Beijing City Feature-Based Airbnb Price Prediction

Team GeekHub

Brian Feng
Mike Liu
Tony Li
Logan Luo

Outline

Motivation

Problem Definition

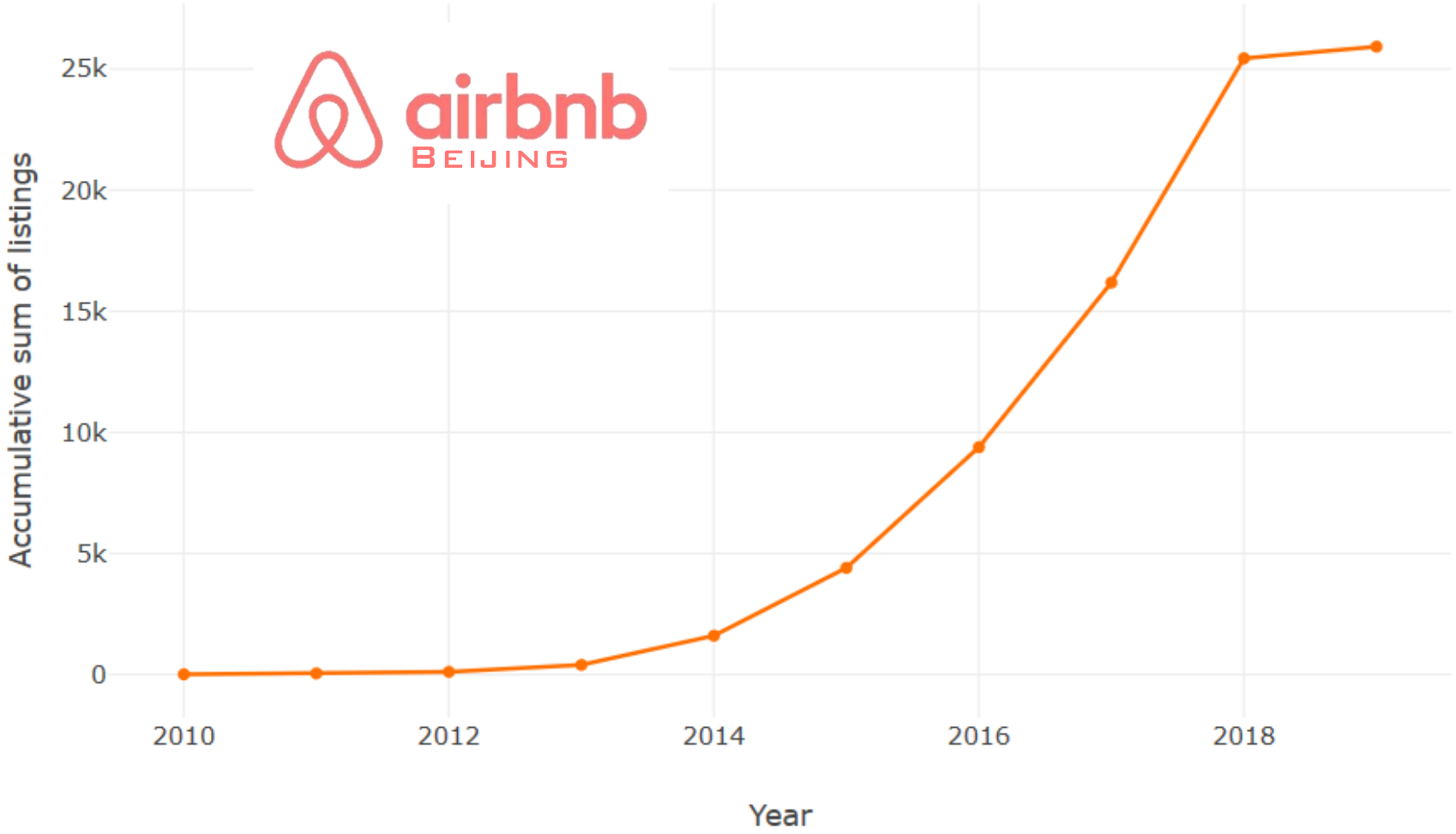
State-of-the-Art

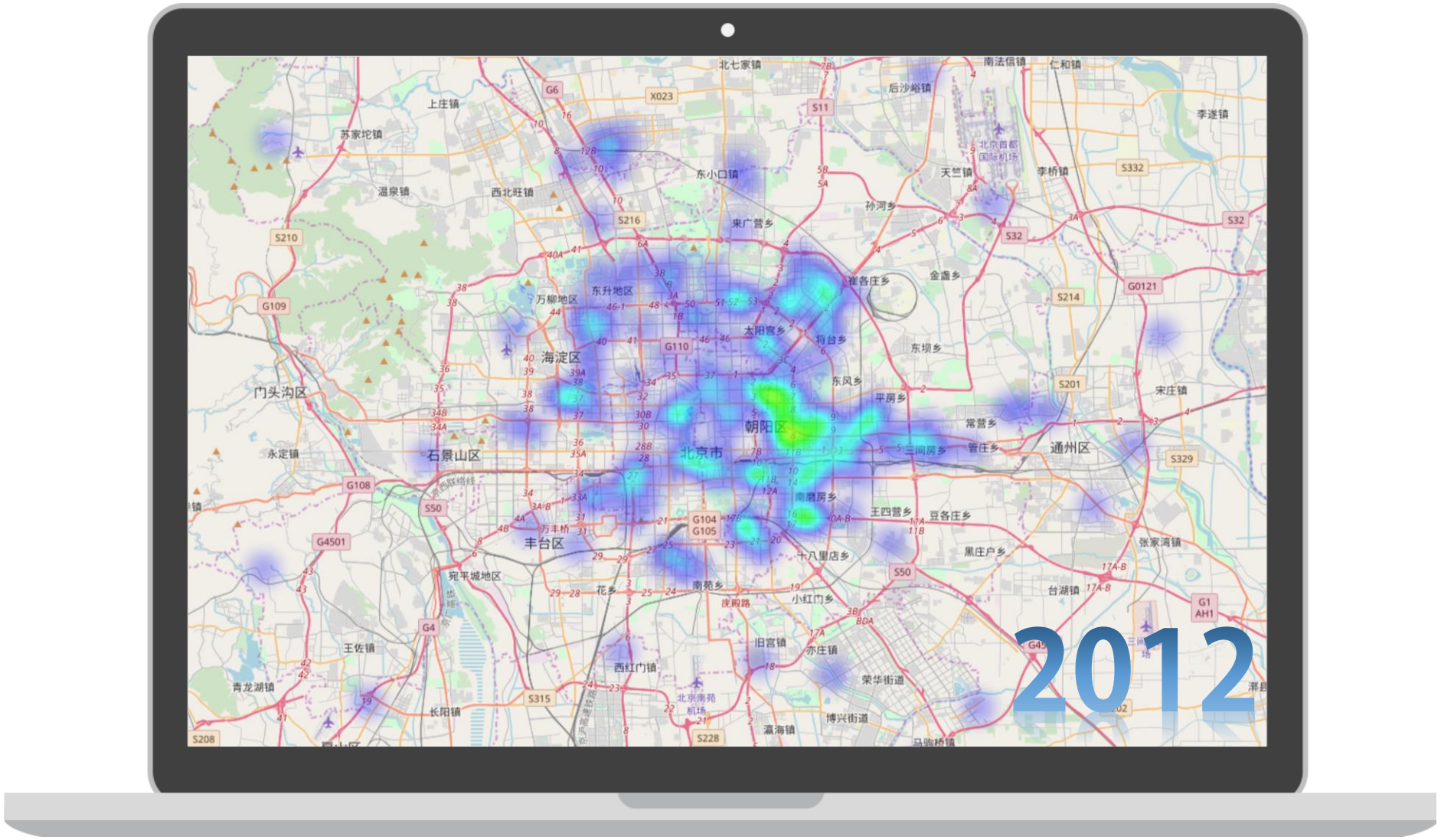
Methodology

Evaluation

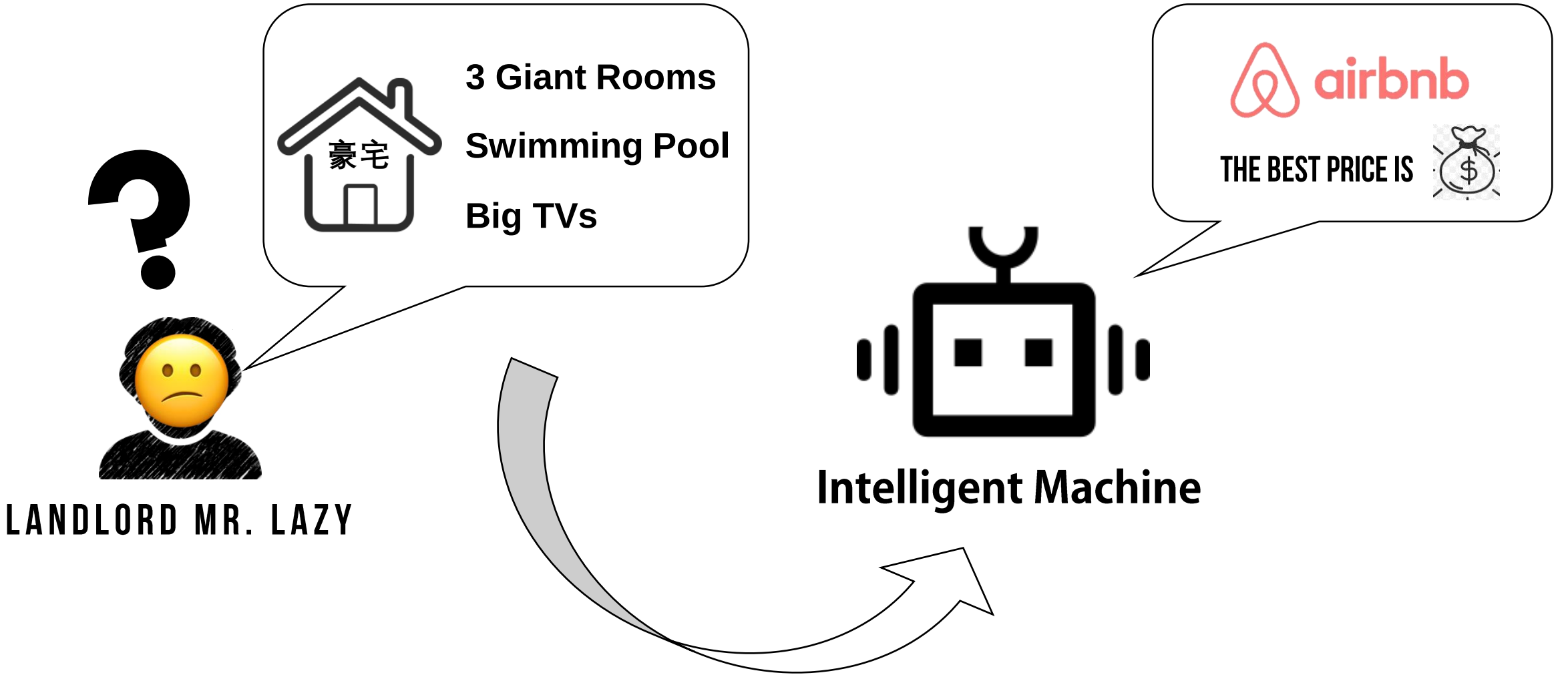
Motivation

Accumulative number of listings year by year since 2010 in Airbnb Beijing




2012
北京市
朝阳区
海淀区
丰台区
石景山区
通州区
门头沟区

2019

