total number of 10m pixels the feeder spans
hftd100m_zone	maximum HFTD the feeder passes through
tfallin10m_cnt	count of 10m pixels the feeder passes through that contain estimated fall-in trees
theight10m_m	90th percentile height of trees along the feeder's path
elevation100m_m	average elevation of the feeder
me_veg_prob_annual	Maxent estimate of expected annual number of vegetation caused outages on the feeder
pop_density100m_sum	total population within 100m of the feeder's lines

Event classification model data flow





Ignition locations

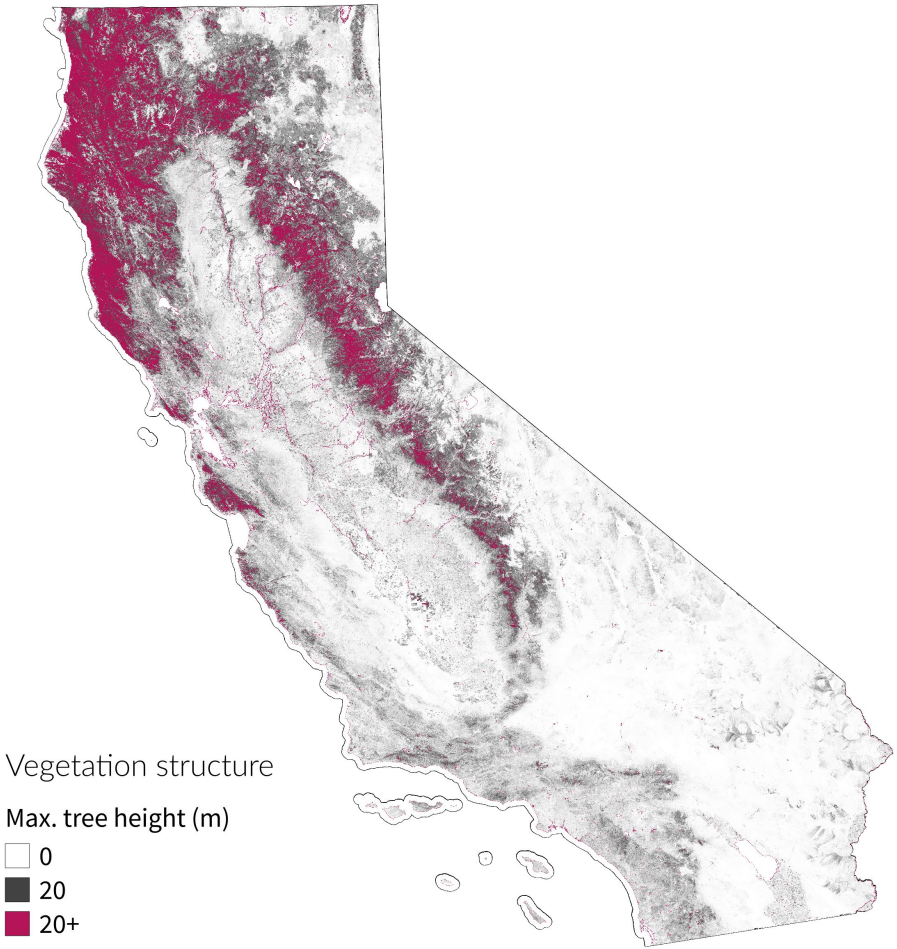
- 2015-2016 ignitions
- 210 points

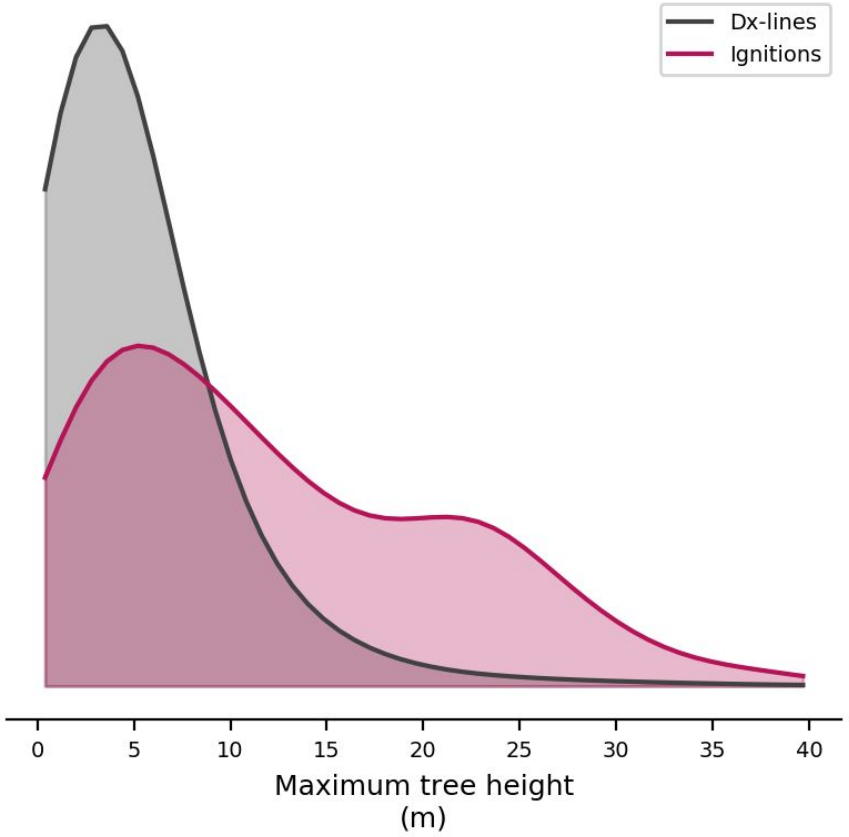
Environmental covariates

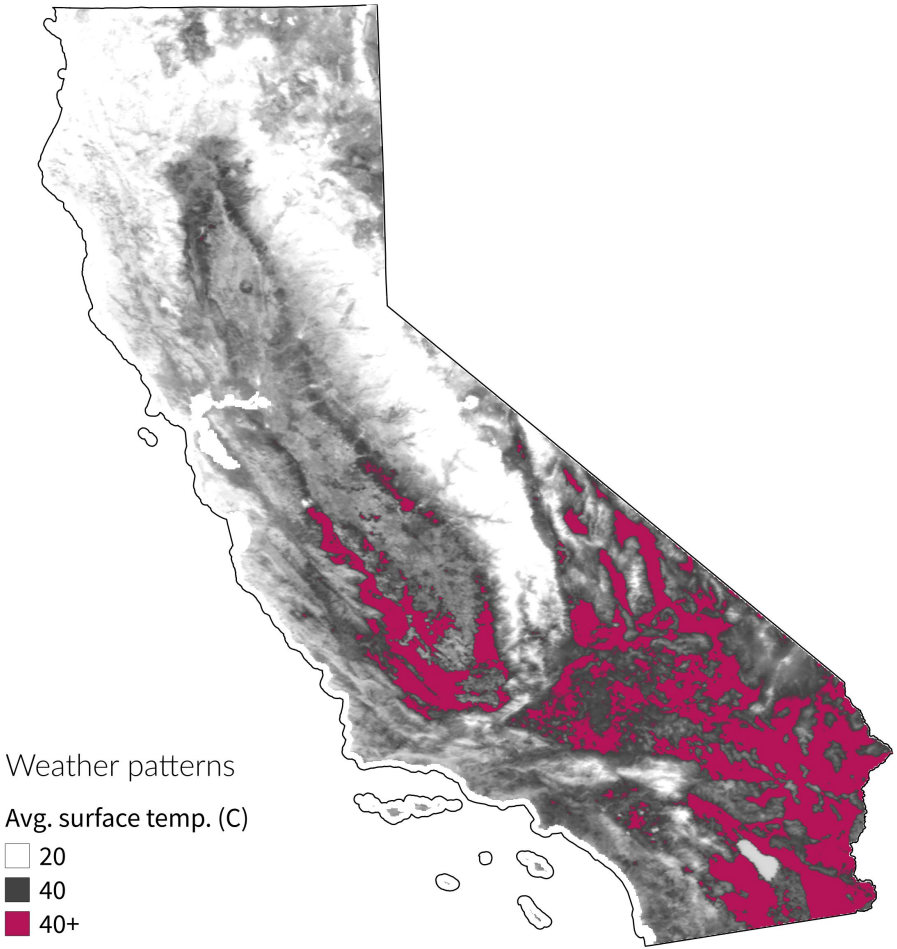
- Vegetation, wind speeds, gust speeds, temperature, topography
- 10 covariates

