DTCC

2017第八届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2017

链家网数据挖掘技术实践

——估价系统的前世今生

演讲者：宋鑫

链家网数据策略部

关于我

硕士
计算机科学与技术专业

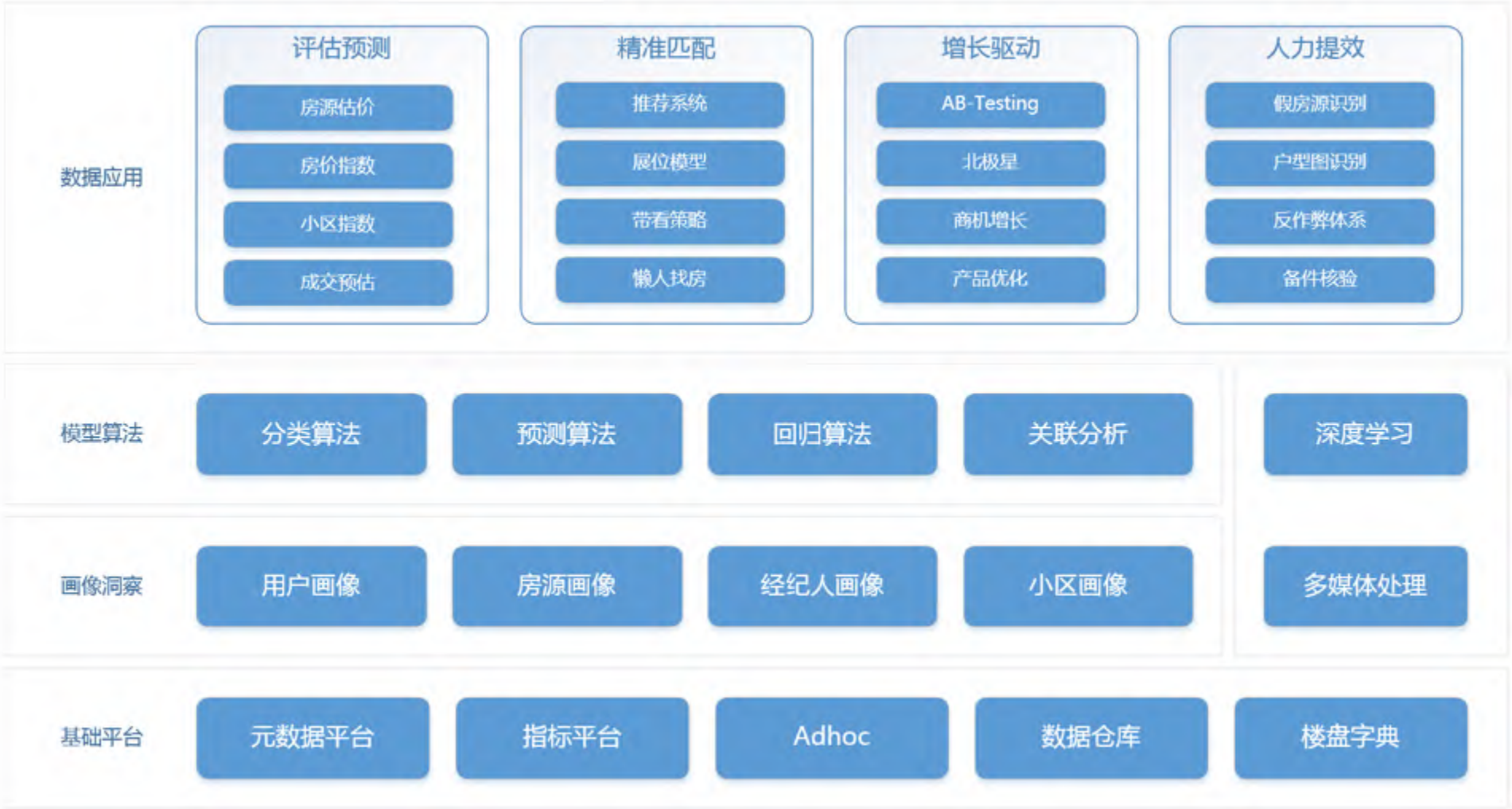
Baidu 百度

高级数据挖掘工程师
凤巢广告自动审核系统

LIANJIA 链家 .com

资深数据挖掘工程师
房屋估价，房源成交预估

链家网数据挖掘体系结构



内容概要

- 为什么要做估价
- 估价系统现状
- 估价系统总体设计
- 估价系统难点及解决方案
- 总结

内容概要

- 为什么要做估价
- 估价系统现状
- 估价系统总体设计
- 估价系统难点及解决方案
- 总结

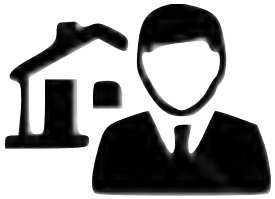
为什么要做估价

链家网：链接人和房产服务

为什么要做估价

二手房交易市场是大宗低频交易，需要一个高频的场景来找到并维系潜在业主与潜在客户

业主、买家需要一个双方认可的价值锚点



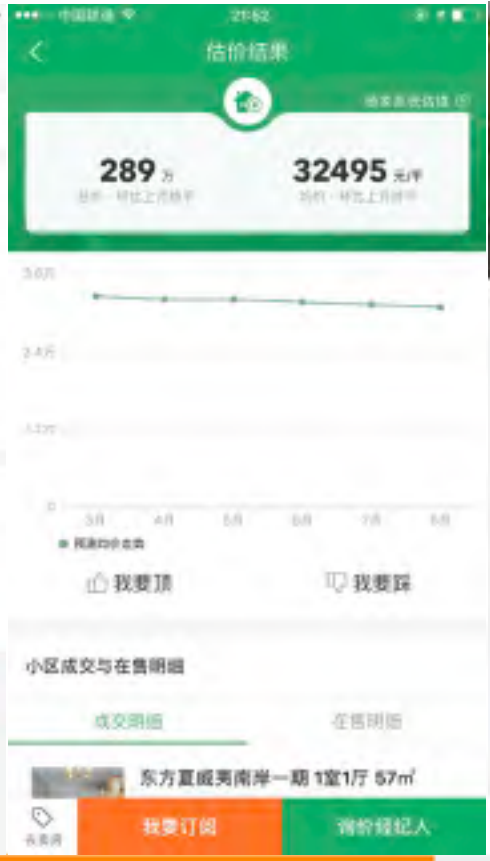
经纪人作业需要估价工具的指导和规范

内容概要

- 为什么要做估价
- 估价系统现状
- 估价系统总体设计