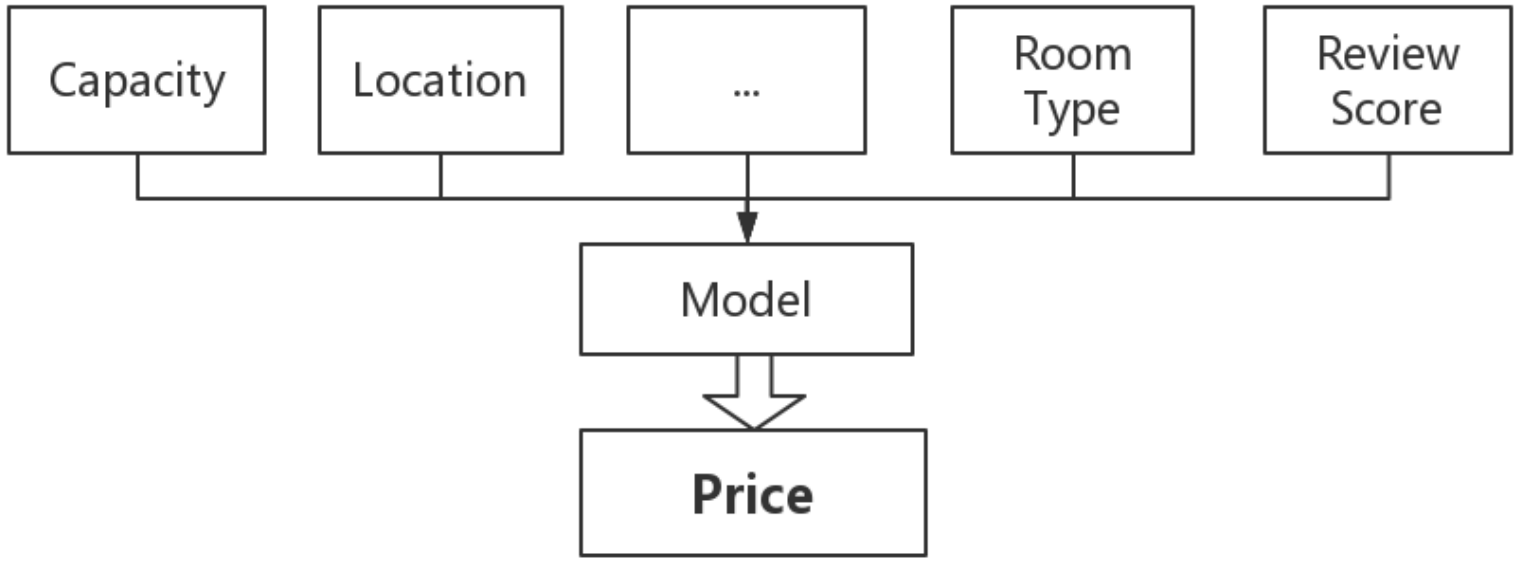

豪宅
3 Giant Rooms
Swimming Pool
Big TVs
LANDLORD MR. LAZY
airbnb
THE BEST PRICE IS
Intelligent Machine

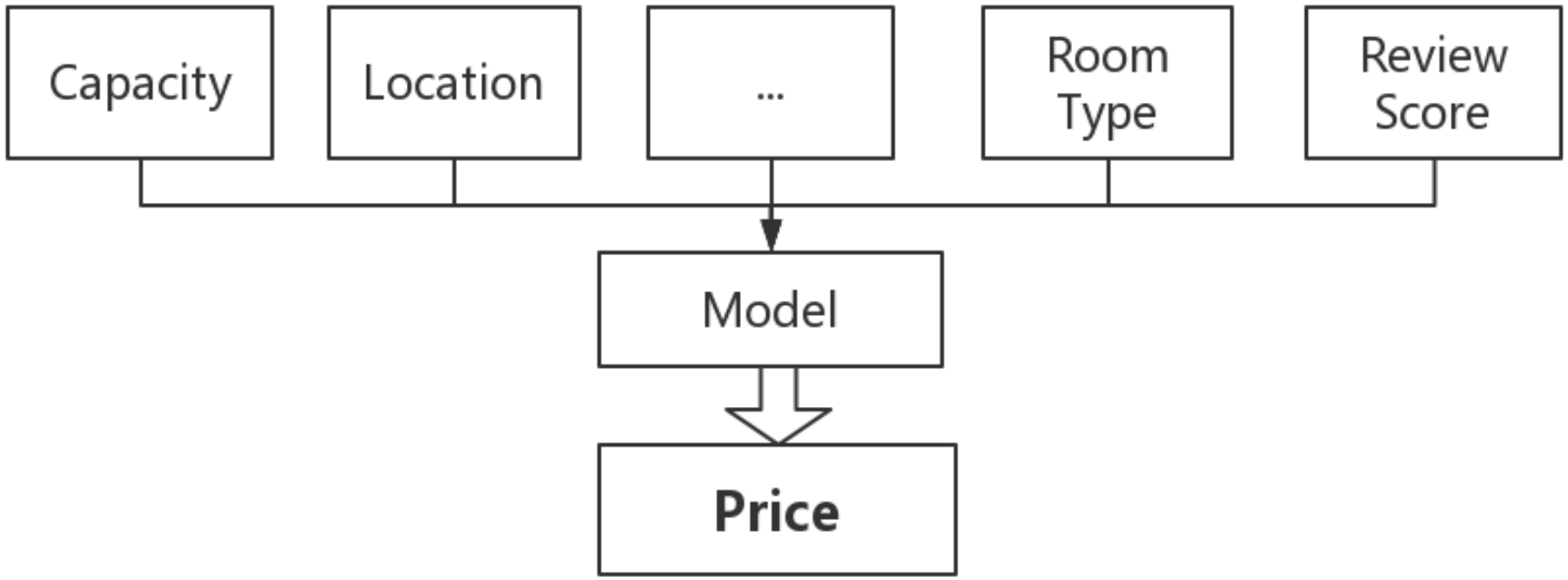
Problem Definition

Problem Statement

Given a series of data describes the property's features.

Output the reasonable/best price point for the host.



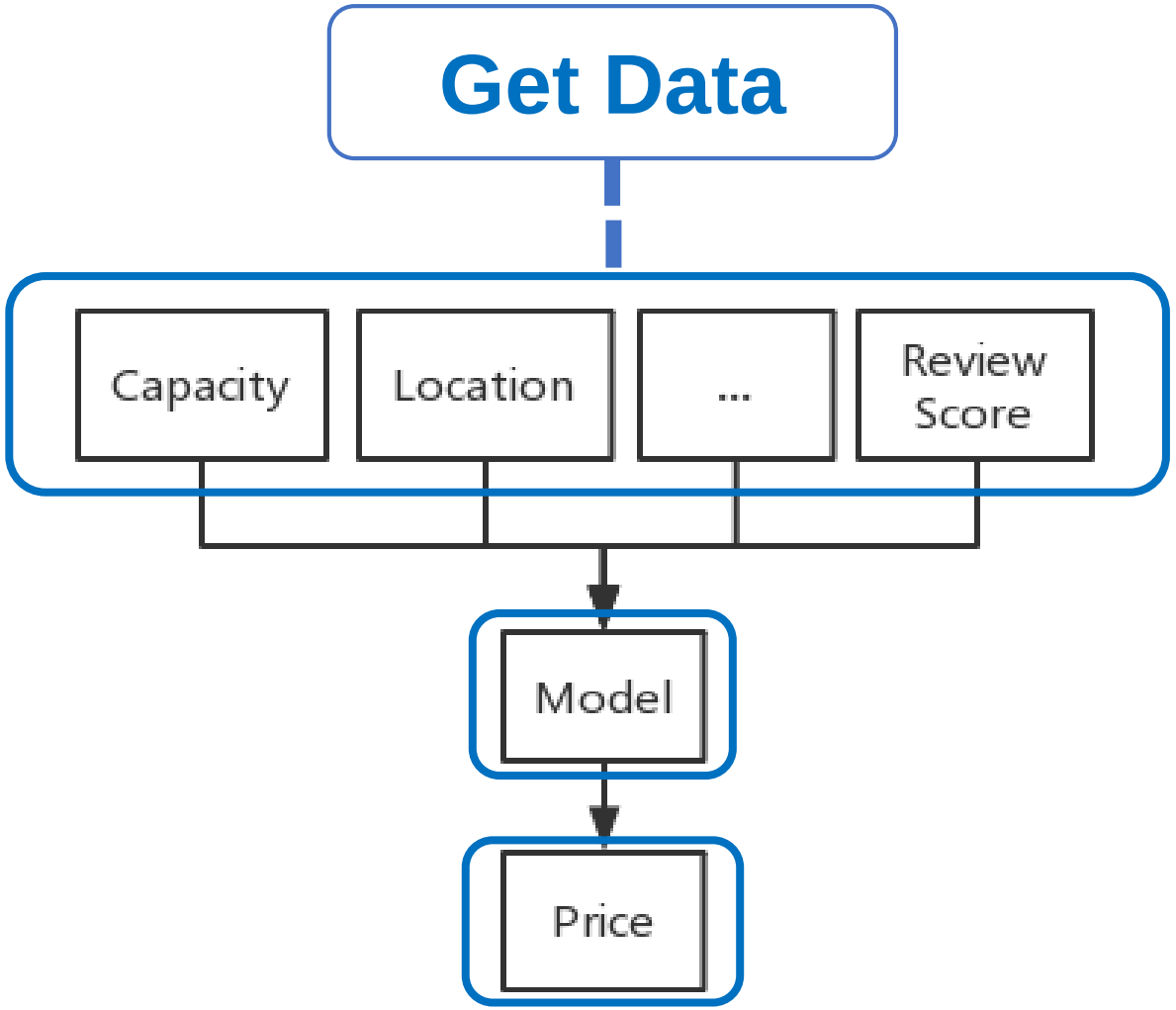


$$f(Capacity, Location, \dots, RoomType, ReviewScore) + \beta_0 \Rightarrow Price$$