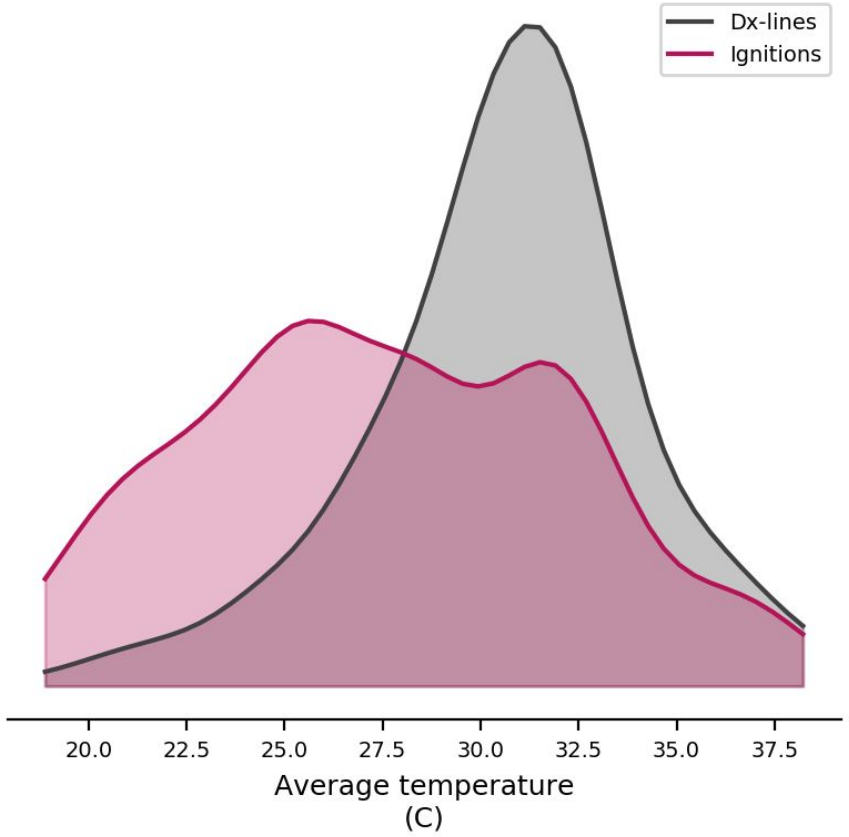
Test locations

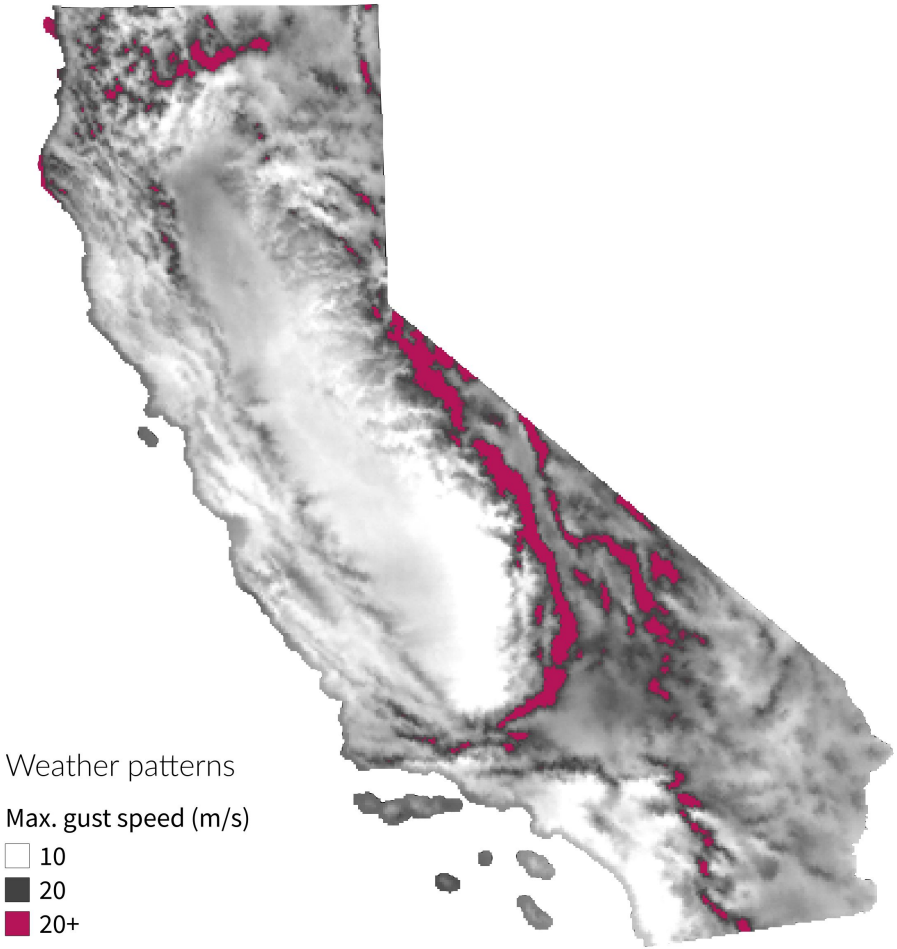
- 2017-2018 ignitions
- 266 points

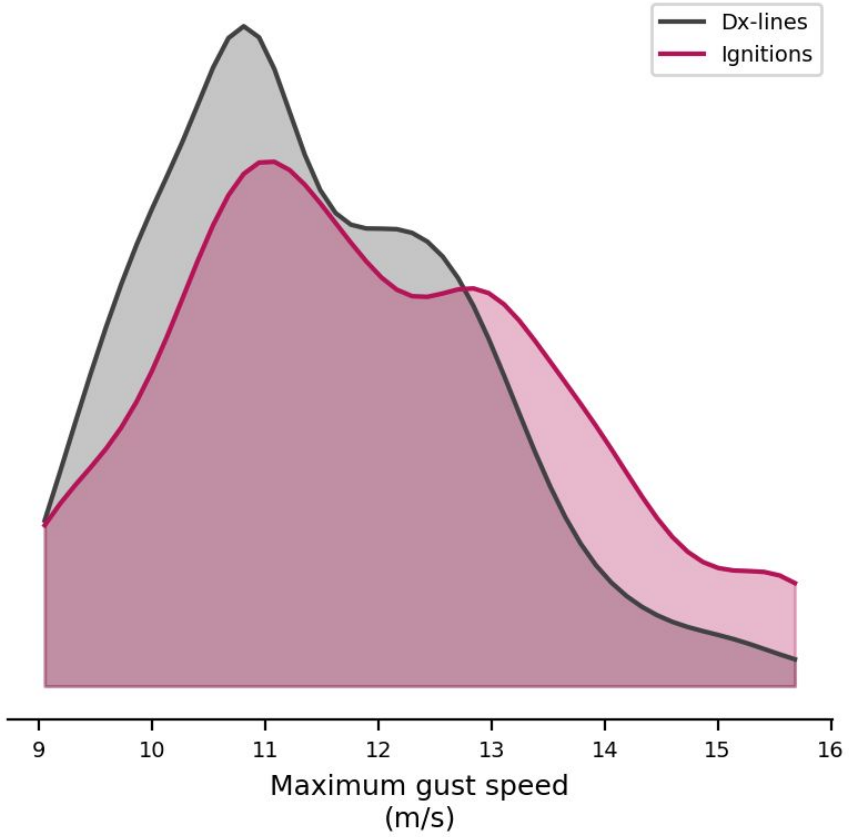

Vegetation structure
Max. tree height (m)
0
20
20+


Dx-lines
Ignitions
0
5
10
15
20
25
30
35
40
Maximum tree height
(m)


Weather patterns
Avg. surface temp. (C)
20
40
40+


Dx-lines
Ignitions
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
Average temperature
(C)


