Appraiser : How Airbnb Generates Complex Models in Spark for Demand Prediction

hector.yee@airbnb.com

@eigenhector

Hector Yee

Selected publications

8 Movies & games (Shrek2, Star Wars, C&C etc)

Google Image Search (4 patents)

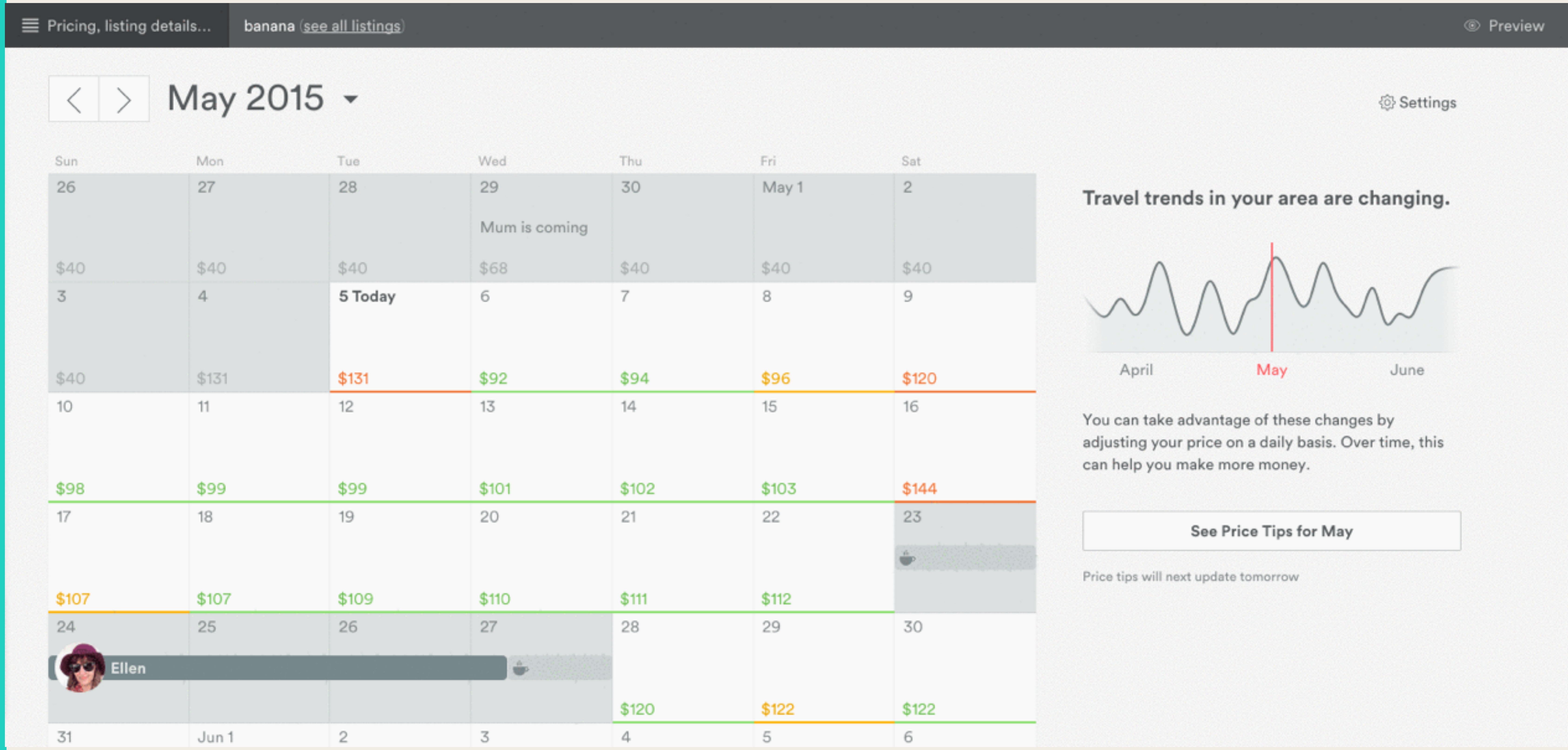
- Image classification, YH Yee et al, US Patent 8,478,052
- Customizing image search for user attributes, YH Yee, CJ Rosenberg,US Patent 8,782,029

Google Self-driving Car - Perception (1 patent)

Youtube Technical Emmy 2014 (4 papers)

- Label partitioning for sublinear ranking, J Weston, A Makadia, H Yee, ICML 2013
- Affinity weighted embedding, J Weston, R Weiss, H Yee, ICML 2014

Price tips



May 2015

Sun	Mon	Tue	Wed	Thu	Fri	Sat
26 $40	27 $40	28 $40	29 Mum is coming $68	30 $40	May 1 $40	2 $40
3 $40	4 $131	5 Today $131	6 $92	7 $94	8 $96	9 $120
10 $98	11 $99	12 $99	13 $101	14 $102	15 $103	16 $144
17 $107	18 $107	19 $109	20 $110	21 $111	22 $112	23
24 Ellen	25	26	27	28 $120	29 $122	30 $122
31	Jun 1	2	3	4	5	6

Travel trends in your area are changing.

You can take advantage of these changes by adjusting your price on a daily basis. Over time, this can help you make more money.

Price tips will next update tomorrow

Pricing in a Two-Sided Marketplace

- **Goal:** Equip hosts with a tool to make better informed pricing decisions to meet their needs
- Price tips only

Overview

- **Modeling**
- **Aerosolve:** open-source ML stack

Modeling

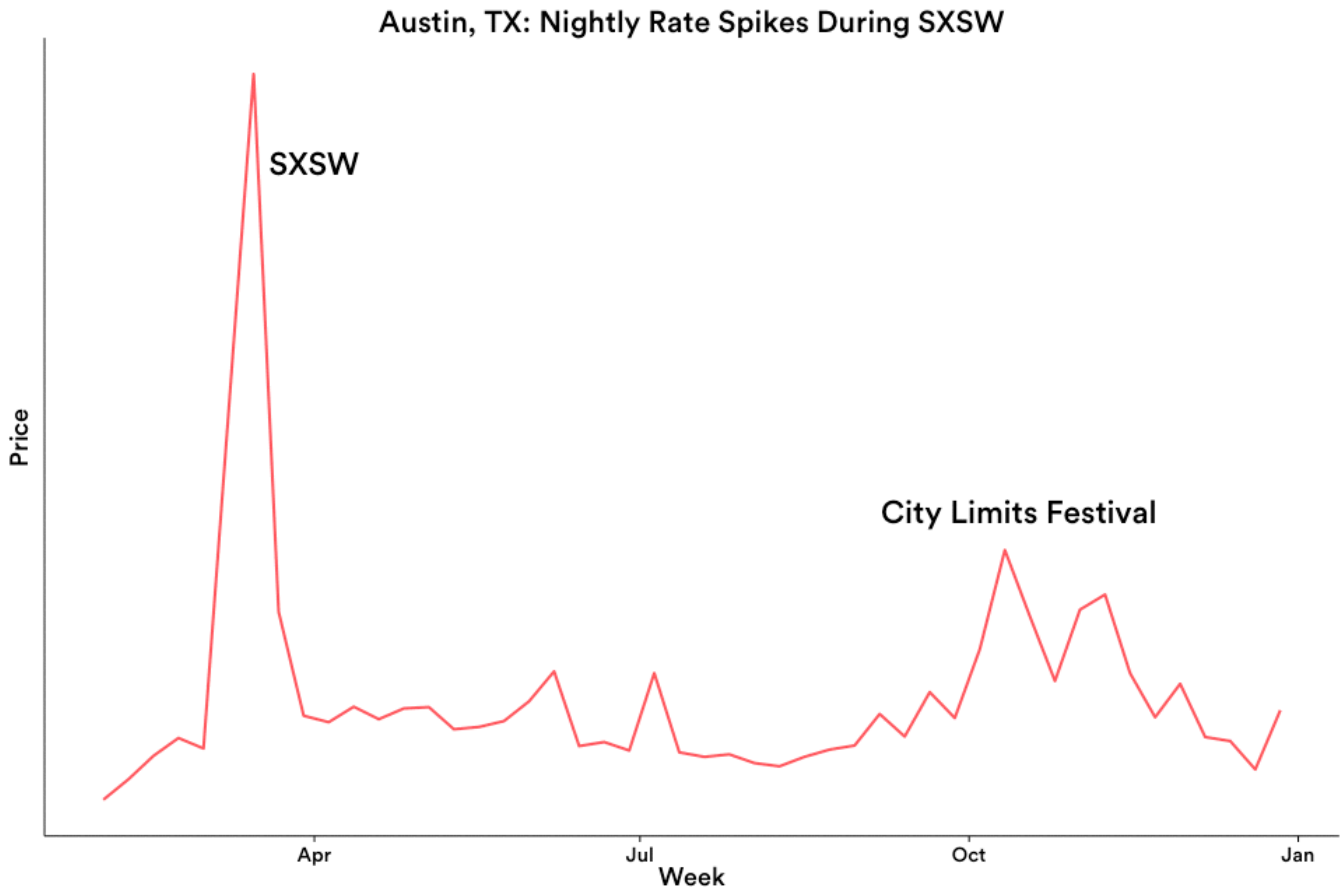
Modeling Approach and Scale

- **Modeling goal: Predict the probability of a booking for each price**
- Scale
  - O(Million) of derived features
  - Over 5B training data points. Proportional to (# listings) x (# days)

What Affects Prices?