State-of-the-Art

A brief review of the literature on price influencing factors

Dimension	Factors	Effects	Author
External	Price in cities	N/A	Stephen and Andrew(2016)
	Number of owners Distance between owners	Pos/Neg	Balaguer and Pernias(2013), Becerra(2013)
Location	Distance	Negative	Bull(1994), Wang & Nicolau(2017), Li (2016), Huang(2010)
House	Room type	Negative	Wang & Nicolau(2017), Zhang(2017)
	Room type	Positive	
Room	Bedroom, bathroom, bed	Positive	
Rule	Instant bookable	Negative	
	Cancellation policy	Positive	
Sociality	Number of reviews per year	Negative	Wang & Nicolau(2017), Zhang(2017)
		Positive	Zhu Lv(2017), Gutt & Herrmann(2015)
	Host identity verified	Pos/Neg	Wang & Nicolau(2017)
		Positive	Zhu Lv(2017)



Feature Engineering

⇩

Tuning Models

⇩

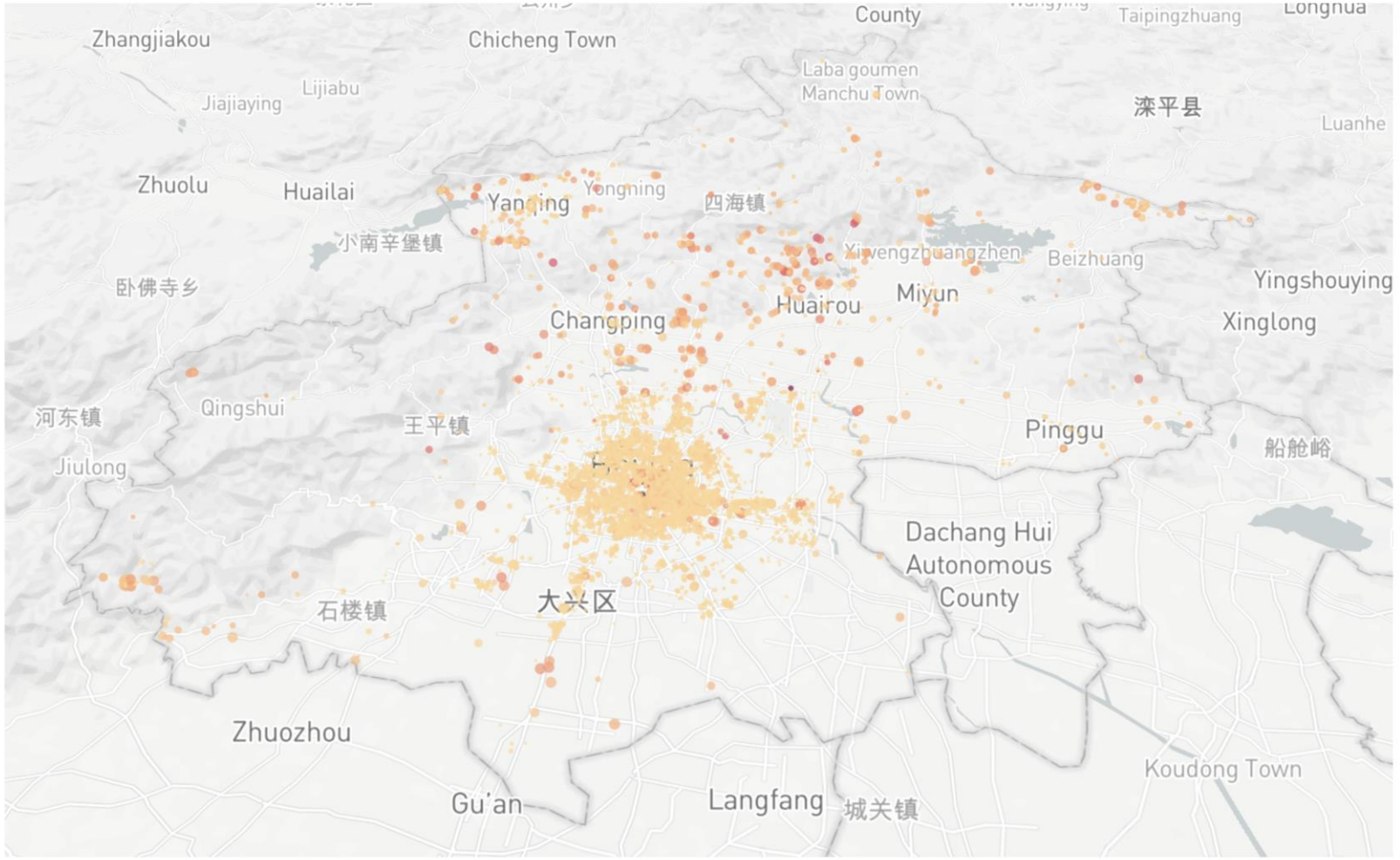
Outputs

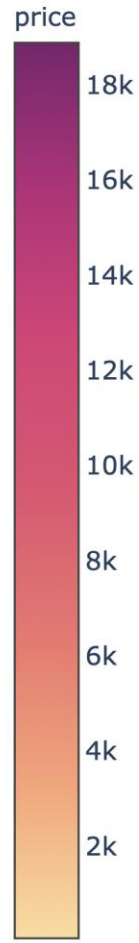
Methodology

Get Data

Raw_Data-Airbnb-Beijing.csv

Total:
19360 Listings





All available Data in Airbnb

Data Pre-Processing

- Drop Properties that lacks too many features (less than 100)
- Drop Properties with unreasonable prices (i.e. the host didn't actually want to Airbnb it)

~~[Price < (￥50 * accommodates) OR (￥50 * beds)]~~

~~[Price > ￥20,000 OR Price = ￥9999]~~

~~[Beds=50 OR Bathrooms=101 OR min_nights > 3]~~

- Fill up a few fixable nulls.

Airbnb-listings-Beijing.csv

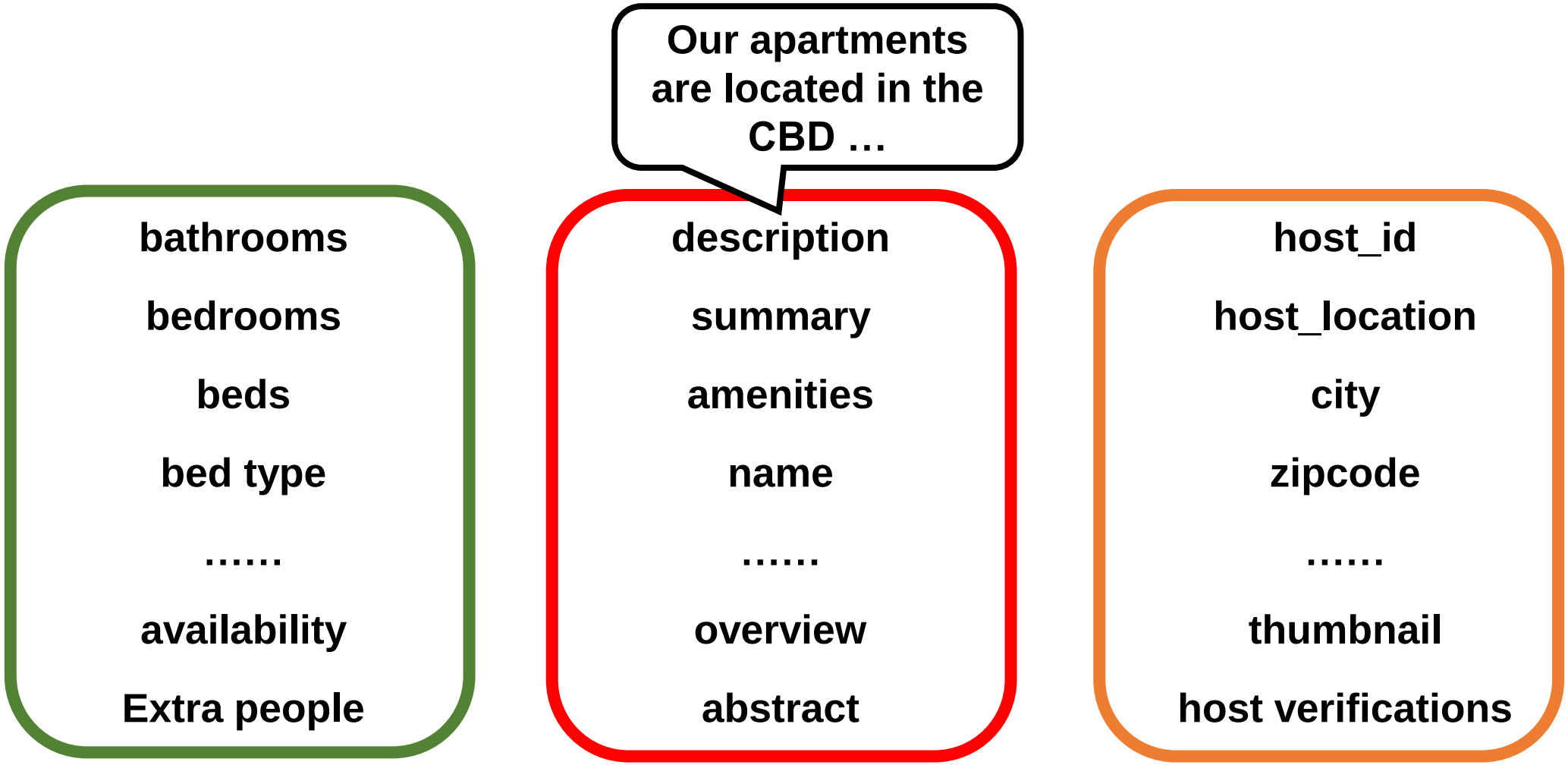
ID	Capacity	Room types	Price	…
44054	9	Entire home	816	…
100213	2	Entire home	1203	…
128496	3	Private room	401	…
161902	2	Private room	387	…
162144	4	Private room	553	…
279078	2	Entire home	401	…
282825	4	Entire home	657	…
287026	3	Entire home	415	…
287511	3	Entire home	415	…
317195	2	Private room	546	…
322292	2	Entire home	436	…
…	…	…	…	…