- 估价系统难点及解决方案
- 总结

估价系统现状

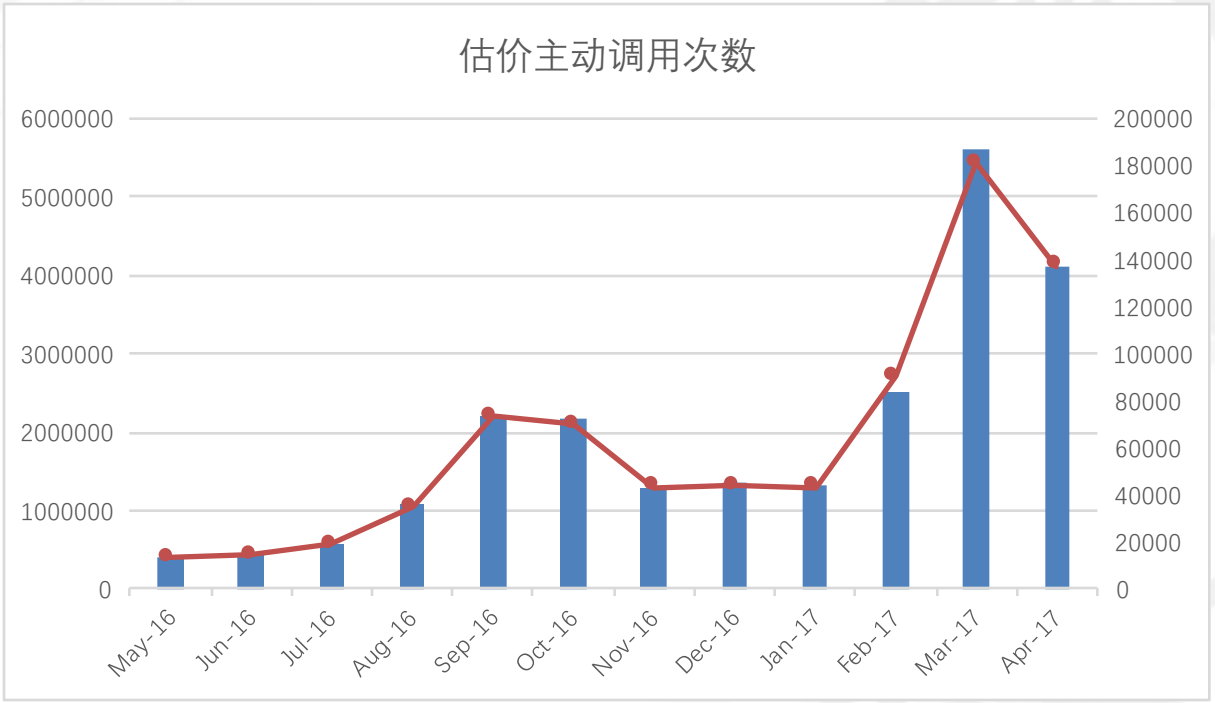
我是房主 → 我要估价



估价系统现状 — 覆盖率和调用次数

北京 燕郊
深圳 南京

小区覆盖率90%



估价系统现状 — 准确率

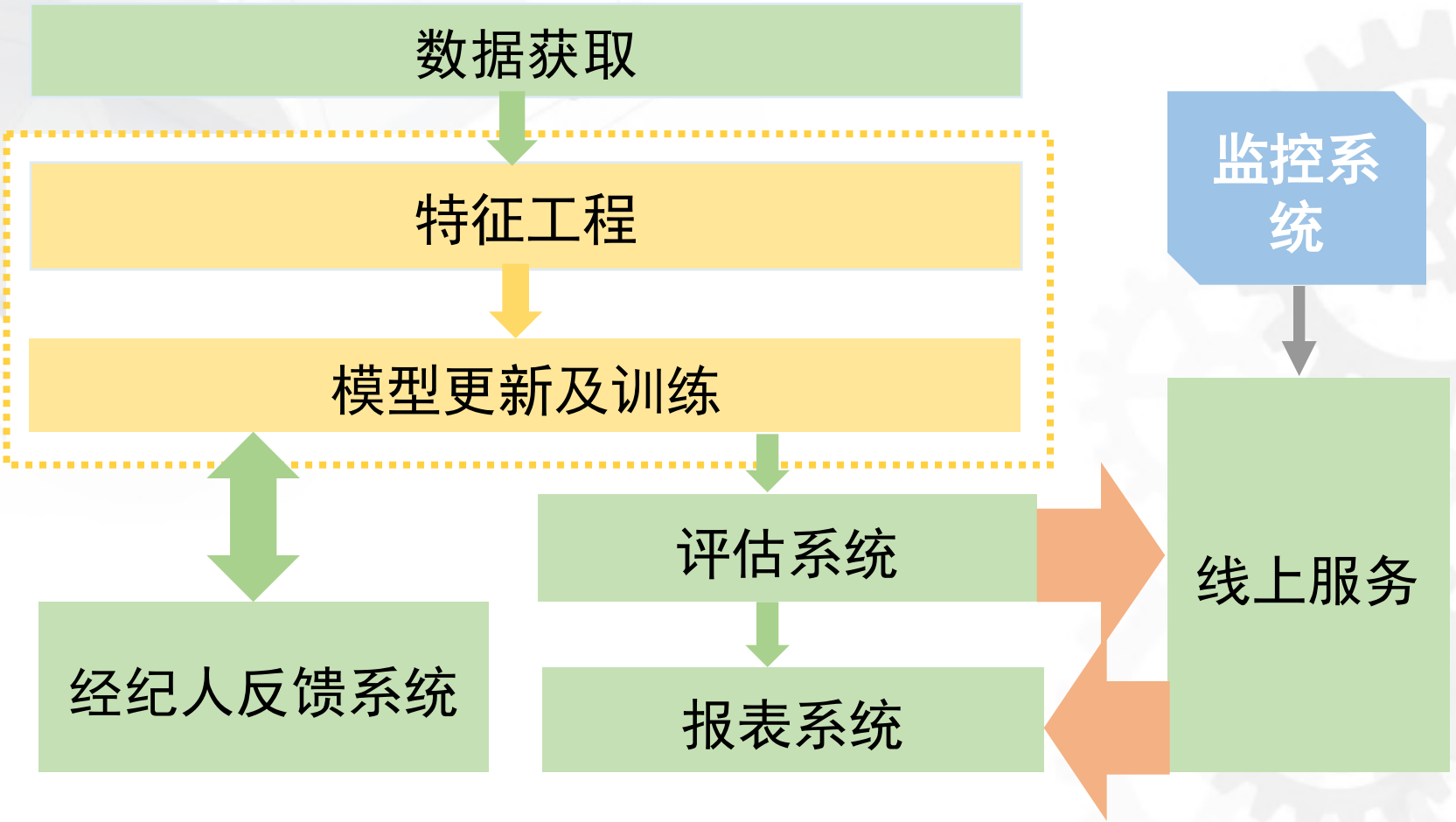
65%

81%

内容概要

- 为什么要做估价
- 估价系统现状
- 估价系统总体设计
- 估价系统难点及解决方案
- 总结

估价系统总体设计



估价系统总体设计 — 特征设计

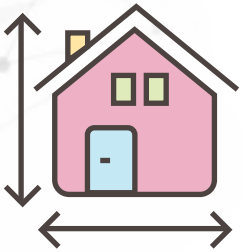
数据和特征决定了机器学习的上限

而模型和算法只是逼近这个上限而已

估价系统总体设计 — 特征设计

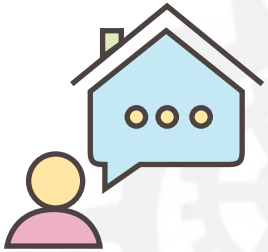
交易特征

- 小区成交均价
- 小区挂牌均价
- 小区分居室成交均价
- 小区分居室挂牌均价
- ……



物理特征

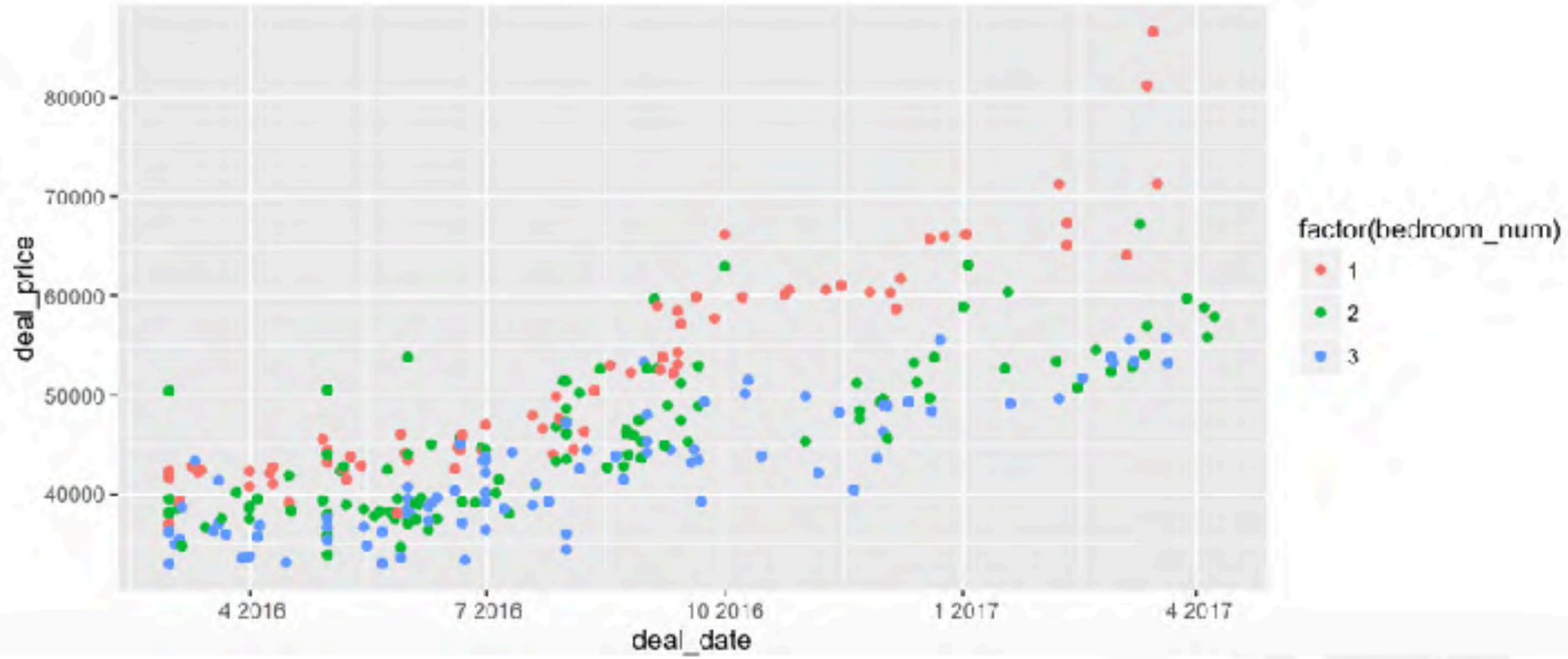
- 几室几厅几卫
- 面积
- 楼层打分/楼层/是否底层/是否顶层
- 建筑类型: 塔楼/板楼/塔板结合
- 建筑年限
- 户型朝向
- 是否精装修
- ……