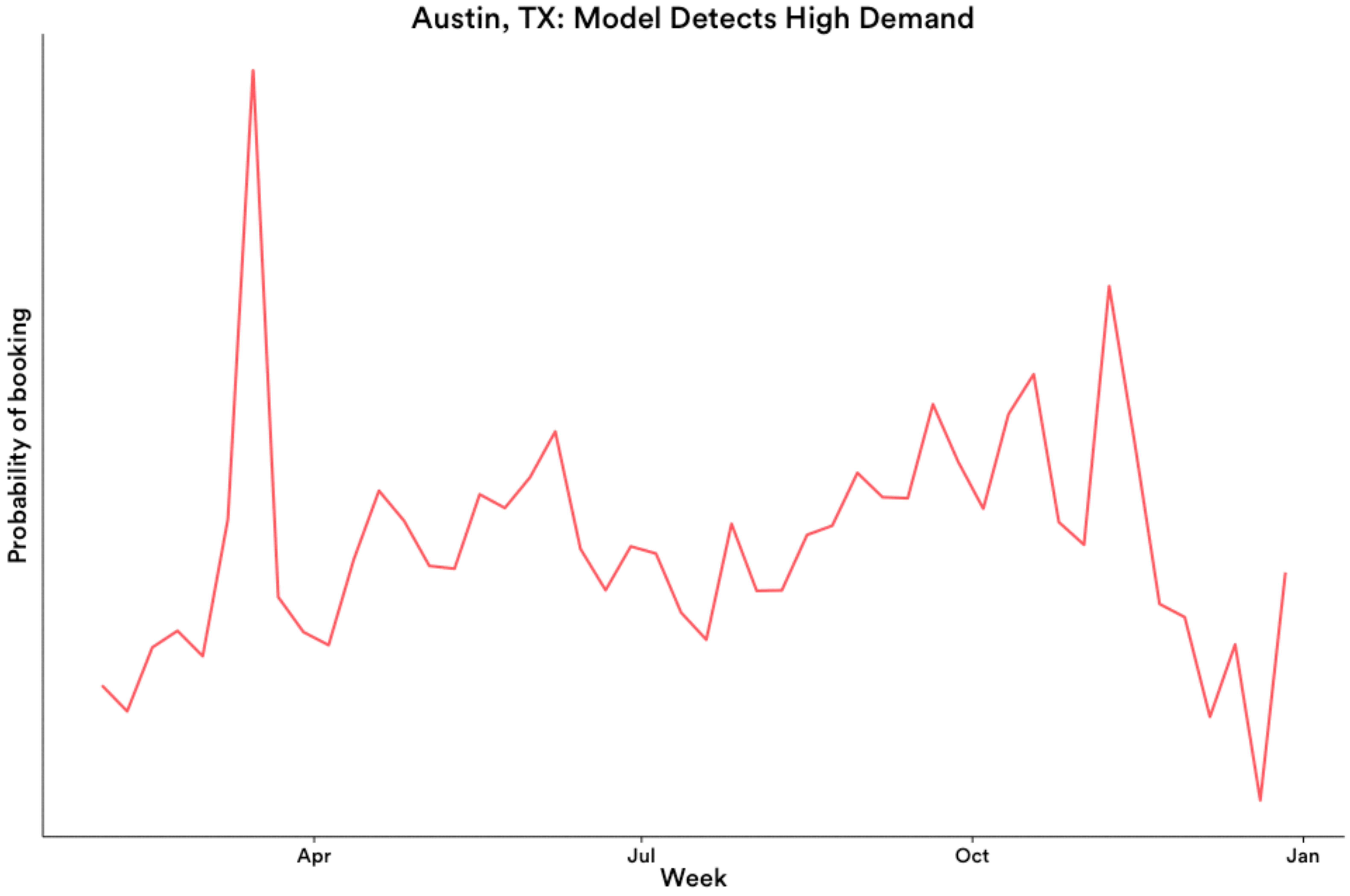
- **Demand** (seasonality, events)
- **Listing location** (market, neighborhood, street block)
- **Listing type and quality**

Seasonality & Events



Seasonality & Events

Demand captured by multiple features



Listing Location

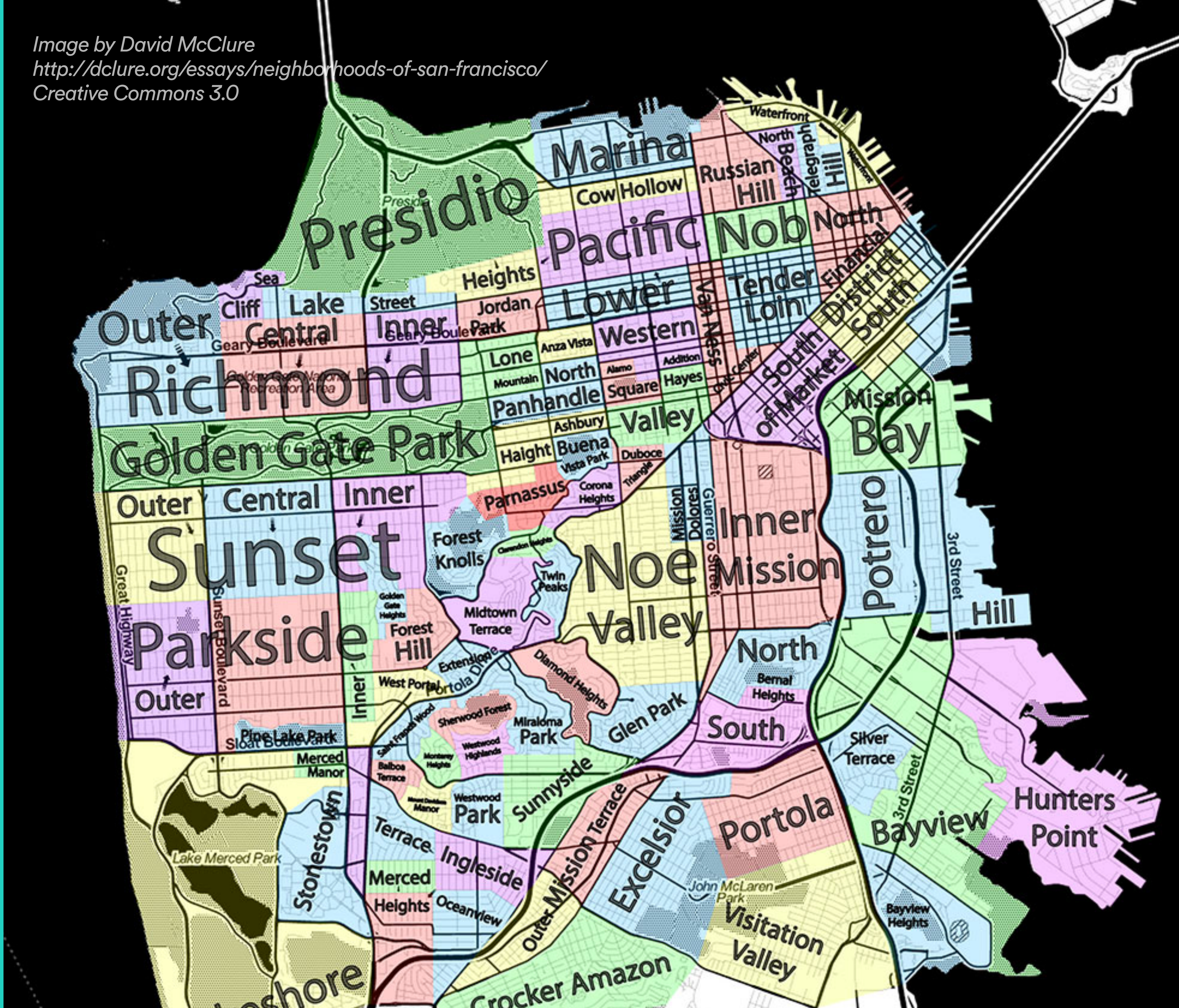
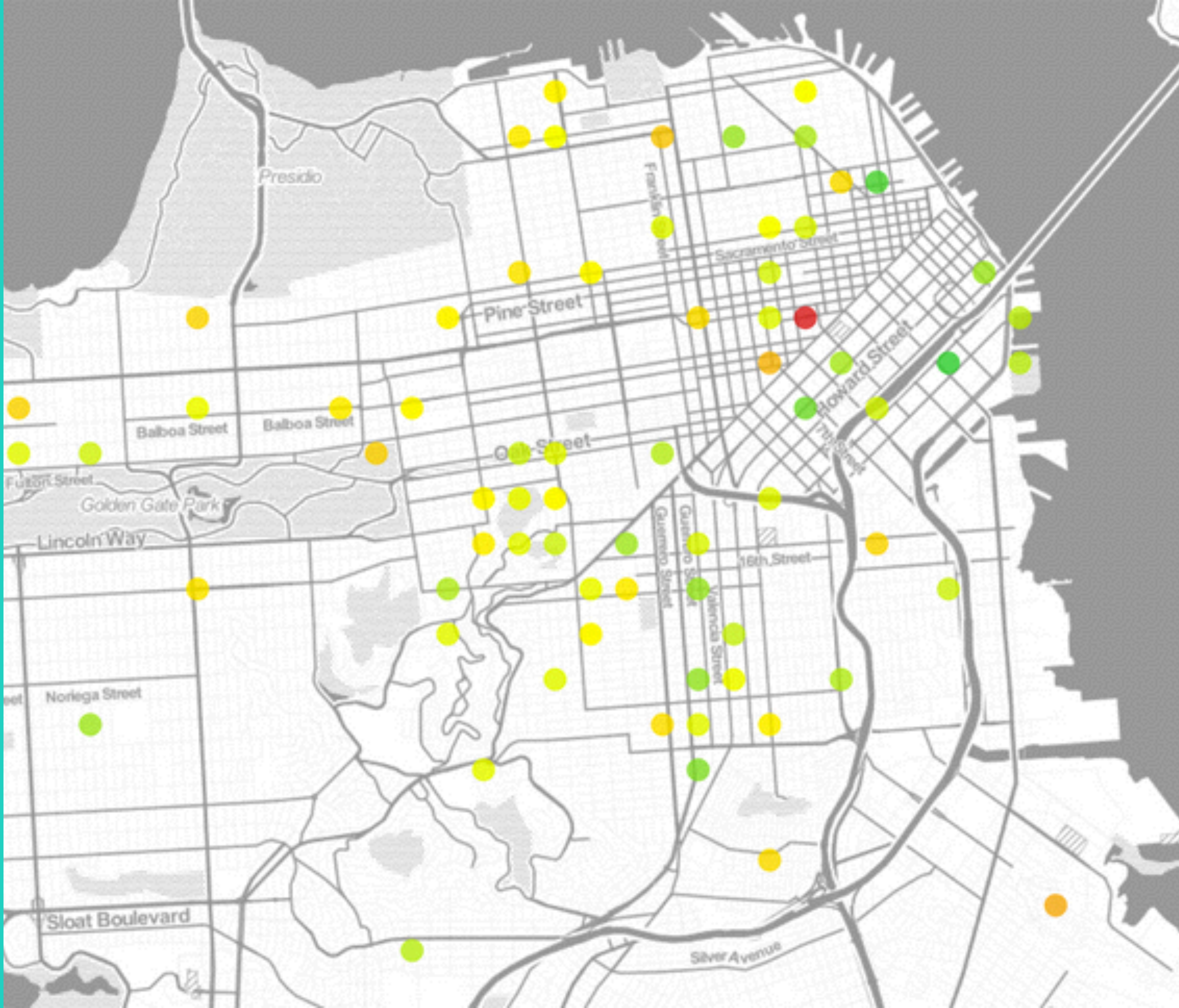