19360 Listings → **17224 Listings**

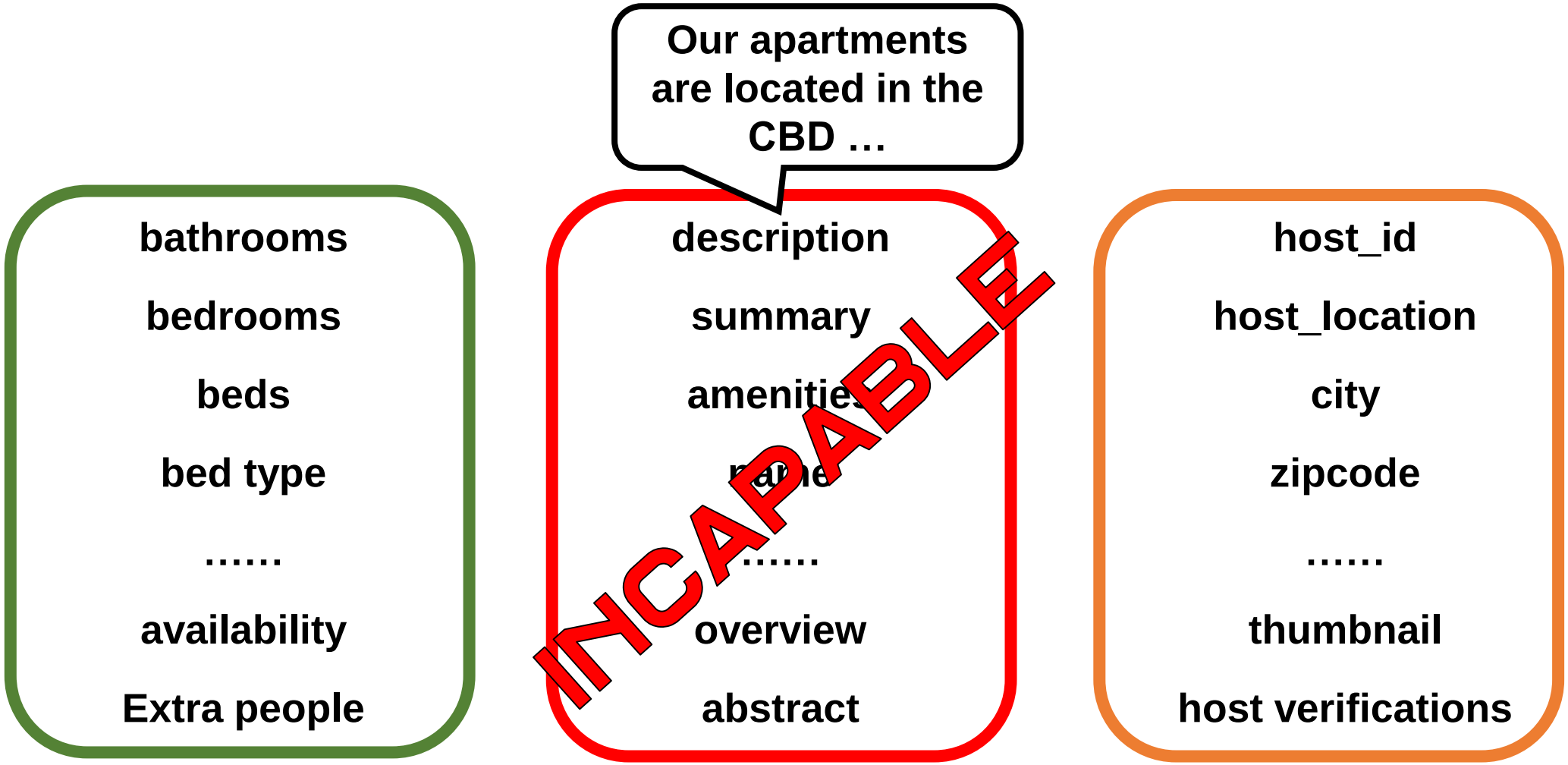
Feature Engineering

bathrooms	description	host_id
bedrooms	summary	host_location
beds	amenities	city
bed type	name	zipcode
……	……	……
availability	overview	thumbnail
Extra people	abstract	host verifications