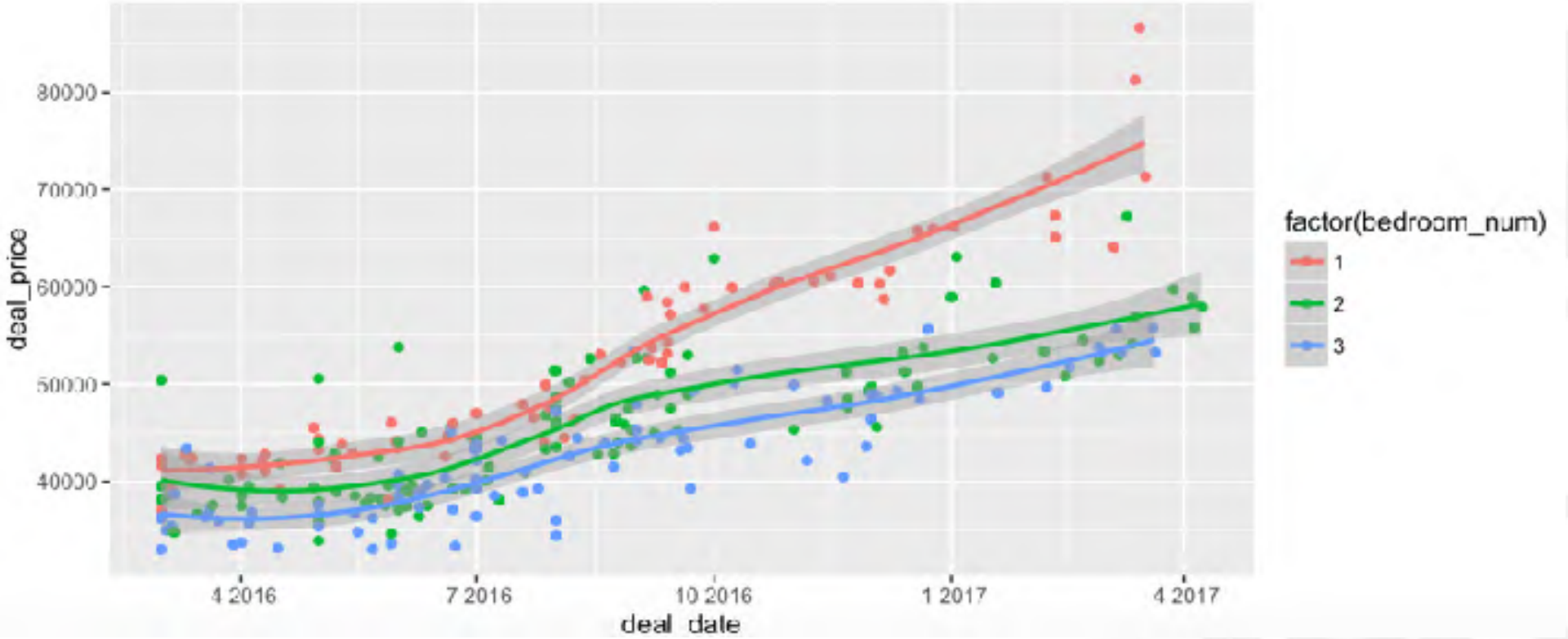
市场供需特征

- 小区平均成交周期
- 小区供需比
- 商圈供需比
- 小区挂牌量
- 小区潜客数
- ….