Image by David McClure
http://dclure.org/essays/neighborhoods-of-san-francisco/
Creative Commons 3.0

Listing Location

Grids / kdtree captures value of locations in San Francisco



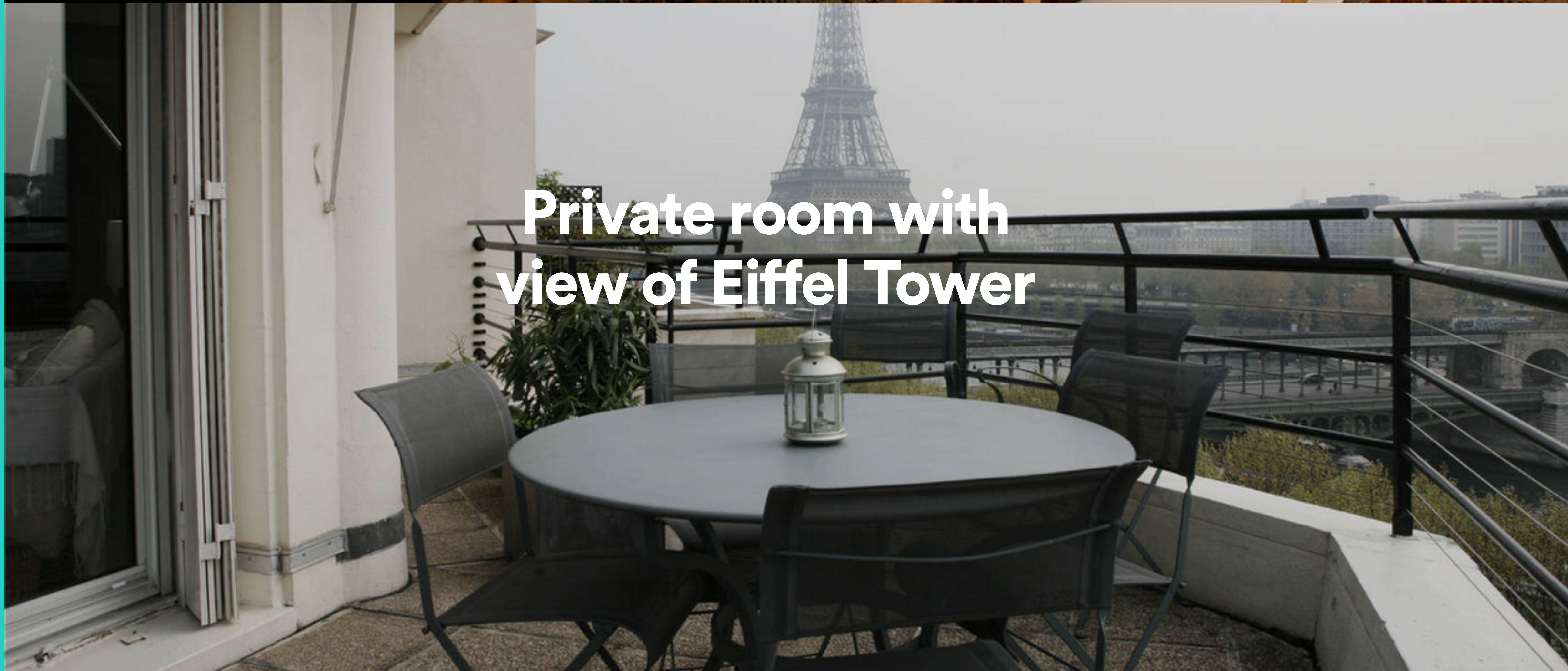
Listing Type & Quality

Entire apt or a room

Amenities

Guest reviews

Entire home/ houseboat across from Eiffel Tower



Private room with view of Eiffel Tower

Aerosolve : Machine Learning *for humans*

Free parking near SFO Airport?

(37.68359375,-122.40234375

```
model.filter(x => x.featureName.contains(“Free Parking”))
```

OUTER MISSION
VALLEY
Geneva Ave
Tunnel Ave
1
35
82
280
WESTLAKE
Daly City
San Bruno
Mountain State Park
Colma
101
1
35
280
Grand Ave
South San
Francisco
King Dr
280
1
The Shops at Tanforan
San Bruno
35

Sabbia (sand Italian)

(20.63671875,-87.06640625)

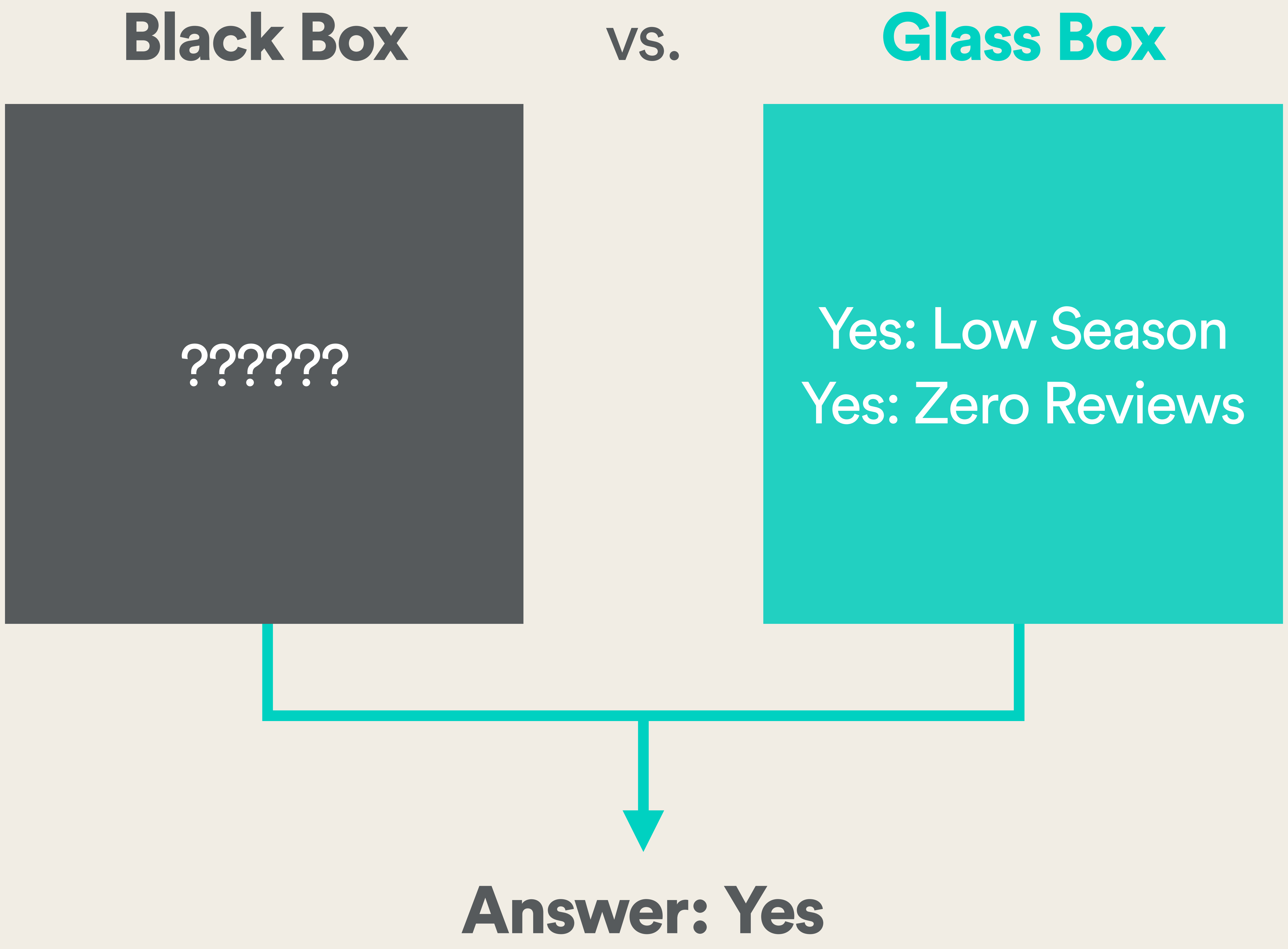
```
model.filter(x => x.featureName.contains(“Sabbia”))
```

Calle J
Grand Coral Riviera Maya
El Marlyn Azul
La Cruz de Servicios - TIERRAZUL
Calle Av 10
Av Constituyentes
30 Avenida Nte
Playa del Carmen
Calle 11 Sur
PLAYACAR
© 2015 Google
© 2015 Google

Machine Learning for Humans

Interpretability

"Is my listing priced too high?"



Black Box

vs.

Glass Box

??????

Yes: Low Season
Yes: Zero Reviews

Answer: Yes

Question : Will I get a booking at this price?