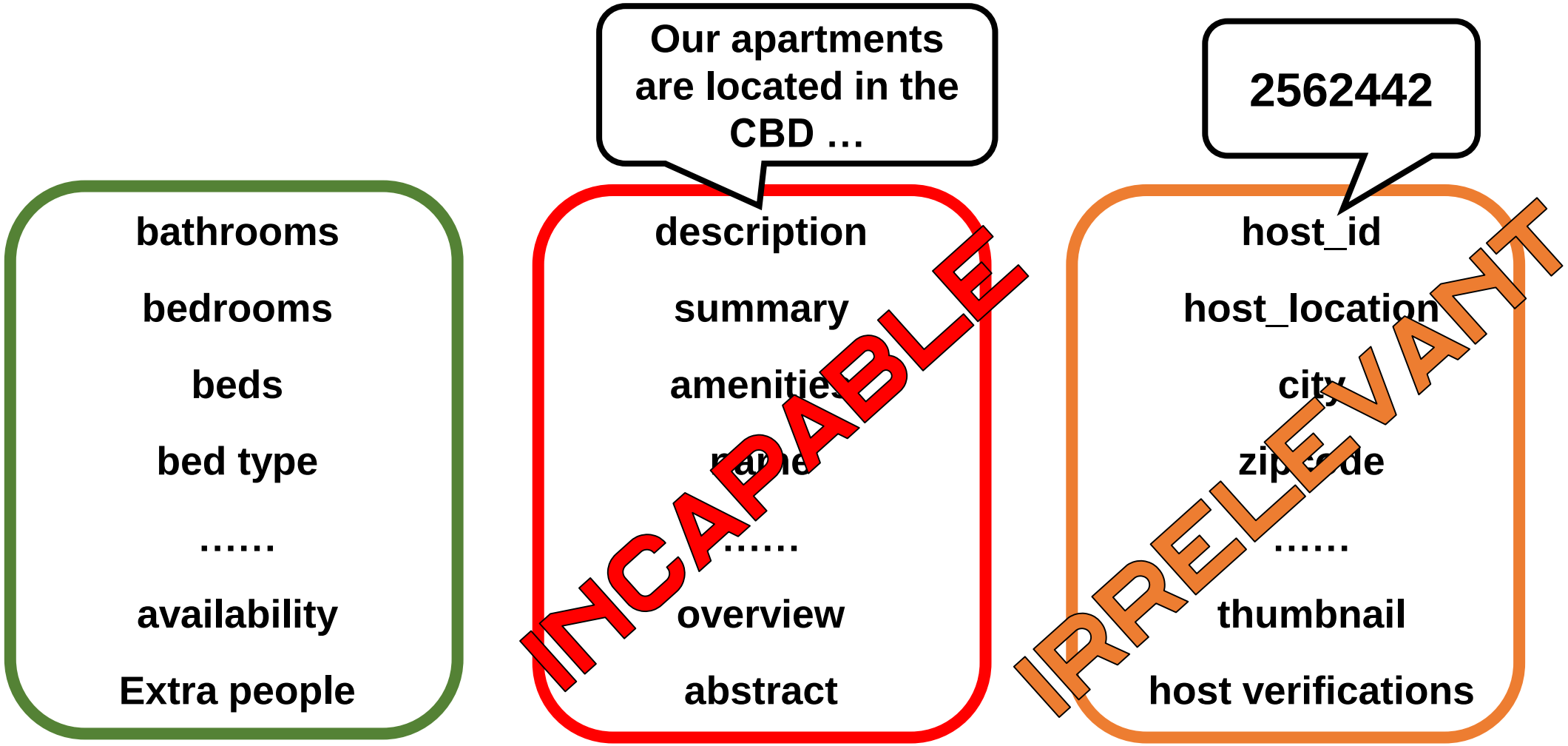
Feature Engineering



Feature Engineering



Feature Engineering



Feature Engineering

Feature Engineering

bathrooms

bedrooms

beds

bed type

……

availability

Extra people

+

Train Station

Airport

Wangjing SOHO

Yanqing IHE

……

Sanlitun

Subway

+

Feature
A / B / C

A	B	C
0	0	0
1	1	1

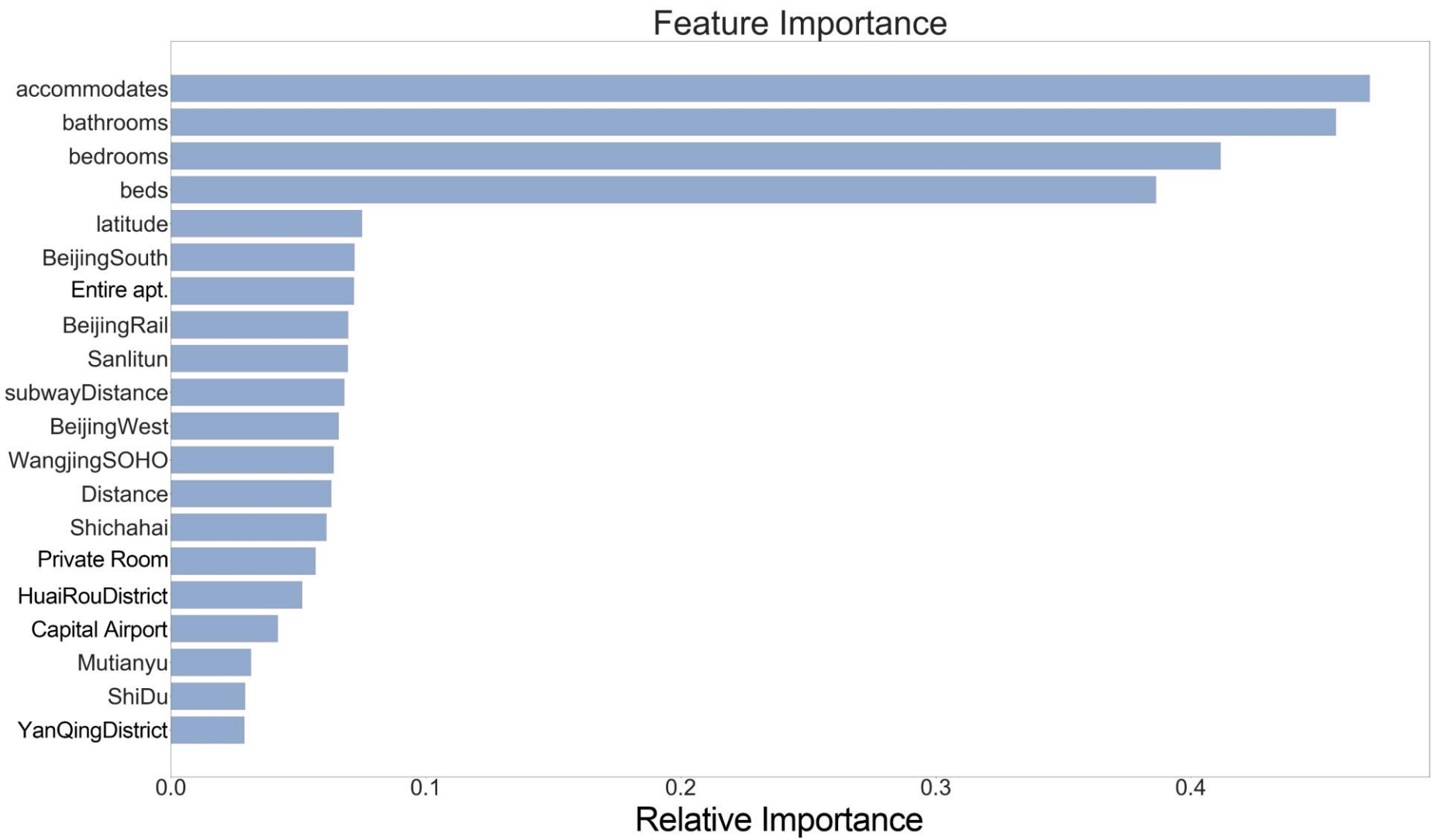
=

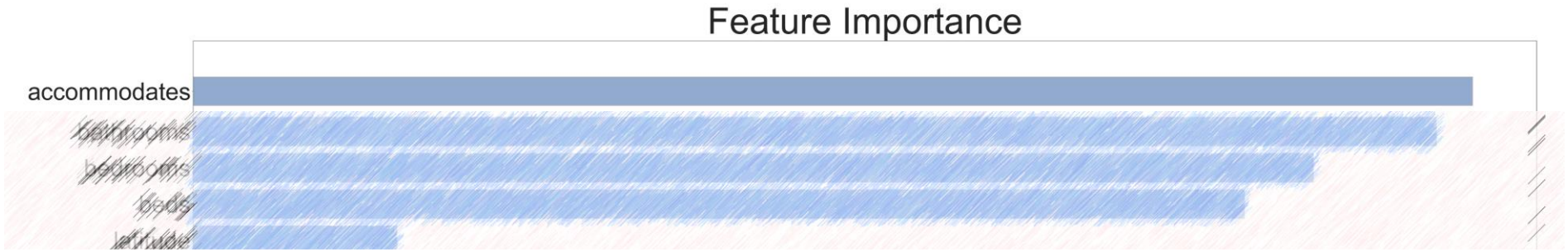
79

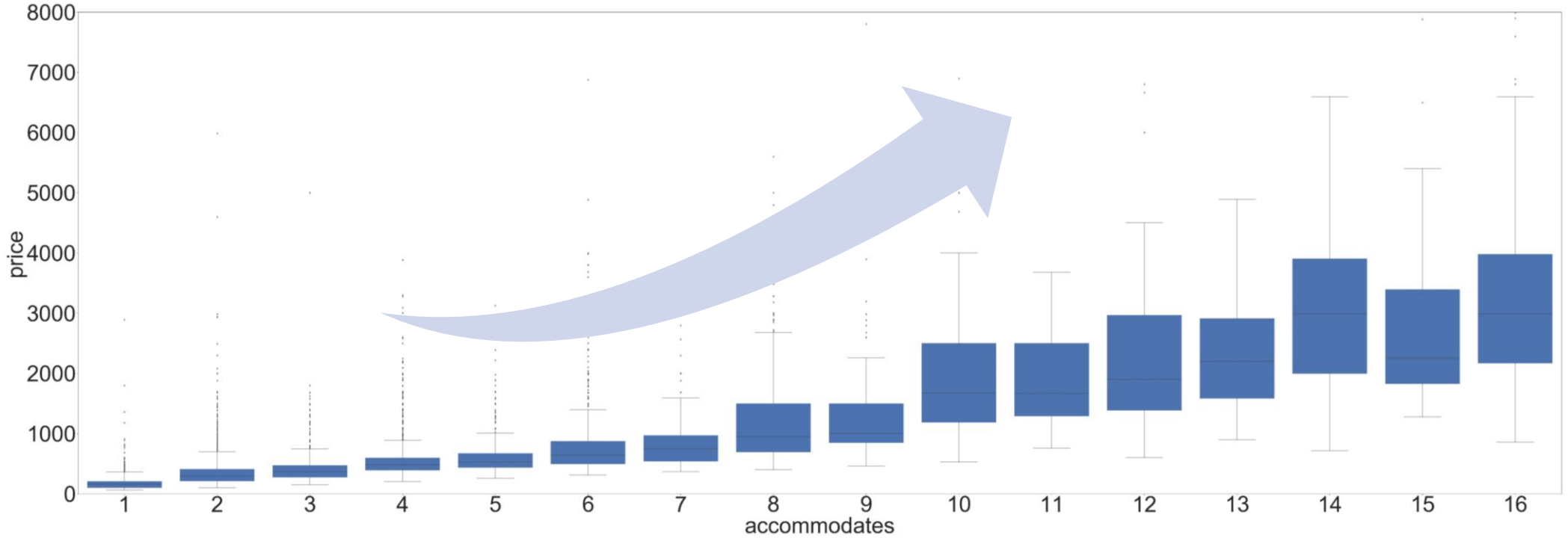
Features in total

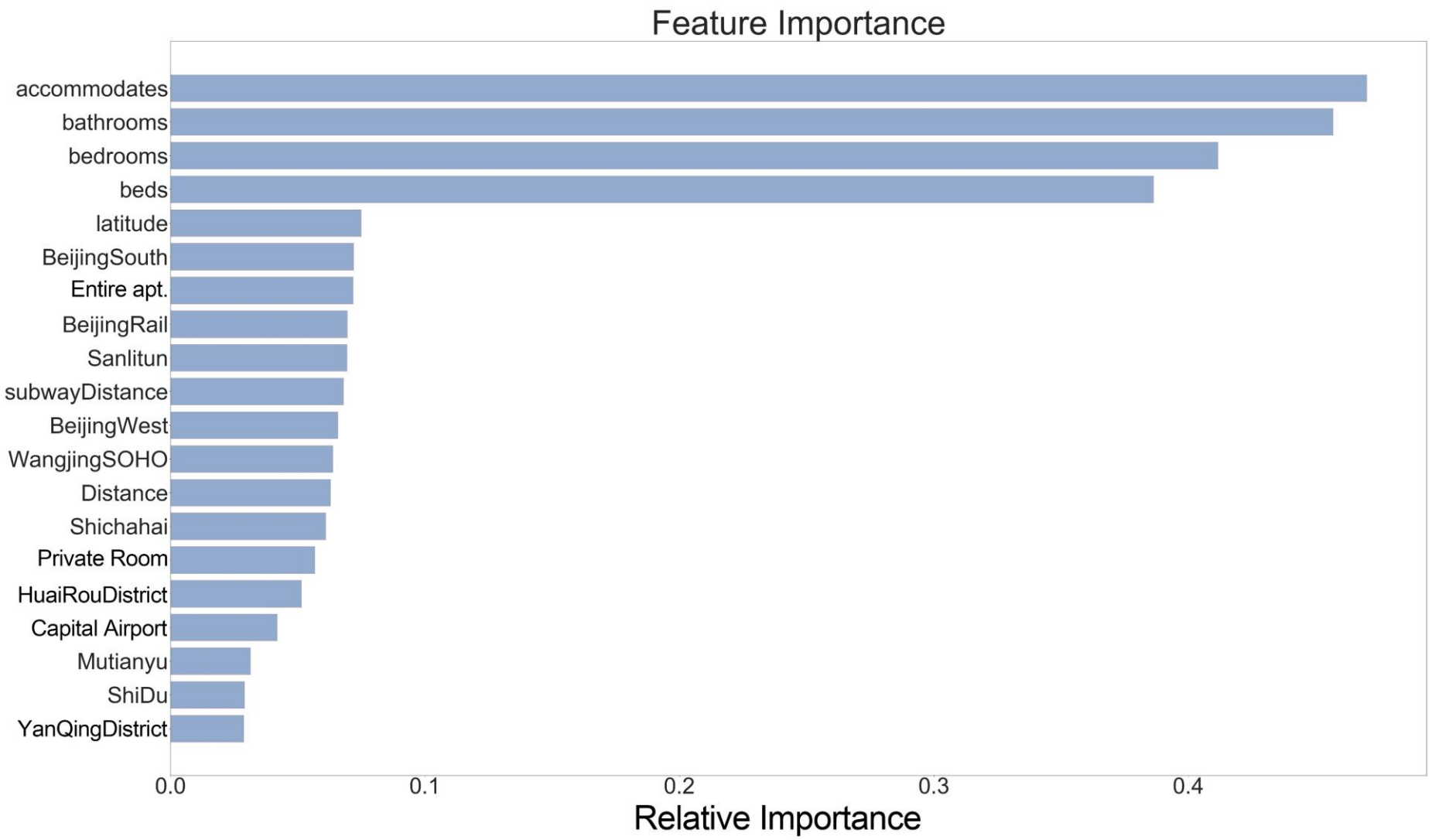
Adding Features

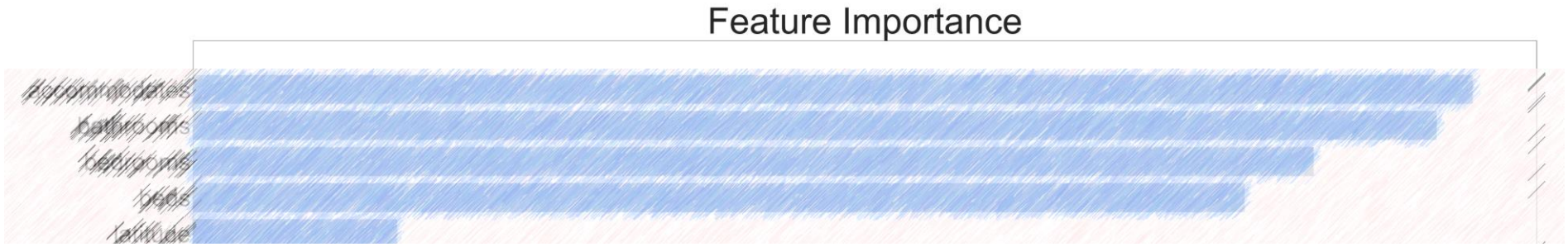
One-Hot Encoding

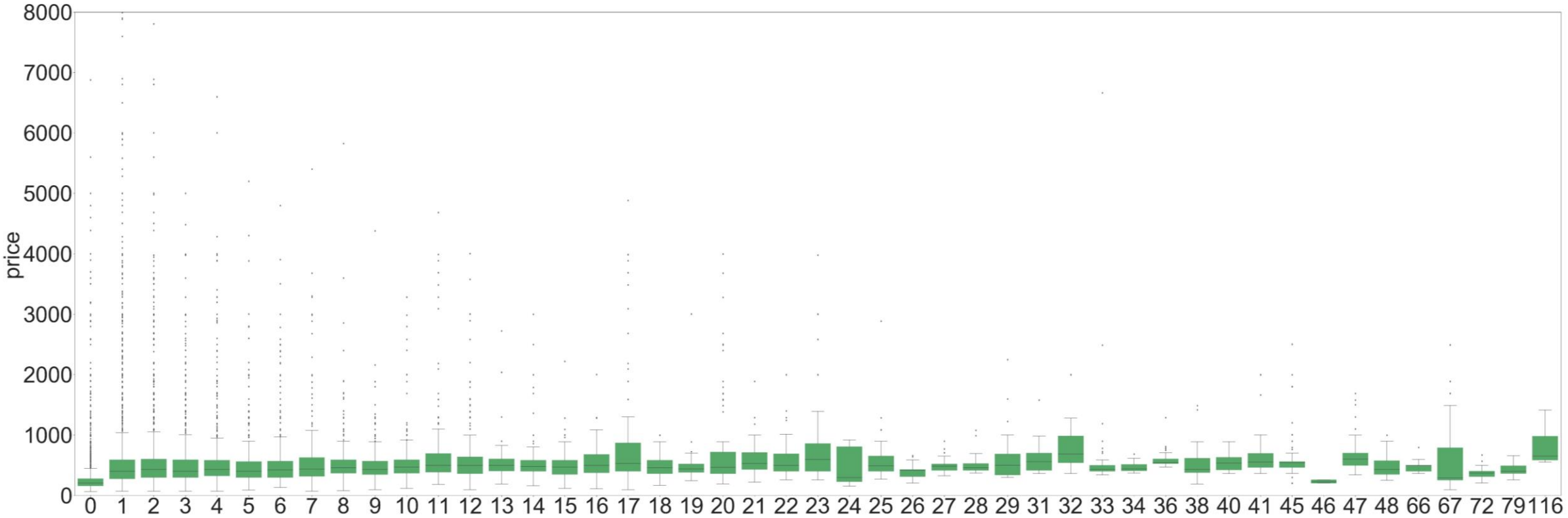

Feature Importance
accommodates
bathrooms
bedrooms
beds
latitude
BeijingSouth
Entire apt.
BeijingRail
Sanlitun
subwayDistance
BeijingWest
WangjingSOHO
Distance
Shichahai
Private Room
HuaiRouDistrict
Capital Airport
Mutianyu
ShiDu
YanQingDistrict
0.0
0.1
0.2
0.3
0.4
Relative Importance


Feature Importance
accommodates


8000
7000
6000
5000
4000
3000
2000
1000
0
price
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
accommodates


Feature Importance
accommodates
bathrooms
bedrooms
beds
latitude
BeijingSouth
Entire apt.
BeijingRail
Sanlitun
subwayDistance
BeijingWest
WangjingSOHO
Distance
Shichahai
Private Room
HuaiRouDistrict
Capital Airport
Mutianyu
ShiDu
YanQingDistrict
0.0
0.1
0.2
0.3
0.4
Relative Importance