估价系统总体设计 — 特征设计



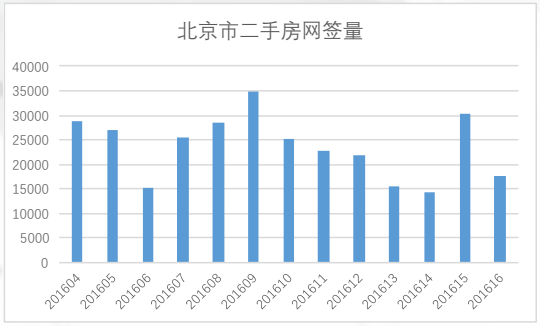
估价系统总体设计 — 特征设计

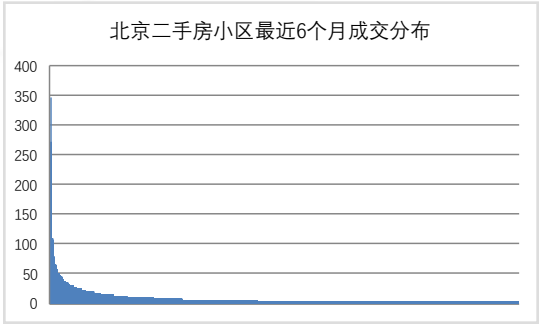


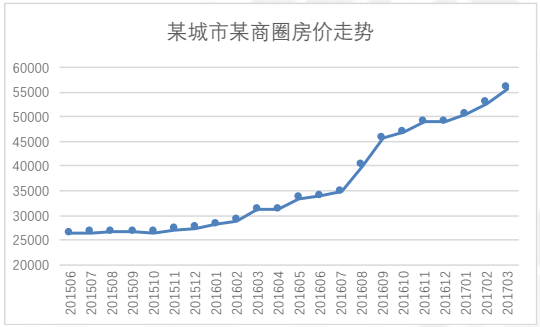
内容概要

- 链家网数据挖掘体系结构
- 为什么要做估价
- 估价系统现状
- 估价系统难点及解决方案
- 总结

估价系统难点及挑战— 交易数据的稀缺、稀疏和时变







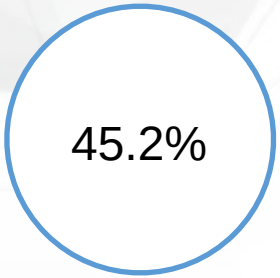
数据总量相对较少

数据分布严重不均

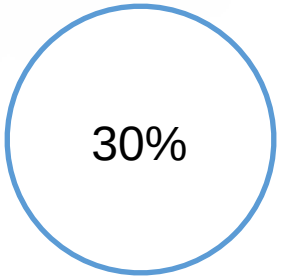
数据时变性强

估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

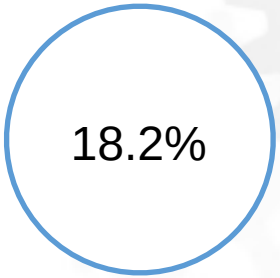
北京市1.2W小区，只有5900个小区最近6个月有成交



近期无成交



近期无挂牌

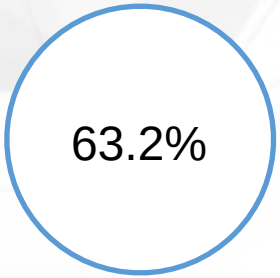


近期既无挂牌也无成交