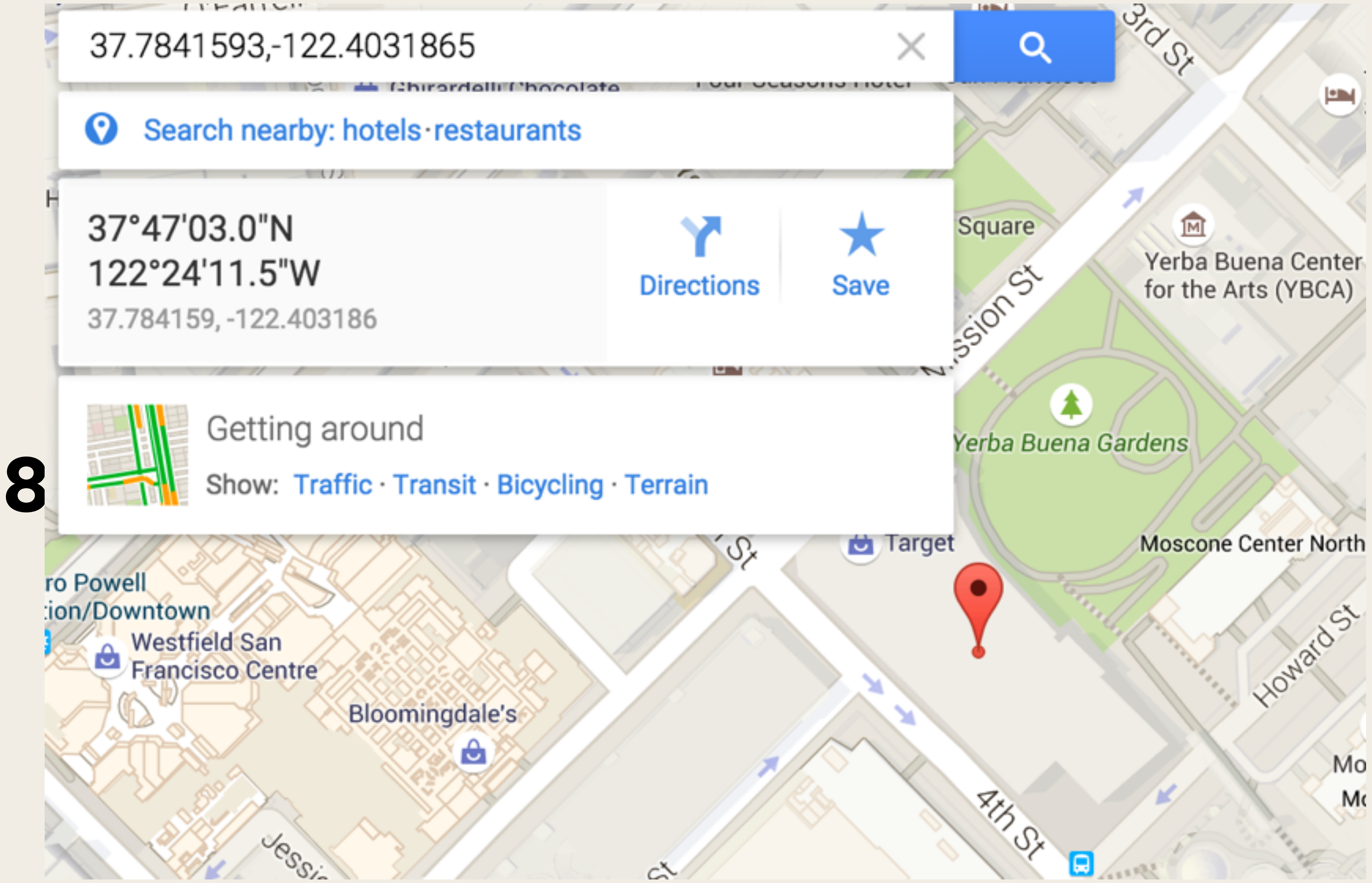
- Latitude
- Longitude
- Price

Metreon

Lat :37.7841593

Long : -122.40318

Price : 500



37.7841593,-122.4031865

Search nearby: hotels · restaurants

37°47'03.0"N
122°24'11.5"W
37.784159, -122.403186

Getting around

Show: Traffic · Transit · Bicycling · Terrain

Black box

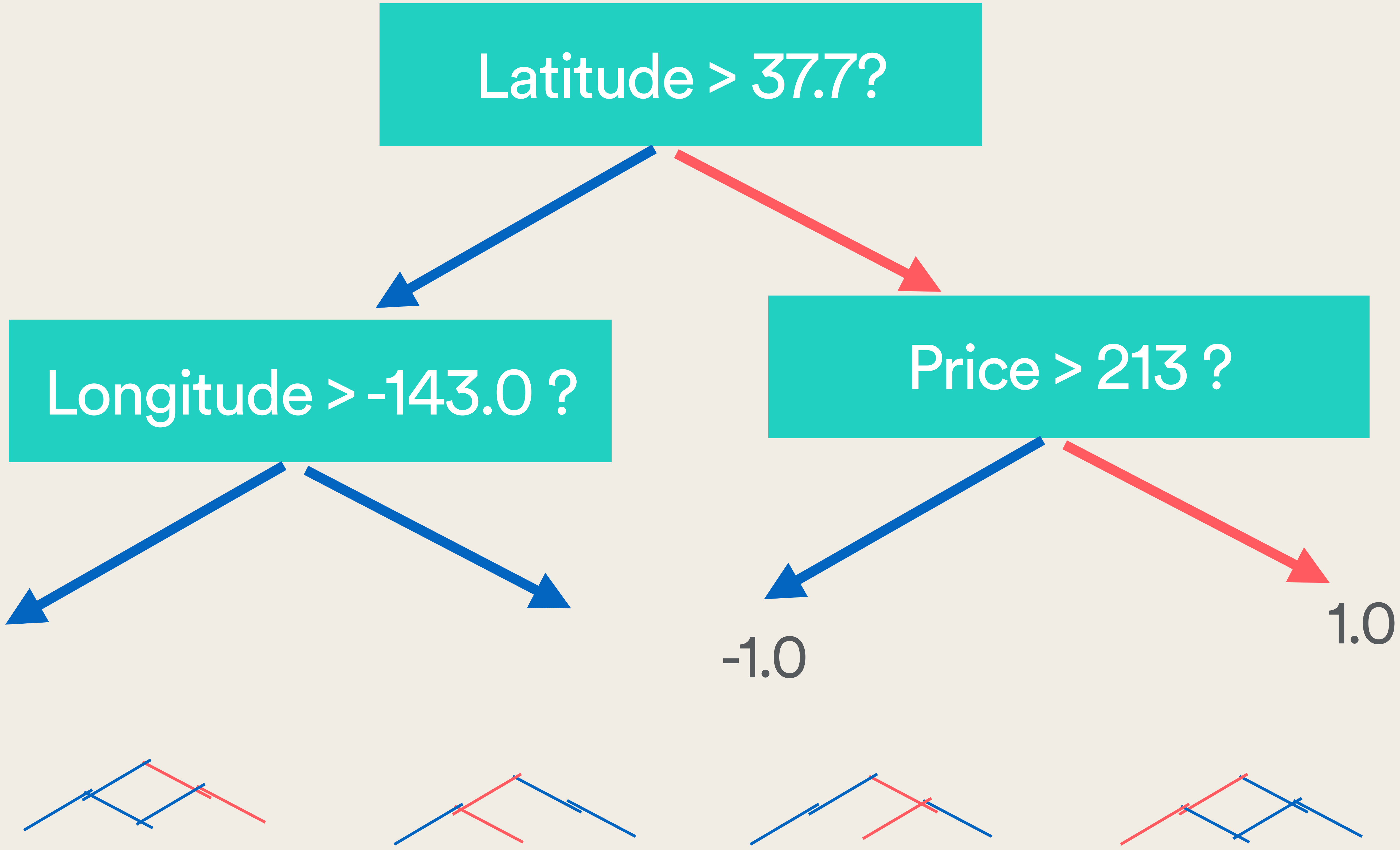
Random forest

Q: Will I get a book?

Lat :37.7841593

Long : -122.4031865

Price : 500



- Difficult to interpret
    - 1000s of trees in a forest
    - Not clear relationship between variables and target