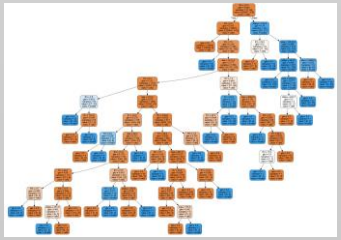


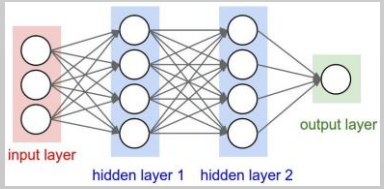


No. listings of host-Price

Model Building

Model





Splitted Data

Train Dataset
80 %

Test Data set
20%

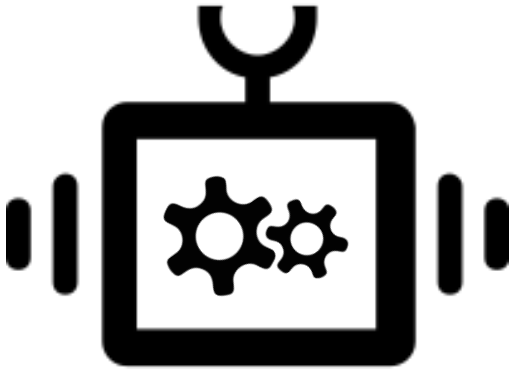
Grid Search
Cross Validation

PIPELINE

Scores

Visuals

Evaluation

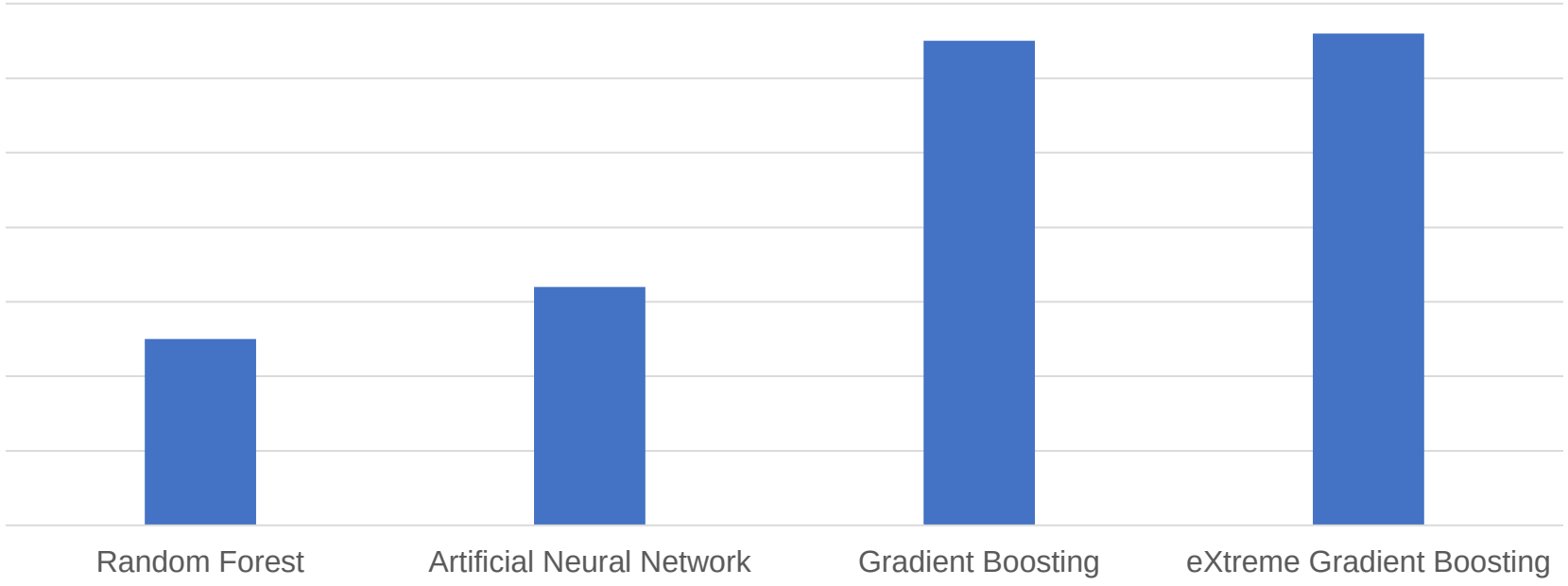


Model

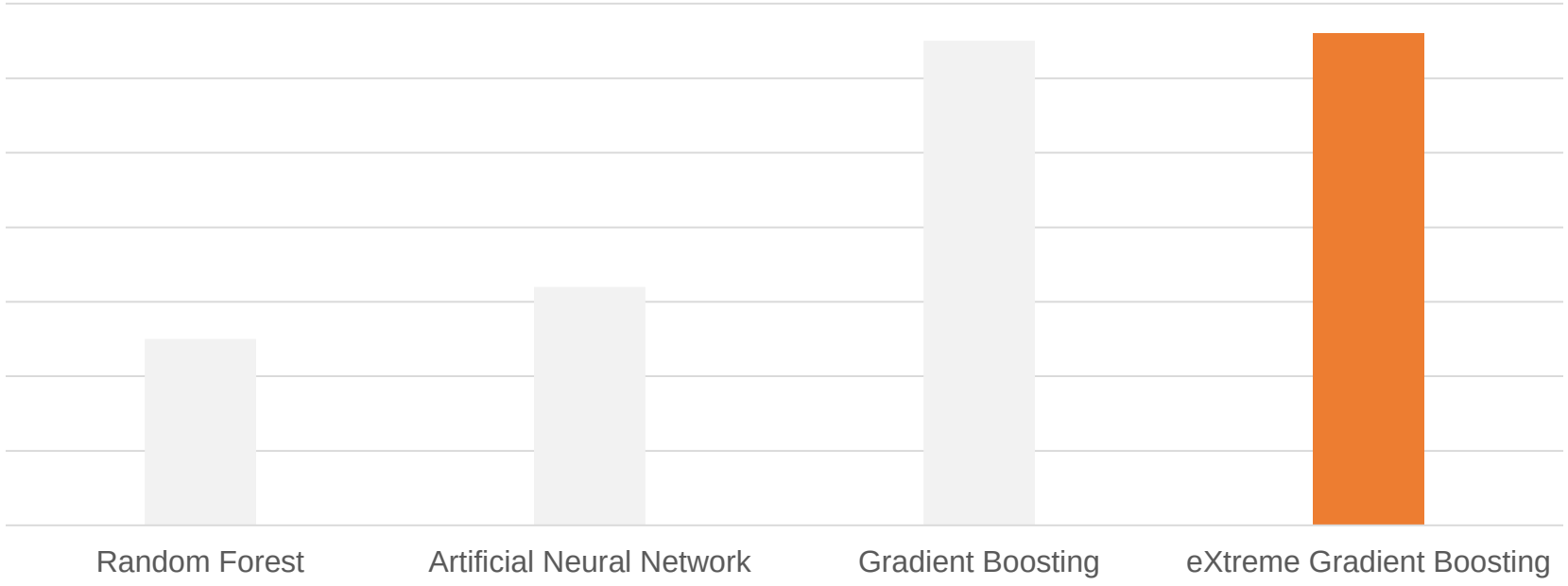
Train Dataset
80 %

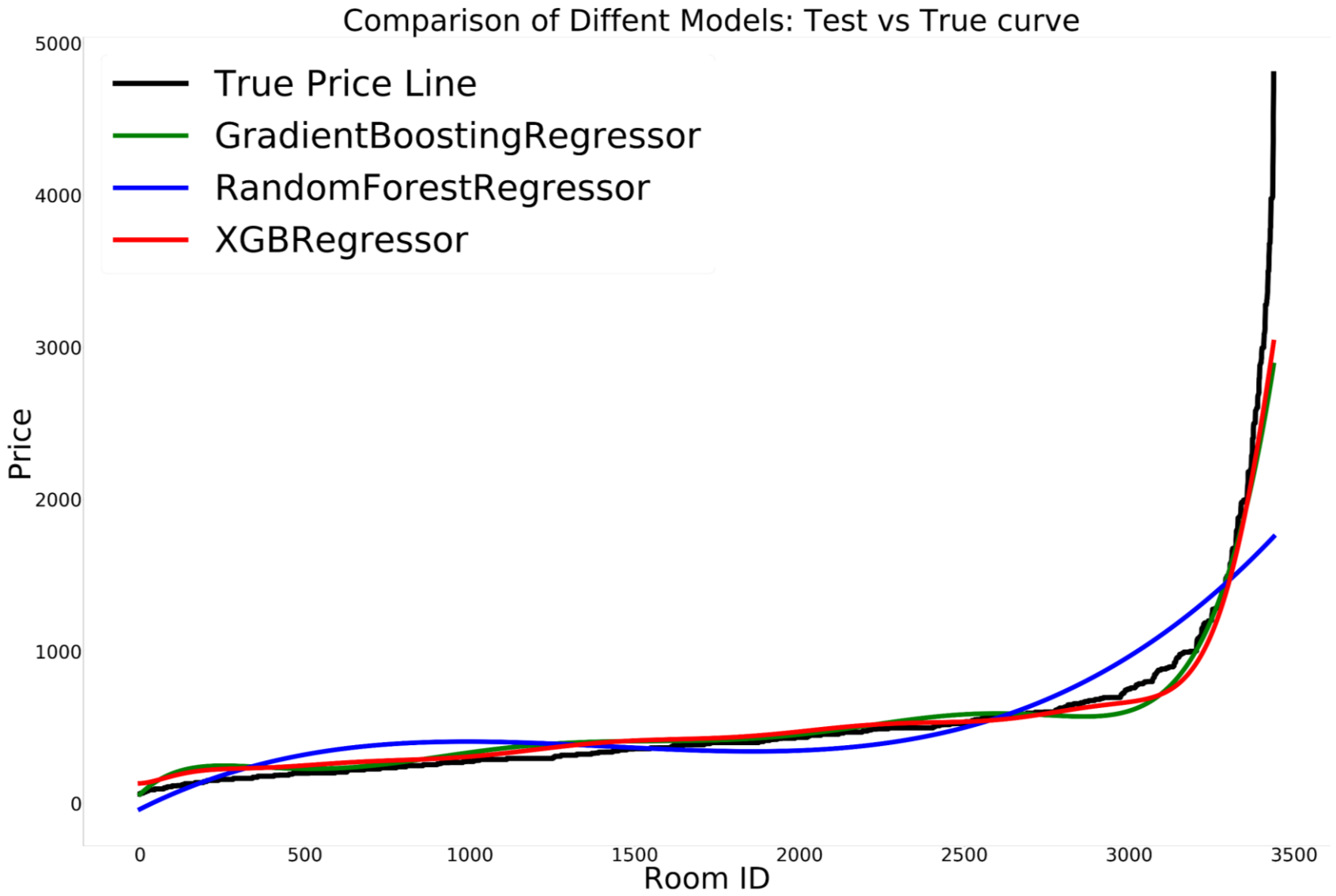
Test Dataset
20%

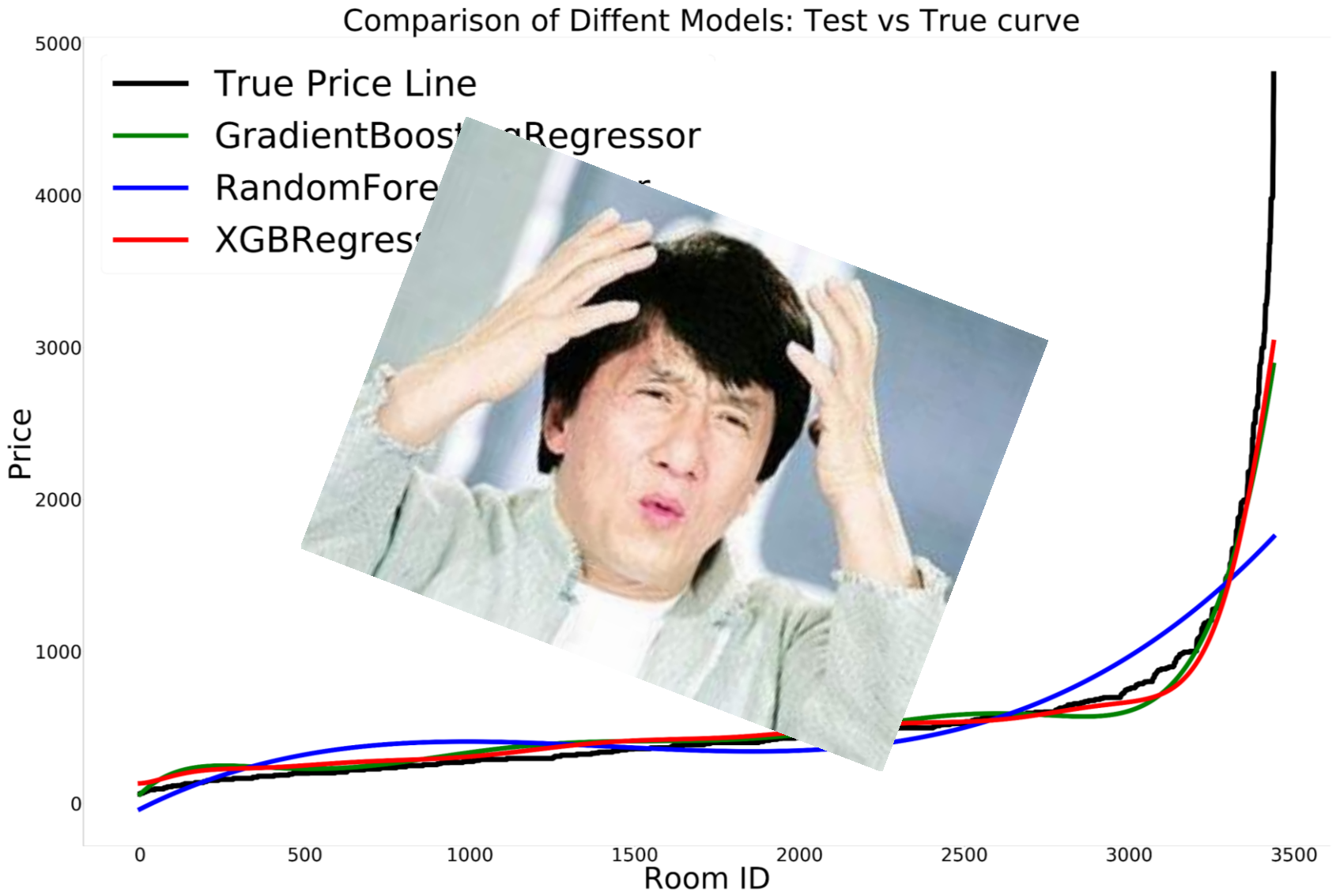
Model Name	R^2 Score
Random Forest	0.705
Artificial Neural Network	0.712
Gradient Boosting	0.745
eXtreme Gradient Boosting	0.746



Model Name	R^2 Score
Random Forest	0.705
Artificial Neural Network	0.712
Gradient Boosting	0.745
eXtreme Gradient Boosting	**0.746**




Comparison of Diffent Models: Test vs True curve
True Price Line
GradientBoostingRegressor
RandomForestRegressor
XGBRegressor
Price
Room ID
0
1000
2000
3000
4000
5000
0
500
1000
1500
2000
2500
3000
3500


Comparison of Diffent Models: Test vs True curve
True Price Line
GradientBoost gRegressor
RandomFore r
XGBRegres
Price
Room ID
0
1000
2000
3000
4000
5000
0
500
1000
1500
2000
2500
3000
3500