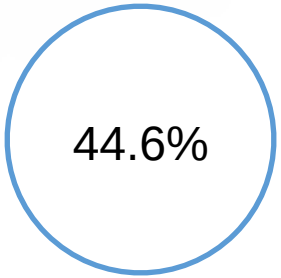
统计时间点: 20170410

估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

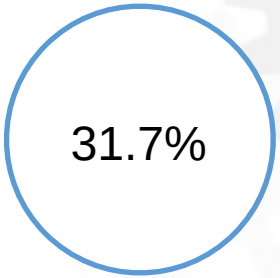
具体到小区居室的数据缺失



近期无成交



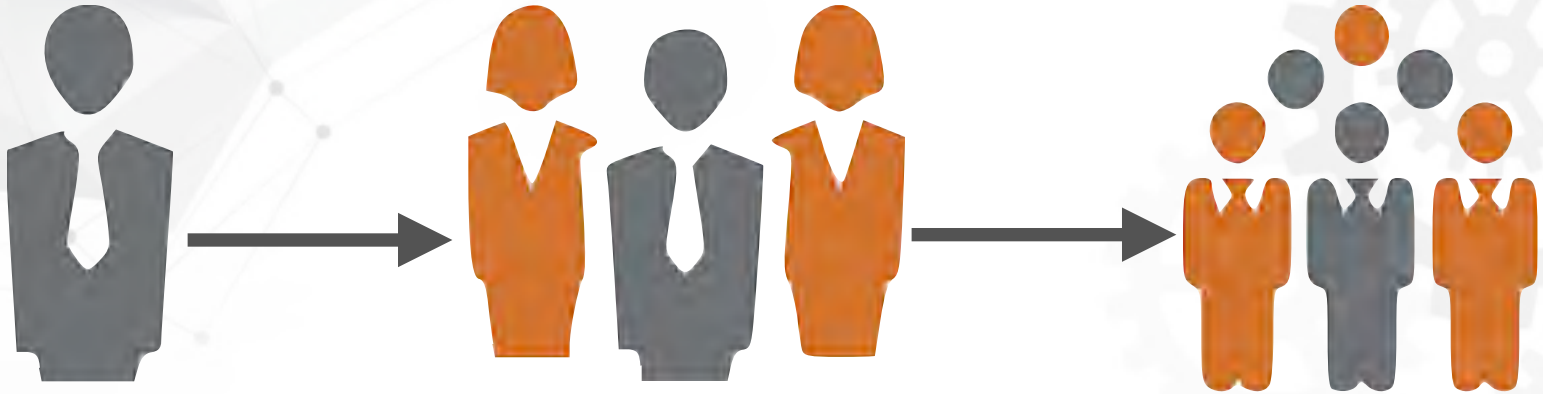
近期无挂牌



近期既无挂牌也无成交

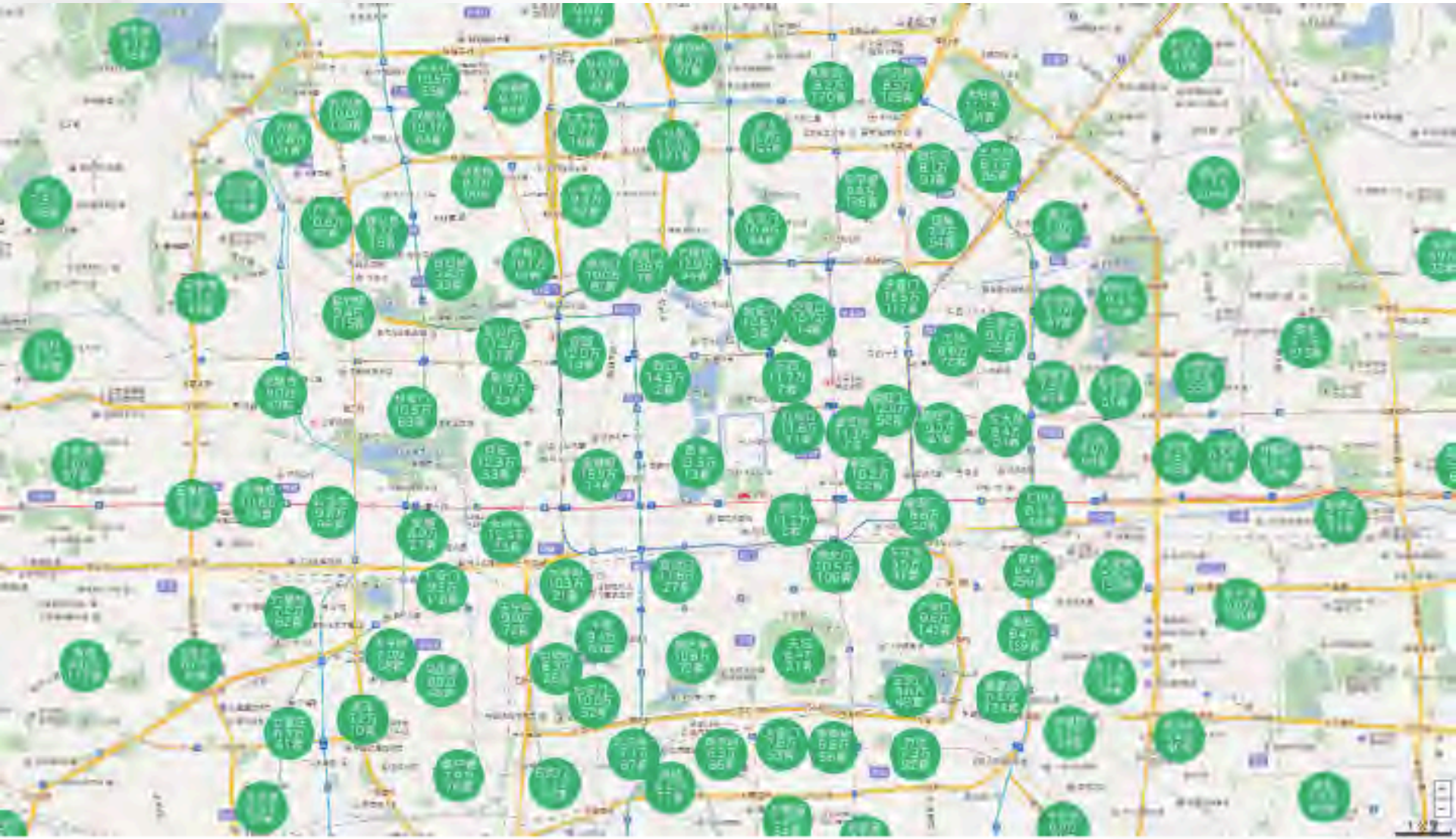
数据缺失如此严重，小区均价如何计算出呢?

估价系统难点及挑战— 交易数据的稀缺、稀疏和时变



面对数据缺失，传统的补足方法是向上汇聚，采用同类众值/均值进行补足

Location！ Location！ Location！



估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

同一商圈内小区，共享商圈内市政、商业和生活设施

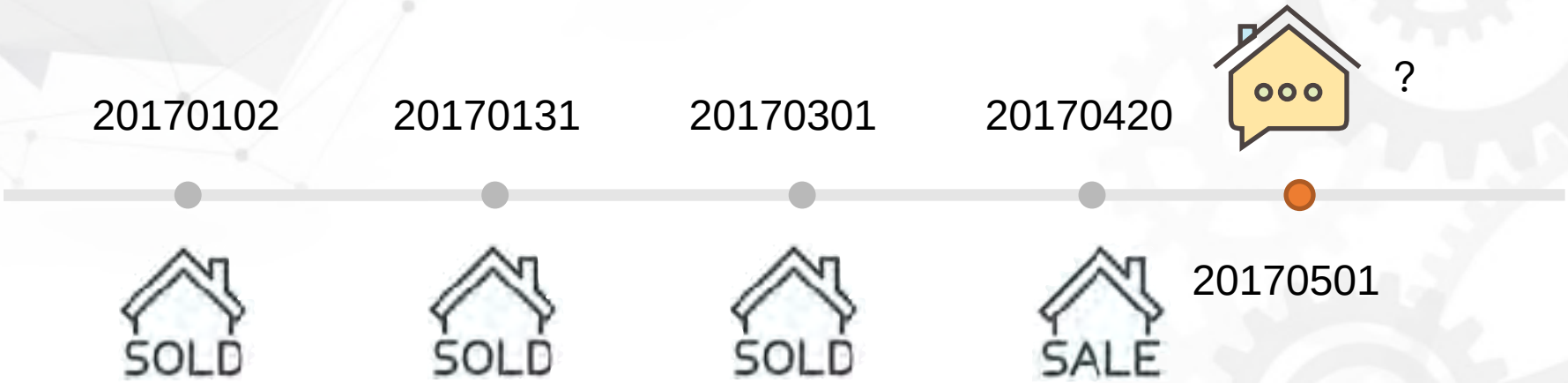
估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