Glass box model

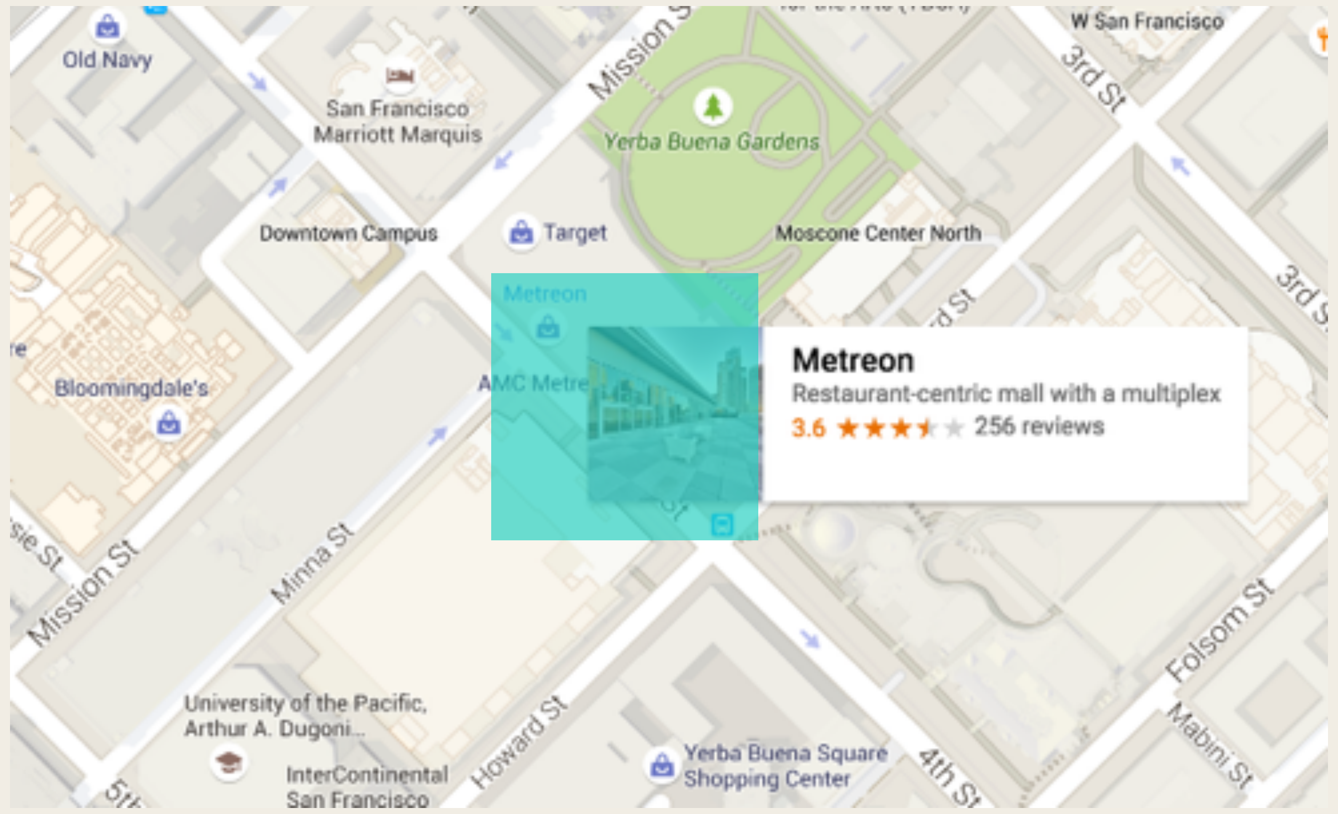
Lat :37.7841593

Long : -122.4031865

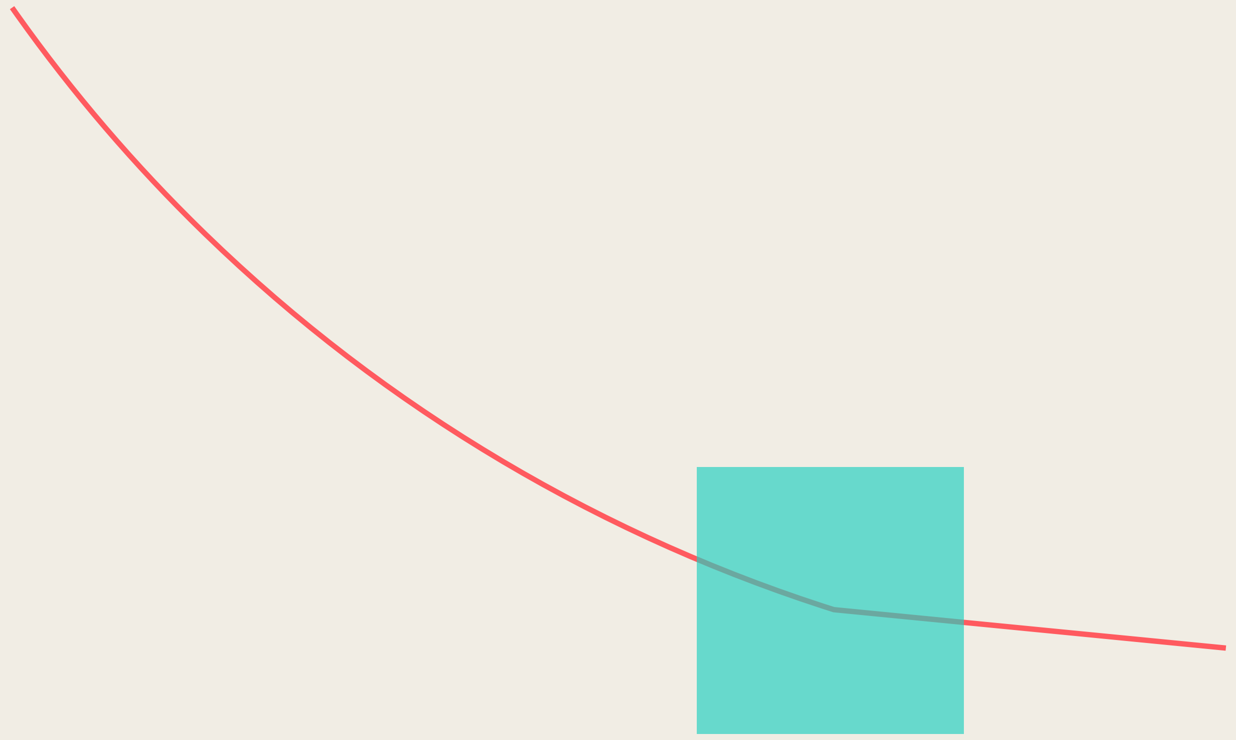
Price : 500

- Control quantization
- Control interaction (crosses)
- O(millions) of sparse parameters
- O(tens) active any time

(37.7, -122.4) AND Price in [500, 550]



Good location



High price!

Feature transforms

Control feature engineering

```
quantize_listing_location {
  transform : multiscale_grid_quantize
  field1: "LOC"
   buckets : [ 0.1, 0.01 ]
  value1 : "Latitude"
  value2 : "Longitude"
  output : "QLOC"
}

quantize_price {
  transform : multiscale_quantize
  field1: "PRICE"
  buckets : [ 10.0, 100.0 ]
  value1 : "$"
   output : “QPRICE" }
```

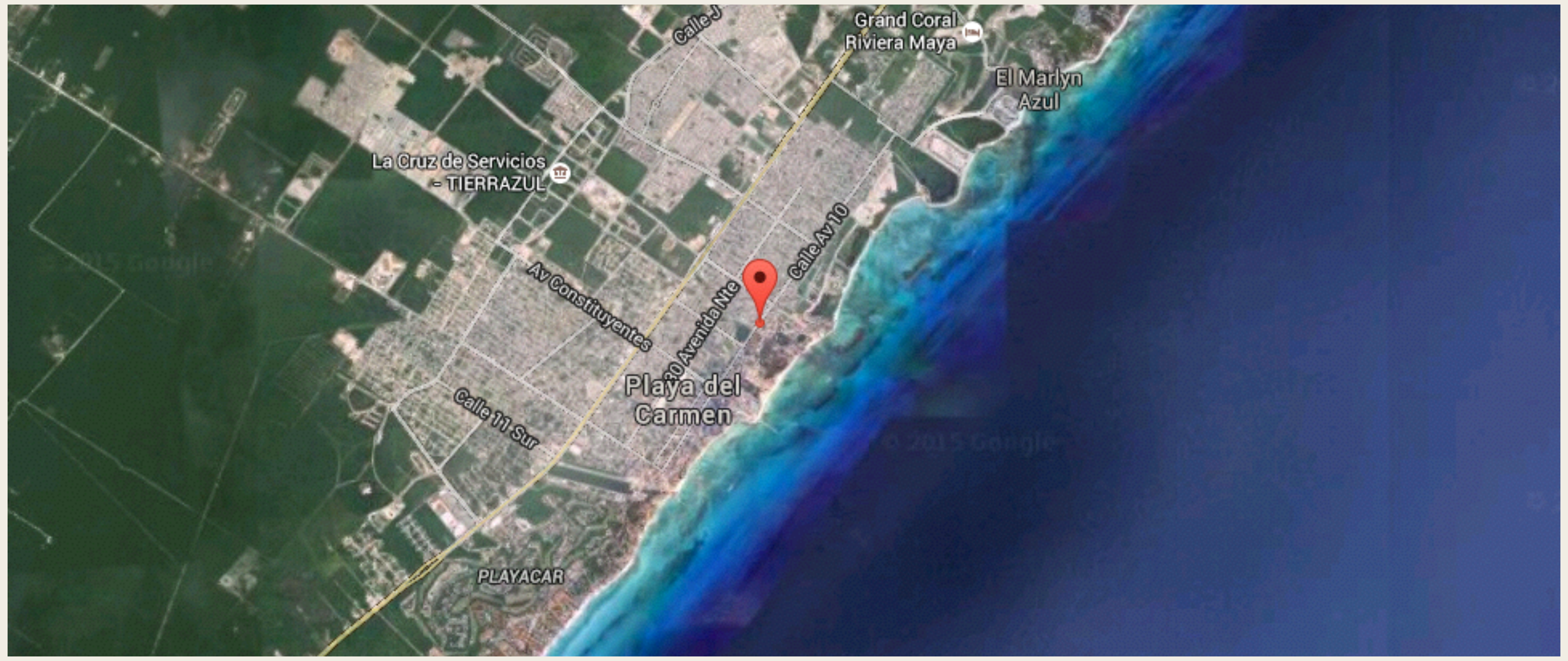
```
Price_X_Location {
  transform : cross
  field1 : "QPRICE"
  field2 : "QLOC"
  output : “PRICE_AND_LOCATION”
}

combined_transform {
  transform : list
  transforms : [
   quantize_listing_location,
   quantize_price,
   Price_X_Location ]
}
```