Weather patterns
Max. gust speed (m/s)
10
20
20+

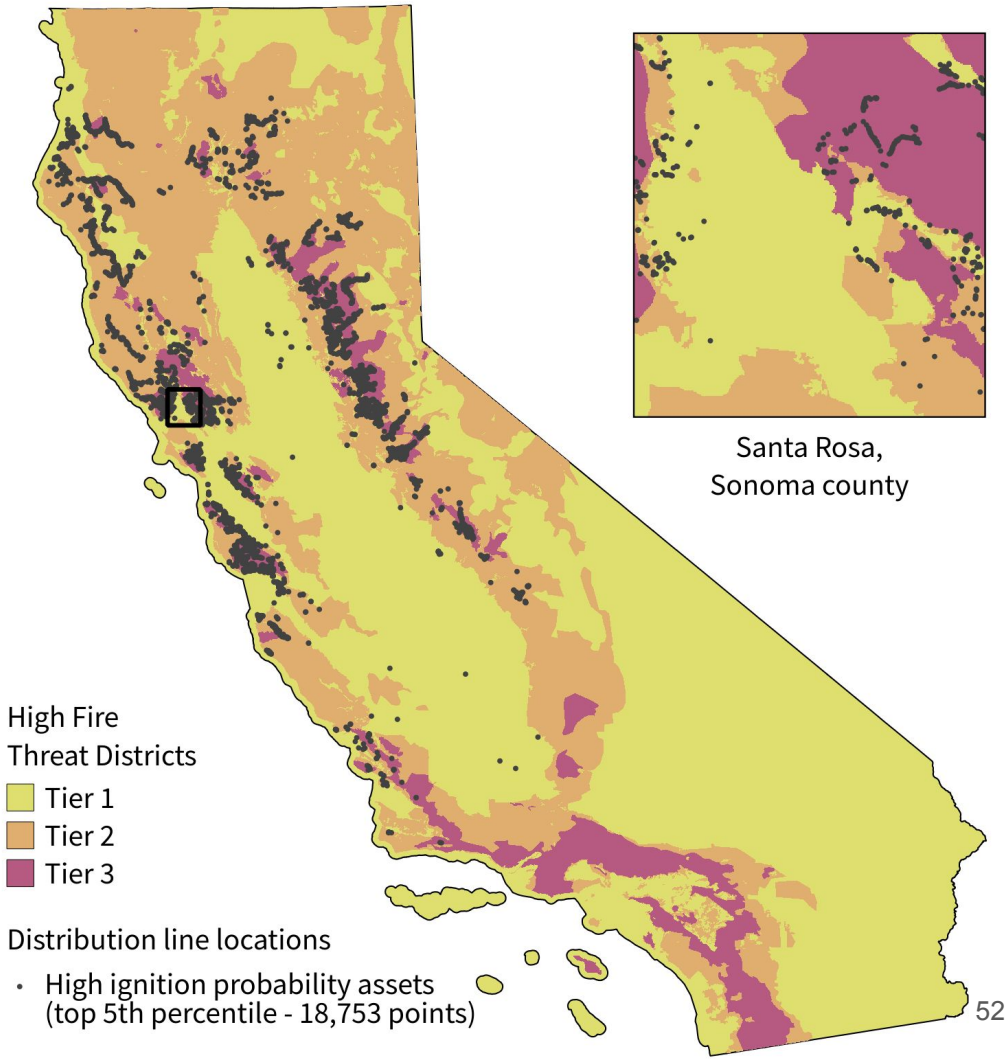

Dx-lines
Ignitions
9
10
11
12
13
14
15
16
Maximum gust speed
(m/s)

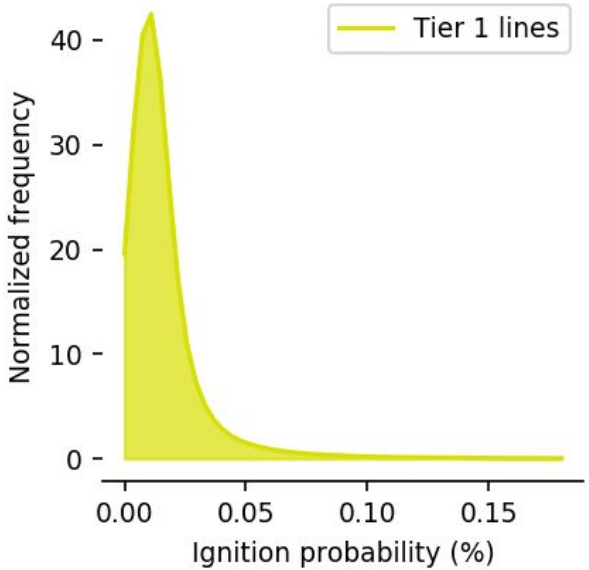
Class	Covariate	Unit	Spatial scale	Notes
Vegetation	Mean tree height	(m)	100 m	Mean tree height of area around asset
	Tallest nearby trees	(m)	100 m	Calculated as maximum tree height in area around an asset
Wind	Mean wind speed	(m/s)	2,500 m	From RTMA
	Local wind speed maximum	(m/s)	2,500 m	Calculated as the 99th percentile of local wind speeds
Gust	Mean gust speed	(m/s)	2,500 m	From RTMA
	Local gust speed maximum	(m/s)	2,500 m	Calculated as the 99th percentile of local gust speeds
Temperature	Mean temperature	(°C)	1,000 m	From MODIS LST
	Local temperature maximum	(°C)	1,000 m	Calculated as the 99th percentile of local temperatures
Topography	Local topographic position	unitless	100 m	From the topographic position index (TPI)
	Landscape topographic position	unitless	1,000 m	Calculating TPI at fine and large scales allows distinguishing multiple landforms (i.e. difference in local and landscape topography)