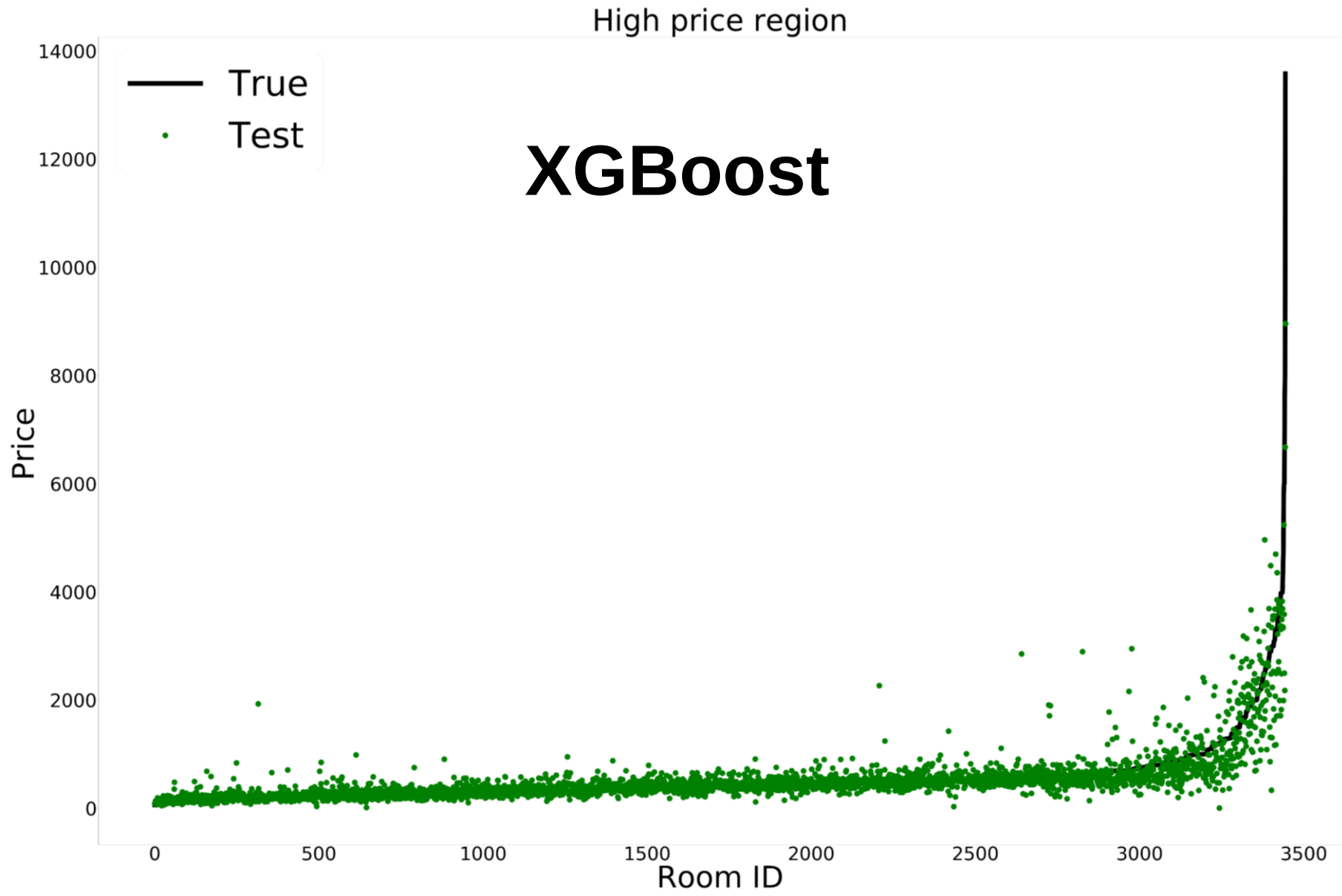
Model Name	R^2 Score	Prediction
Random Forest	0.705	¥ 1798.06
Artificial Neural Network	0.712	¥ 1820.95
Gradient Boosting	0.745	¥ 1882.24
eXtreme Gradient Boosting	**0.746**	**¥ 1887.61**

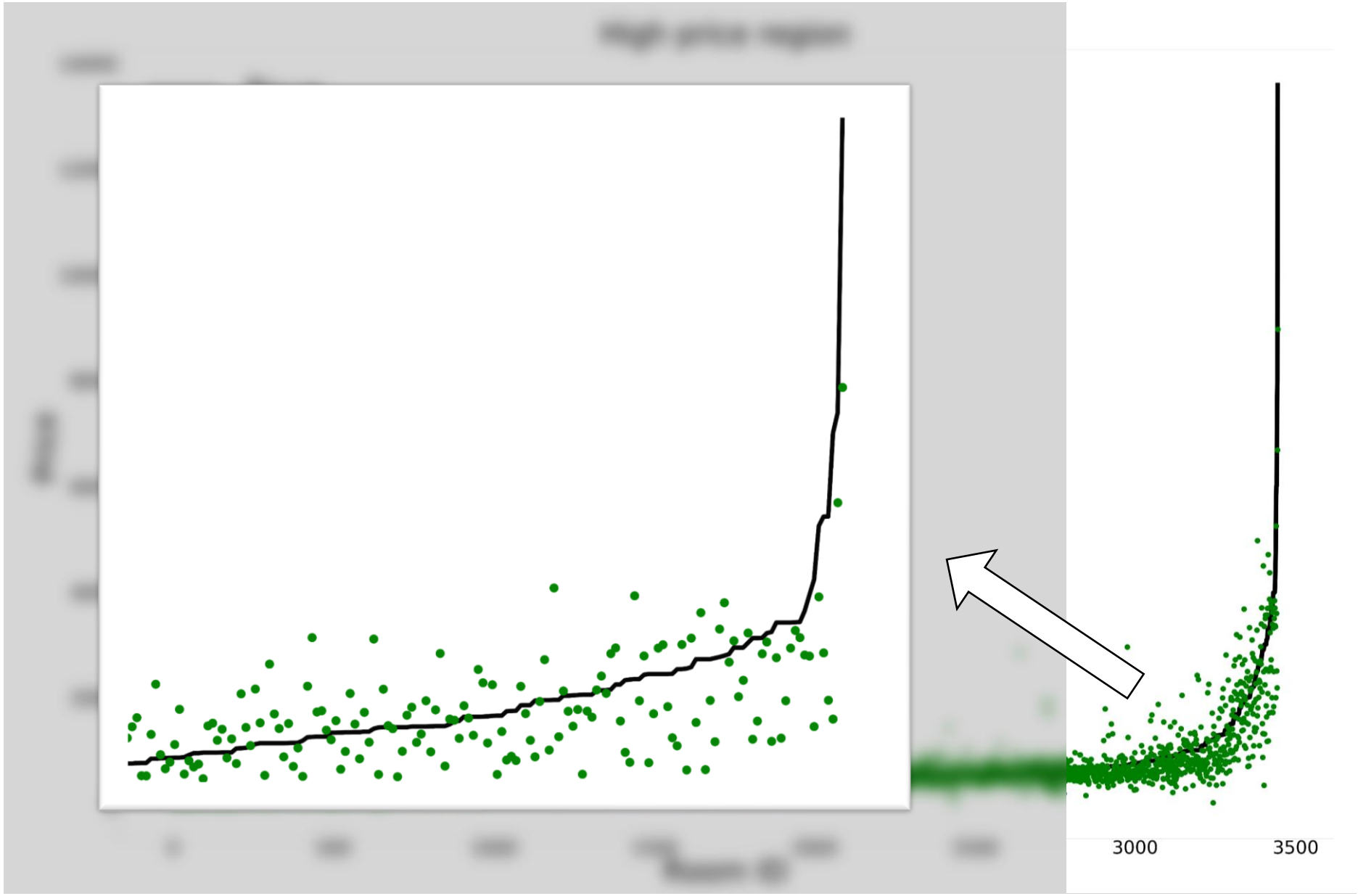
Beijing

Entire house

12 guests 6 bedrooms 6 beds 2 baths

¥ 1,888 per night


High price region
True
Test
XGBoost
Price
Room ID
14000
12000
10000
8000
6000
4000
2000
0
0
500
1000
1500
2000
2500
3000
3500


3000
3500

Error Analysis

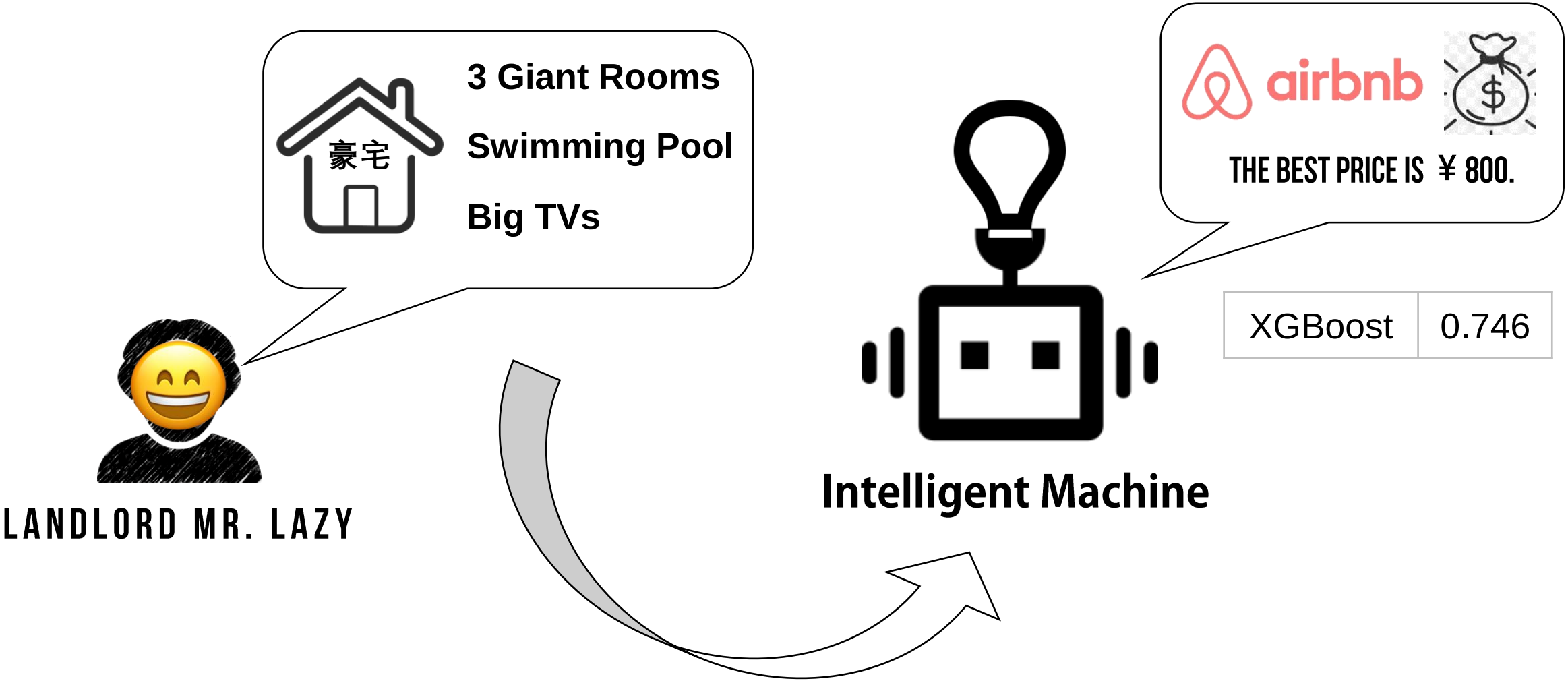
- Expensive listings have **less** data compared to inexpensive listings

- **Feature isn't everything for price.**

¥ 200

¥ 2348

To wrap up,



XGBoost	0.746

Thank you!

AI HUB PBL

Thank you!
AI HUB 😘

Q & A