同一商圈小区，虽然共享市政和商业设施，但各小区内公共设施不共享

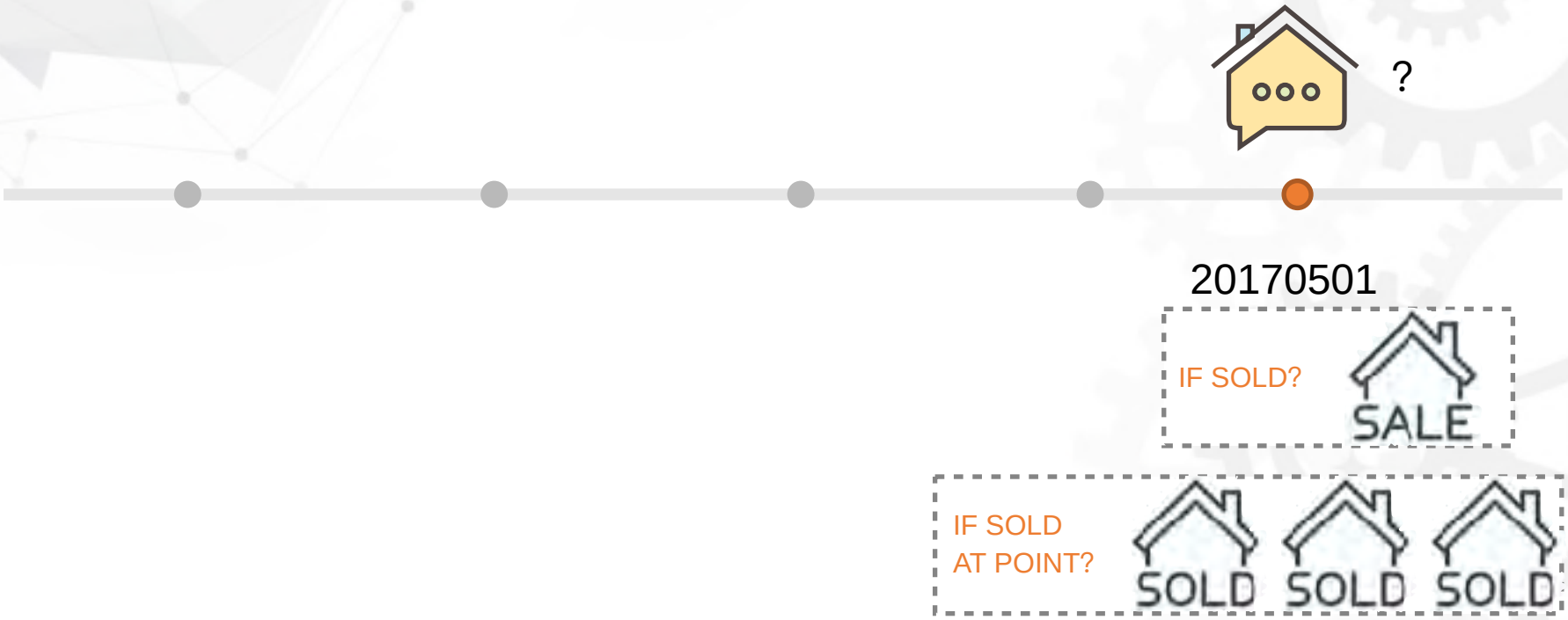
同一商圈，小区均价绝对差异很大

基本只要使用商圈价格补，就意味着你放弃了这个小区！

估价系统难点及挑战— 交易数据的稀缺、稀疏和时变



估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

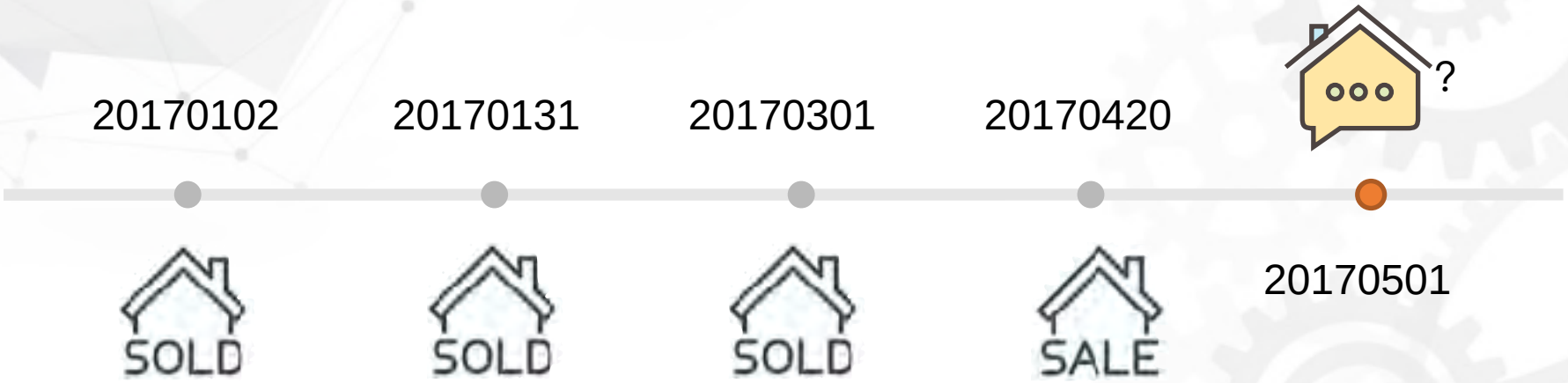


SALE —> IF SOLD ?

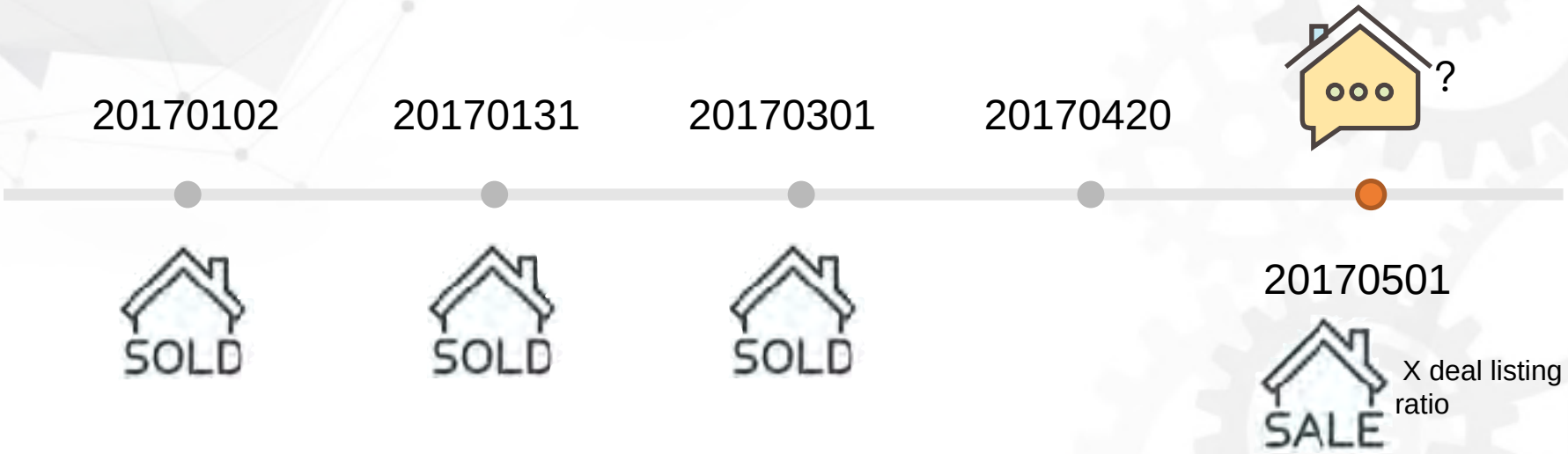
$$deal\ listing\ ratio = \frac{1}{n} \times \sum_{i=0}^{n} max(min\left(\frac{deal\_price\_i}{listing\_price\_i}, 1\right), LOWER\_BOUND)$$