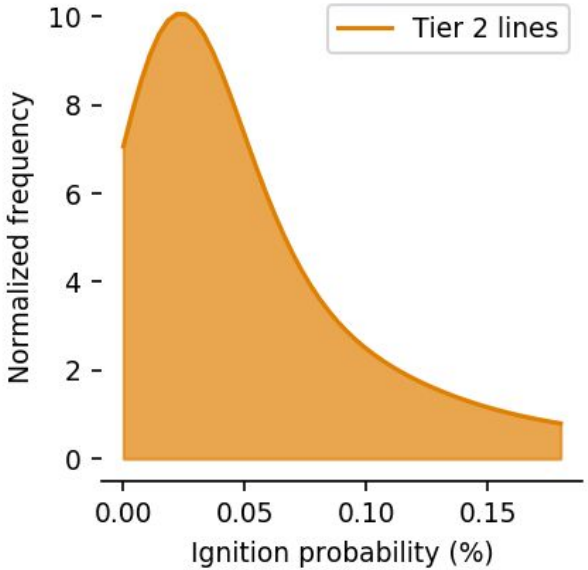
Model outputs

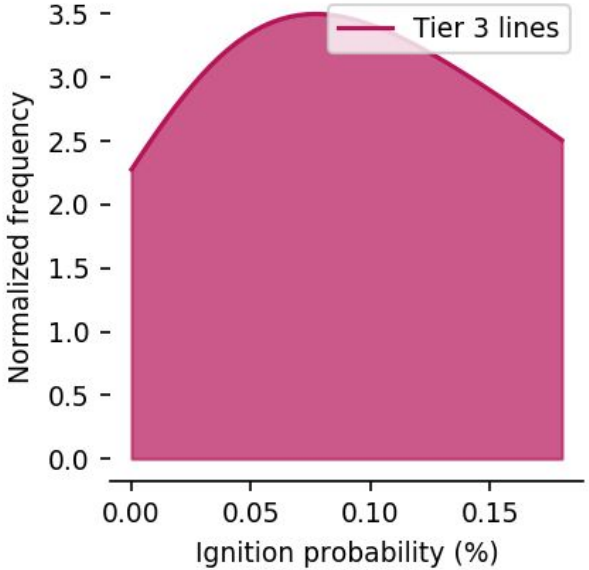
1. Relative probability scores
   - Units: arbitrary
   - Computes ignition probability for each asset using raw probability distributions
   - Evaluated using AUC scores
2. Omission rates
   - Units: %
   - Scales relative probability scores based on the total area evaluated
   - Can threshold rates to evaluate likely/unlikely in binary sense
   - Threshold set to > 5%
   - Evaluated using recall scores
3. Occurrence probability scores
   - Units: %
   - Scales relative probability scores to probability of ignition scores via logistic transformation of raw scores
   - Done via scaling parameter, τ, (the probability of ignition at 'average' ignition locations)
   - τ calculated as (number of total ignitions) / (number of Dx assets evaluated)
   - Evaluated by summing probability scores and comparing to number of ignitions