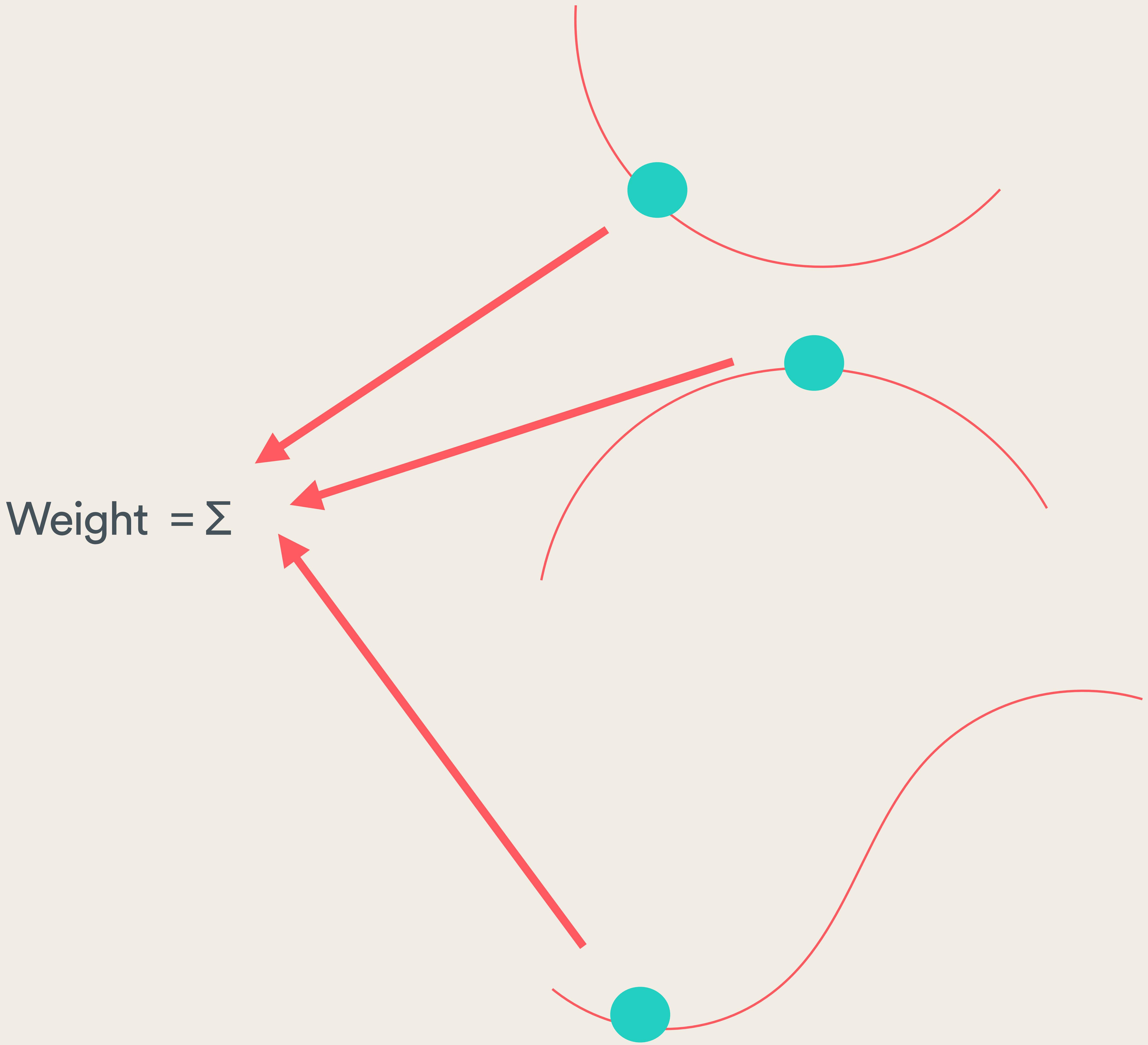
Benefits of feature transforms

- Write training data once
- Iterate feature transforms on the fly
- Control quantization
- Control interaction (crosses)
- Debuggable models (graphs + readable)

(20.6,-87.0) ^ Sabbia = Good!

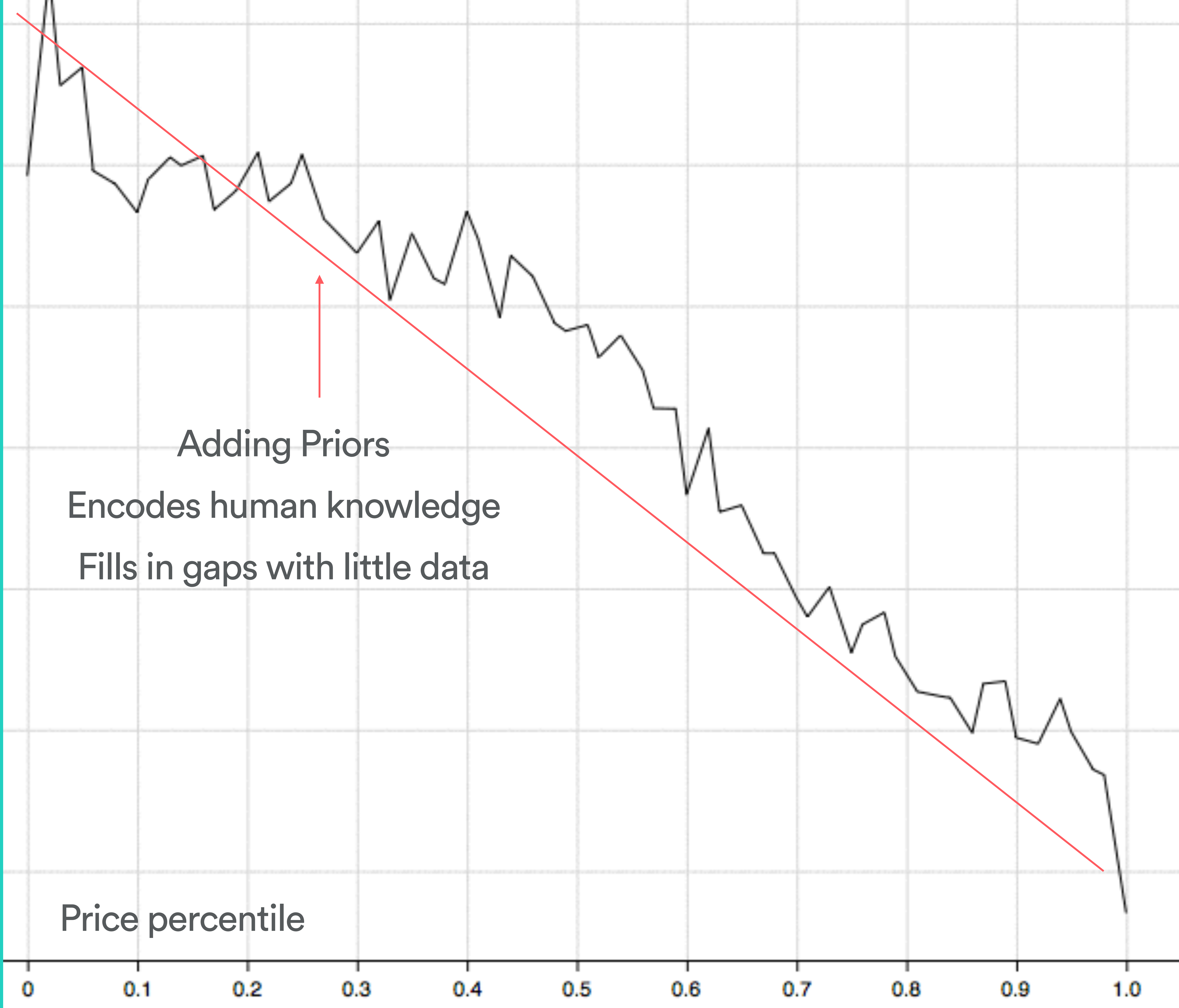


Additive splines



Price percentile

Approximate price percentile in market



Adding Priors

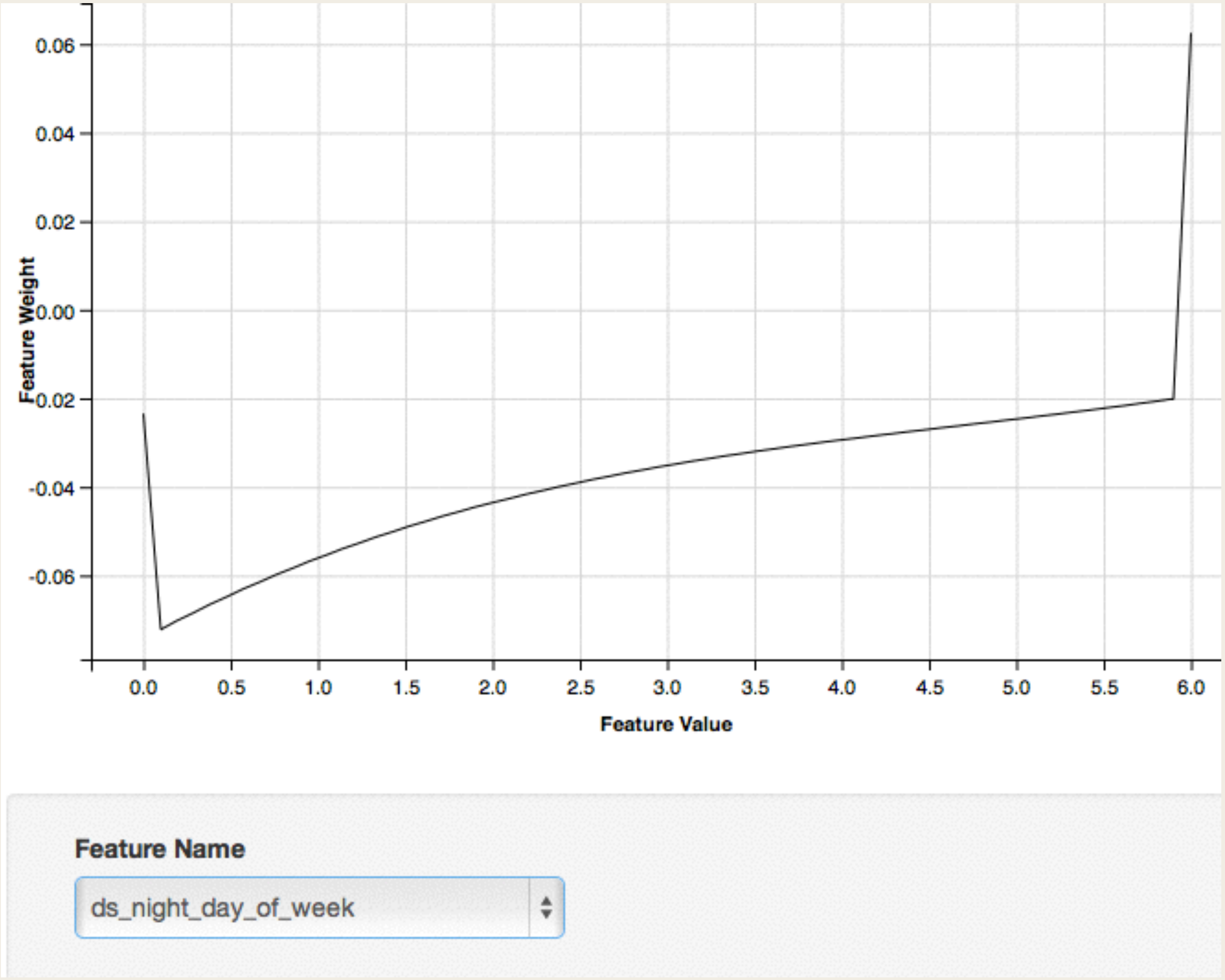
Encodes human knowledge

Fills in gaps with little data

Bézier Cubic

With Dirac Deltas

End points can vary a lot

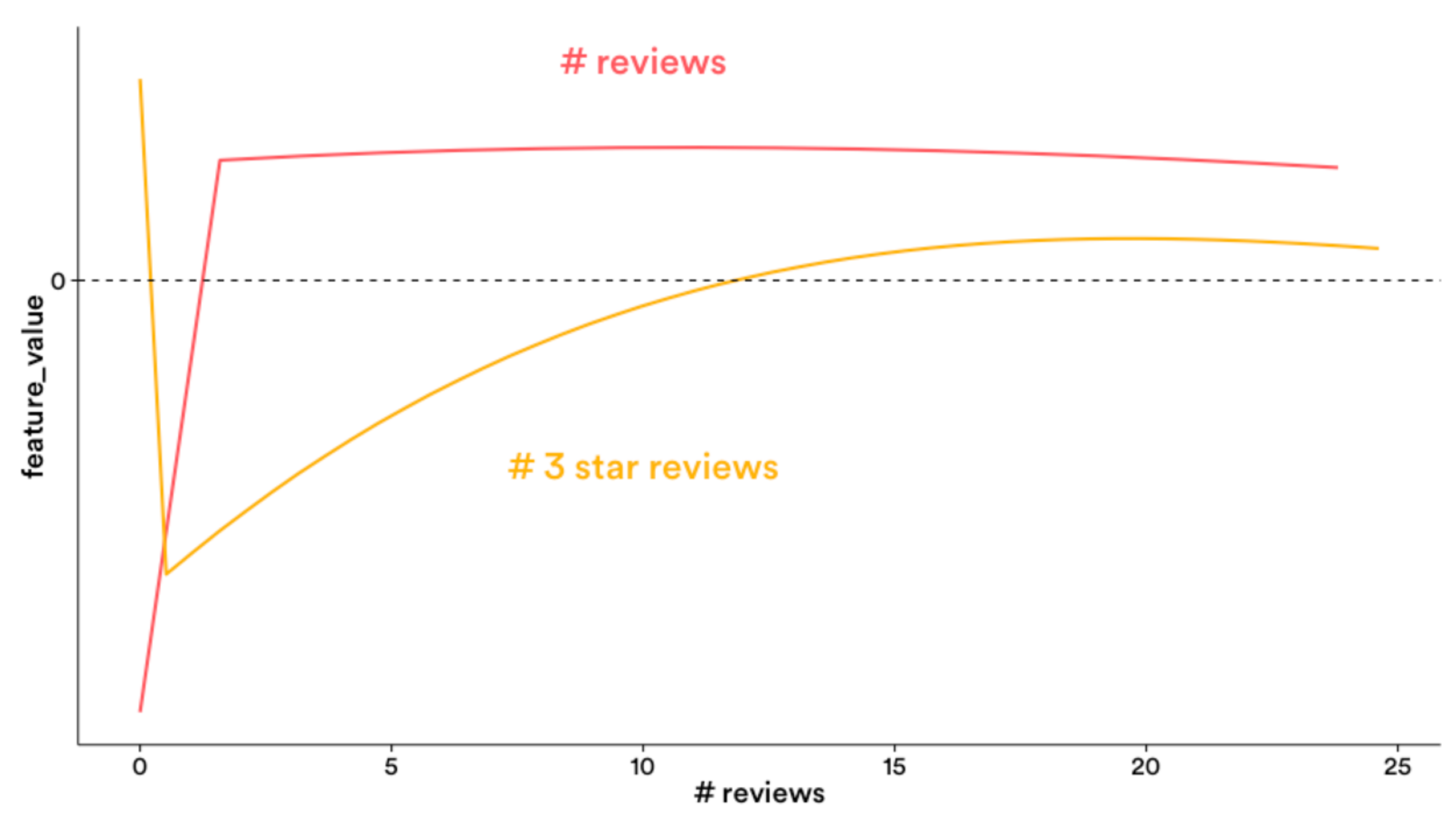


Feature Name

ds_night_day_of_week

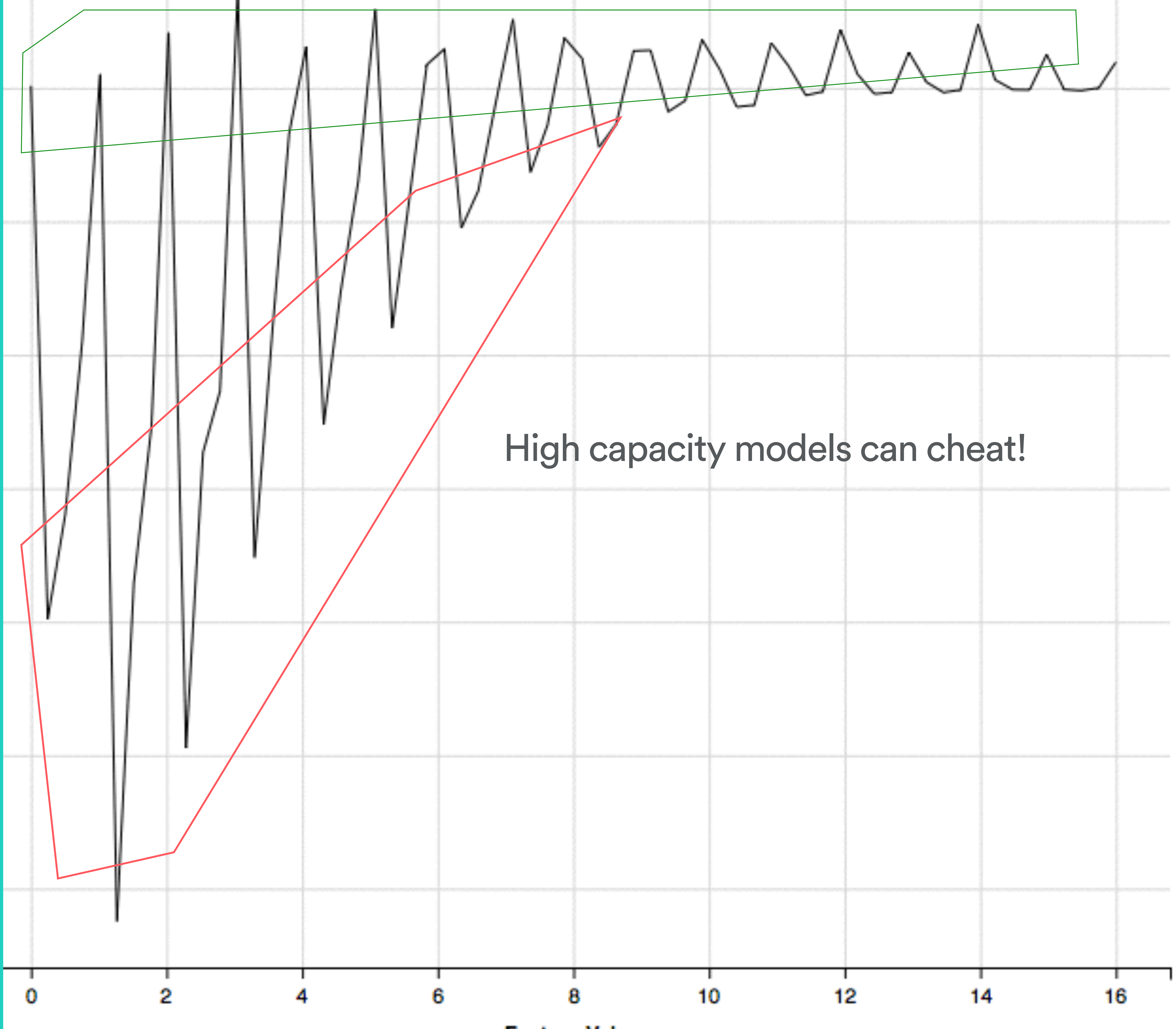
Listing quality

Dirac + Cubic spline captures effect of reviews



Debugging Splines

When splines go bad and weebly wobbly



High capacity models can cheat!

Overfitting

Dangers of high capacity models

L1 regularization

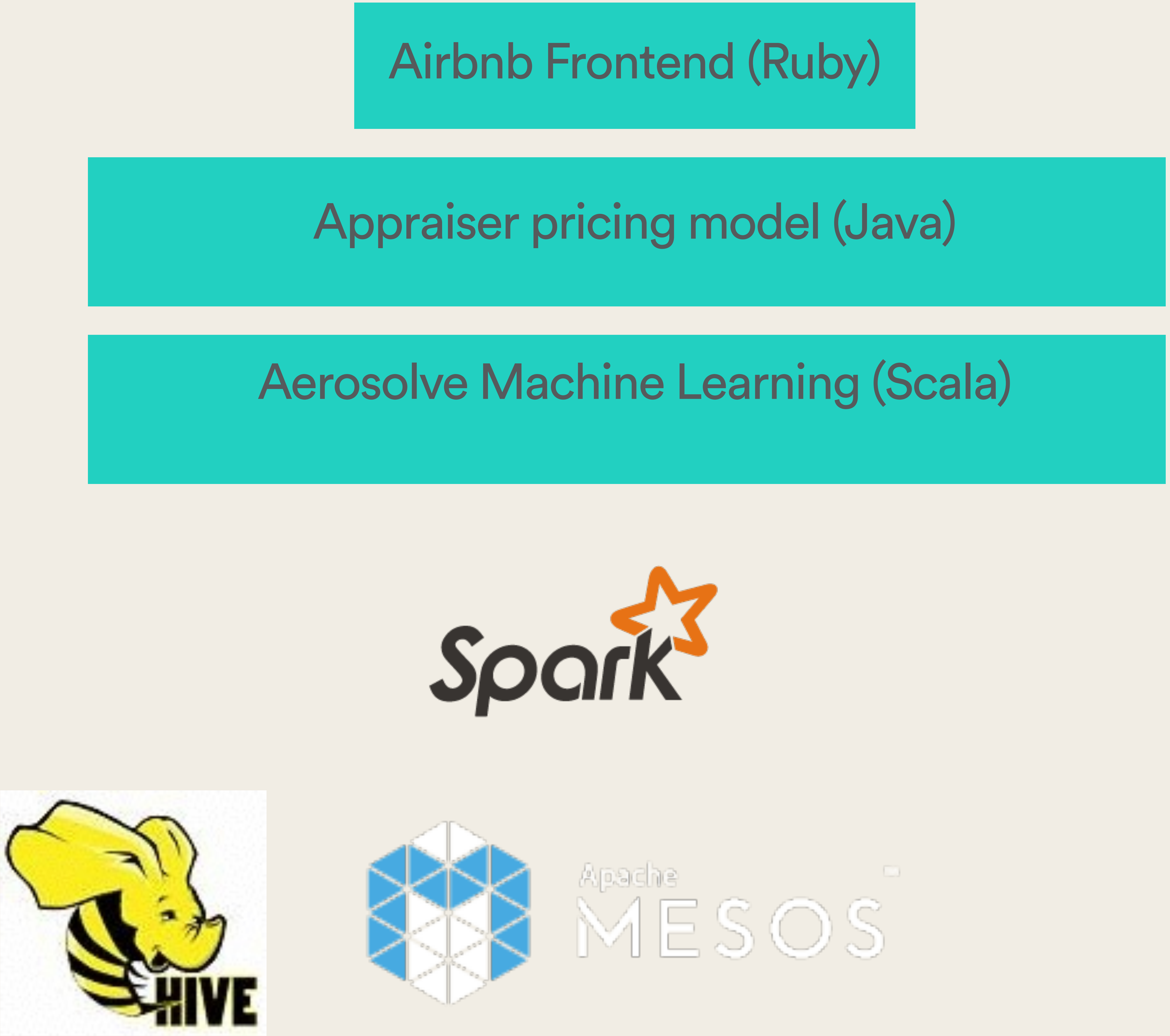
- Starts learning ID like features
  - Exact location
  - Time of creation
  - Damages Splines