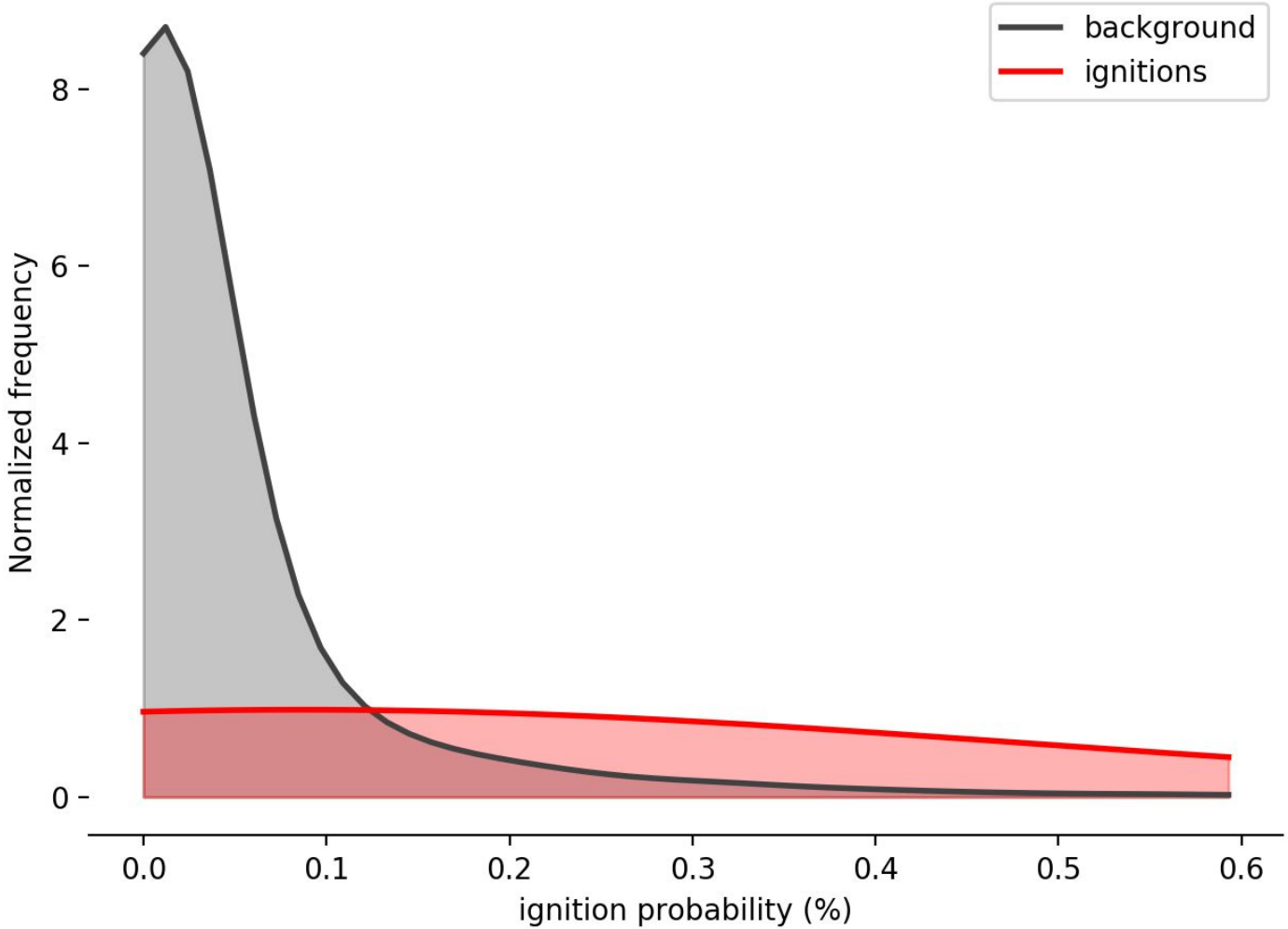
	Training 2015-2016	**Testing** 2017-2018
Predicted ignition count	229.1	200.0
Observed ignition count	210	266

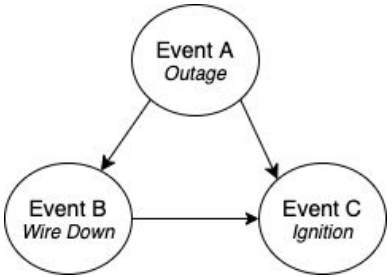



Tier 1 lines
Normalized frequency
40
30
20
10
0
0.00
0.05
0.10
0.15
Ignition probability (%)


Tier 2 lines
Normalized frequency
10
8
6
4
2
0
0.00
0.05
0.10
0.15
Ignition probability (%)


Tier 3 lines
Normalized frequency
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.00
0.05
0.10
0.15
Ignition probability (%)


background
ignitions
Normalized frequency
8
6
4
2
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
ignition probability (%)


Event A
Outage
Event B
Wire Down
Event C
Ignition

Vegetation ignitions

Variable	Percent contribution	Permutation importance
tree-height-max	35.1	38.6
tree-fall-in	31.9	7.2
hftd	8.8	3.5
local-topography	5	9.4
canopy-stress	4.9	5.7
temperature-avg	4.6	9.9
impervious	2.8	4.8
conductor-count	2.6	5.7
specific-humidity-avg	1.4	1.9
tree-height-avg	1.3	8.1
precipitation-avg	0.9	0.8
wind-avg	0.4	1.8
1000-hour-fuels-avg	0.3	1.1
energy-release-avg	0.1	1
burn-index-avg	0	0.5
wind-max	0	0
vapor-pressure-deficit-avg	0	0
100-hour-fuels-avg	0	0