For Linear model:
L1 + L2 + Random Dropout

For Spline model:
L_infinity spline group norm +
Change of basis (Bézier Cubic - 4 params) +
Dropout for spline model

Miscellaneous

Achitecture

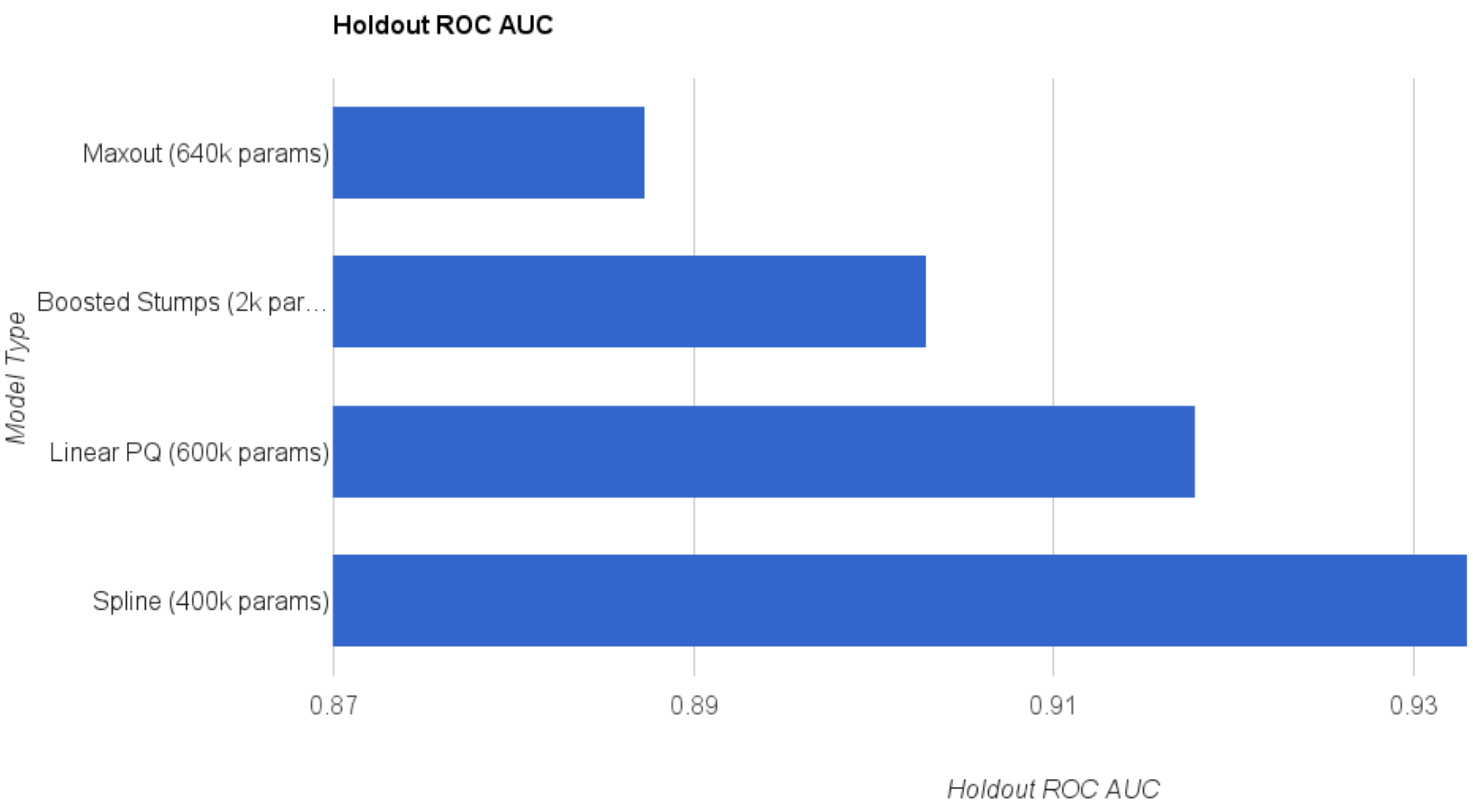


Airbnb Frontend (Ruby)

Appraiser pricing model (Java)

Aerosolve Machine Learning (Scala)

Kinds of models



KD Trees

Bodies of water kept separate

Min Sum P(leaf_i | parent) * n_leaf_i

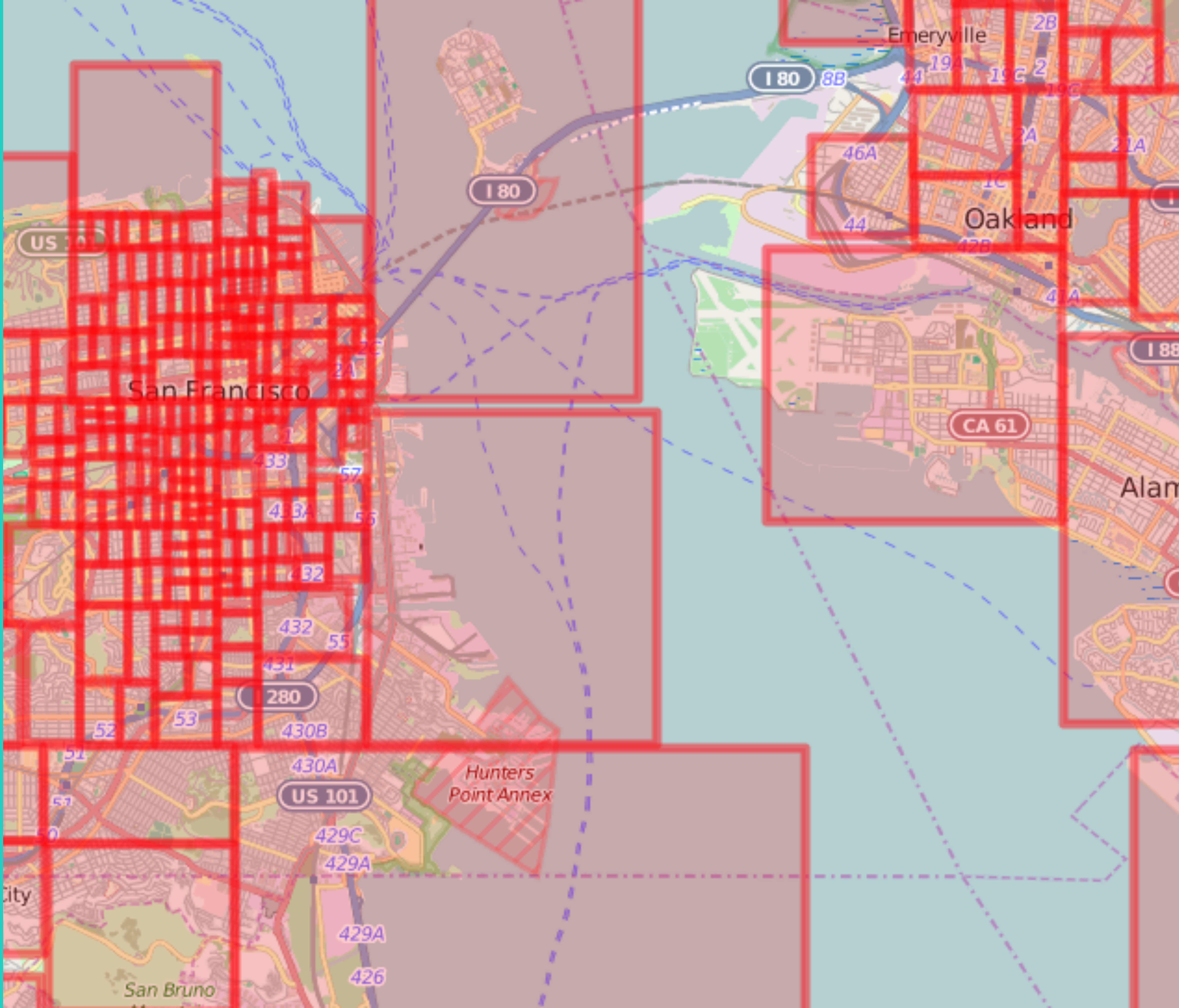


Image Ranking

Human curated vs. organic bookings

Ornate
Living Rooms

vs.

Creature
Comforts

Generate Features

- RGB, HSV, Edge histograms, Texture histograms

Sparsify

- Winner Take All Hash (max of random elements)

Ranking Loss

Demo : Learning to paint

http://airbnb.github.io/aerosolve/

Questions?

@eigenhector