$$future\ deal\ price = listing\ price \times deal\ listing\ ratio$$

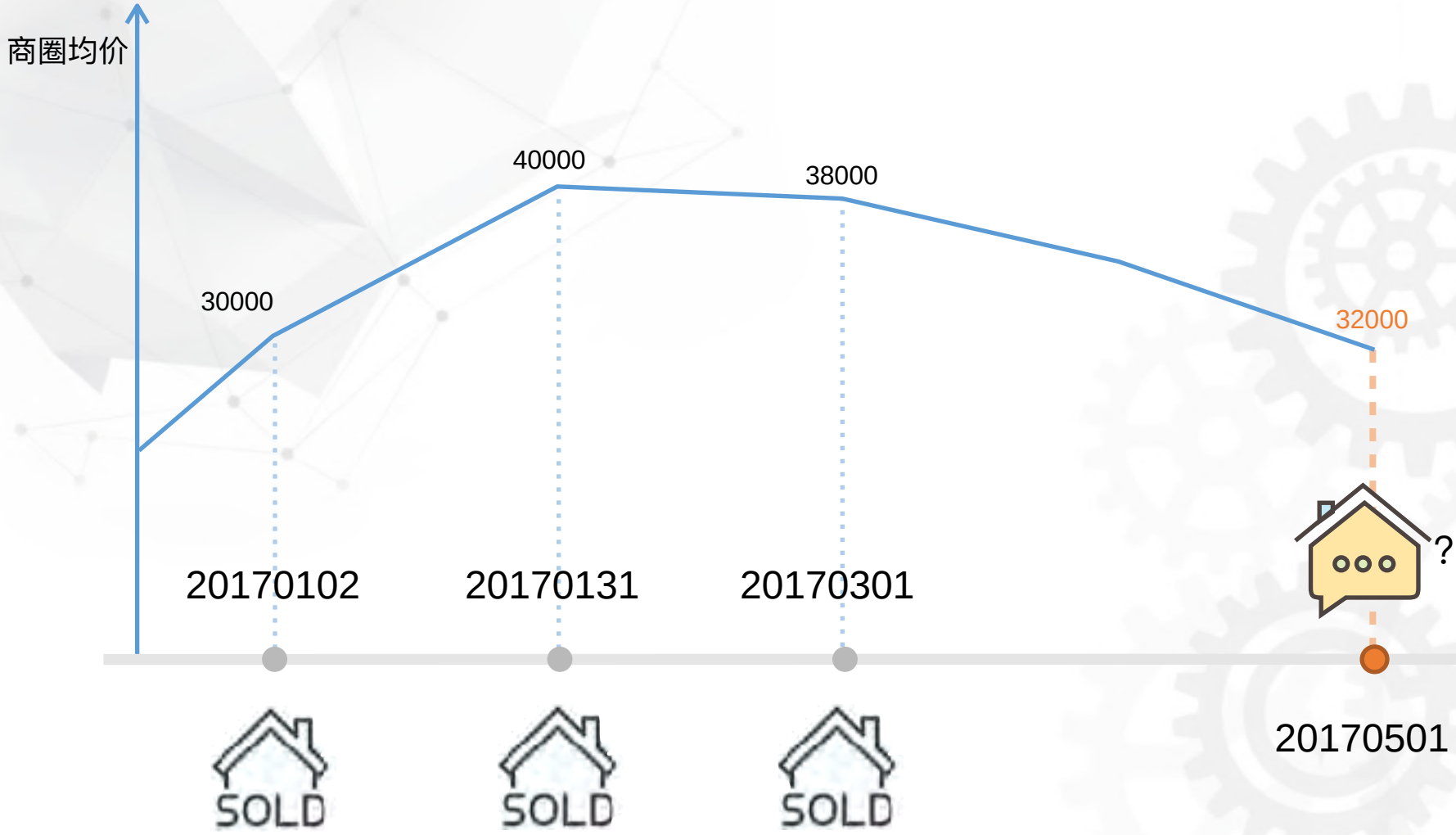
估价系统难点及挑战— 交易数据的稀缺、稀疏和时变



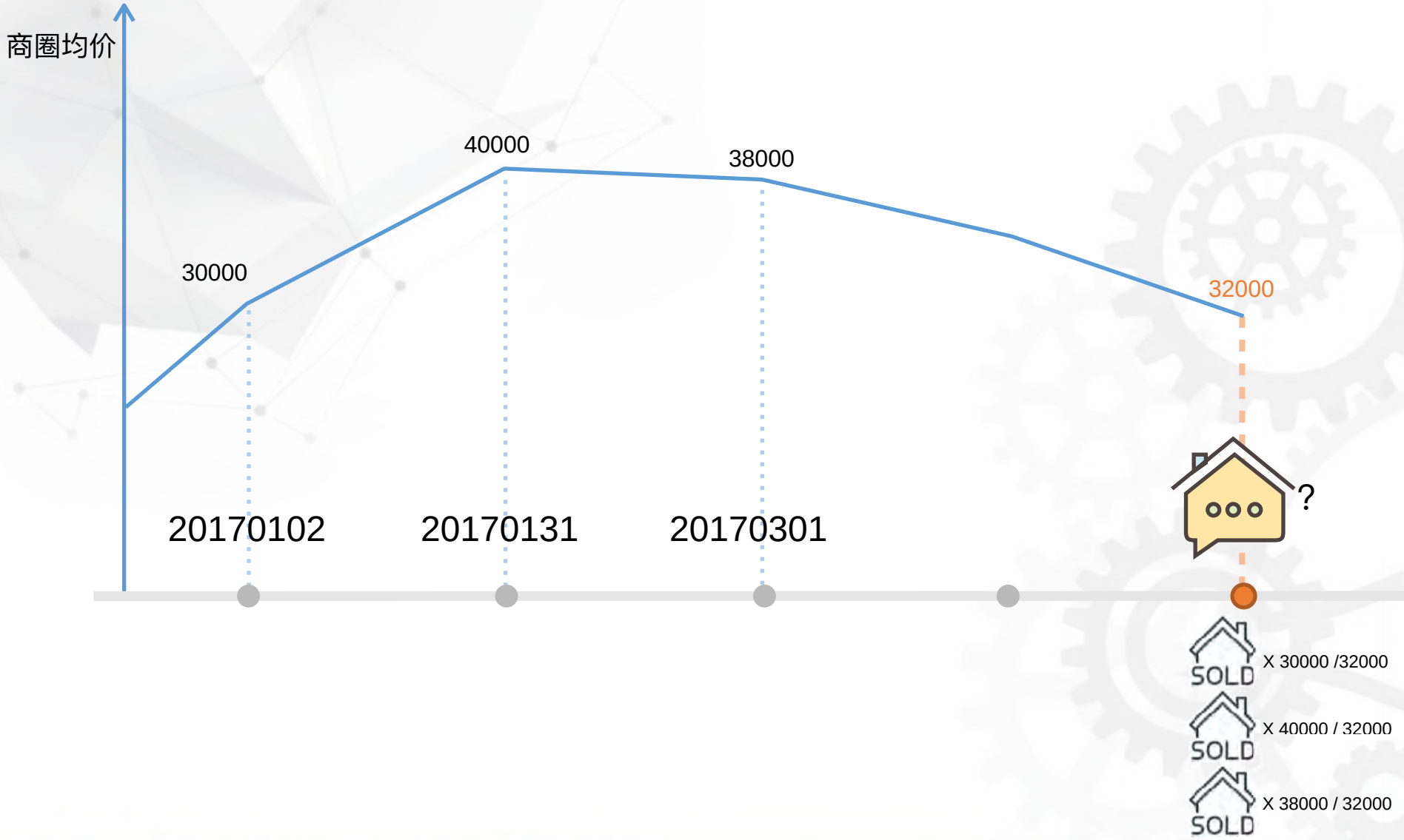
估价系统难点及挑战— 交易数据的稀缺、稀疏和时变



估价系统难点及挑战— 交易数据的稀缺、稀疏和时变



估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

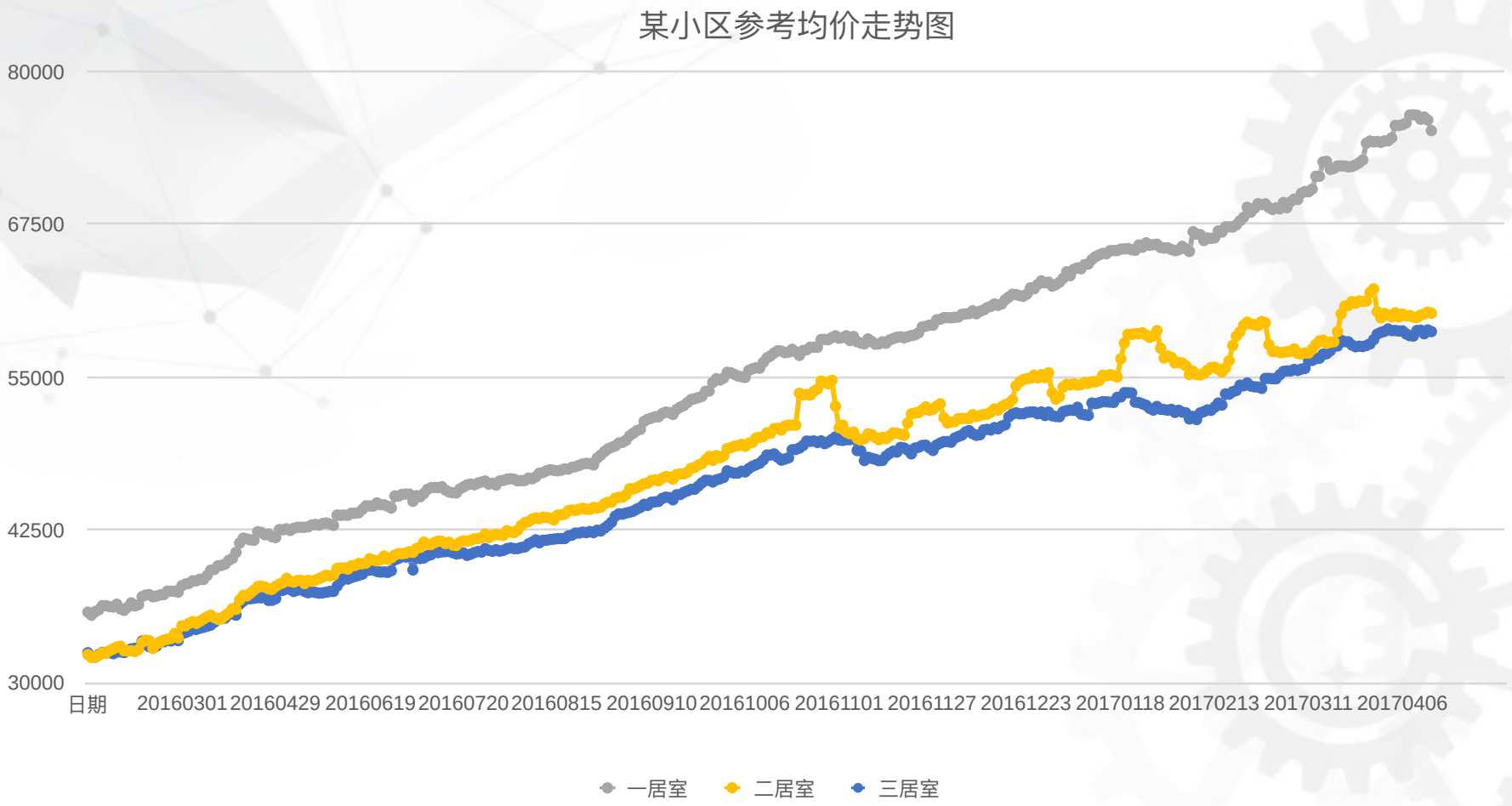


估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

$$\text{当前参考均价} = \frac{\sum_{i=0}^{n} \alpha_i * \text{第}i\text{个历史成交折算均价} + \beta\,\text{当前成交均价} + \gamma\,\text{挂牌折算均价}}{\sum_{i=0}^{n} \alpha_i + \beta + \gamma}$$

同时解决数据时变性和稀疏性问题！

估价系统难点及挑战— 交易数据的稀缺、稀疏和时变



估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

脏数据带来小区均价的阶跃!

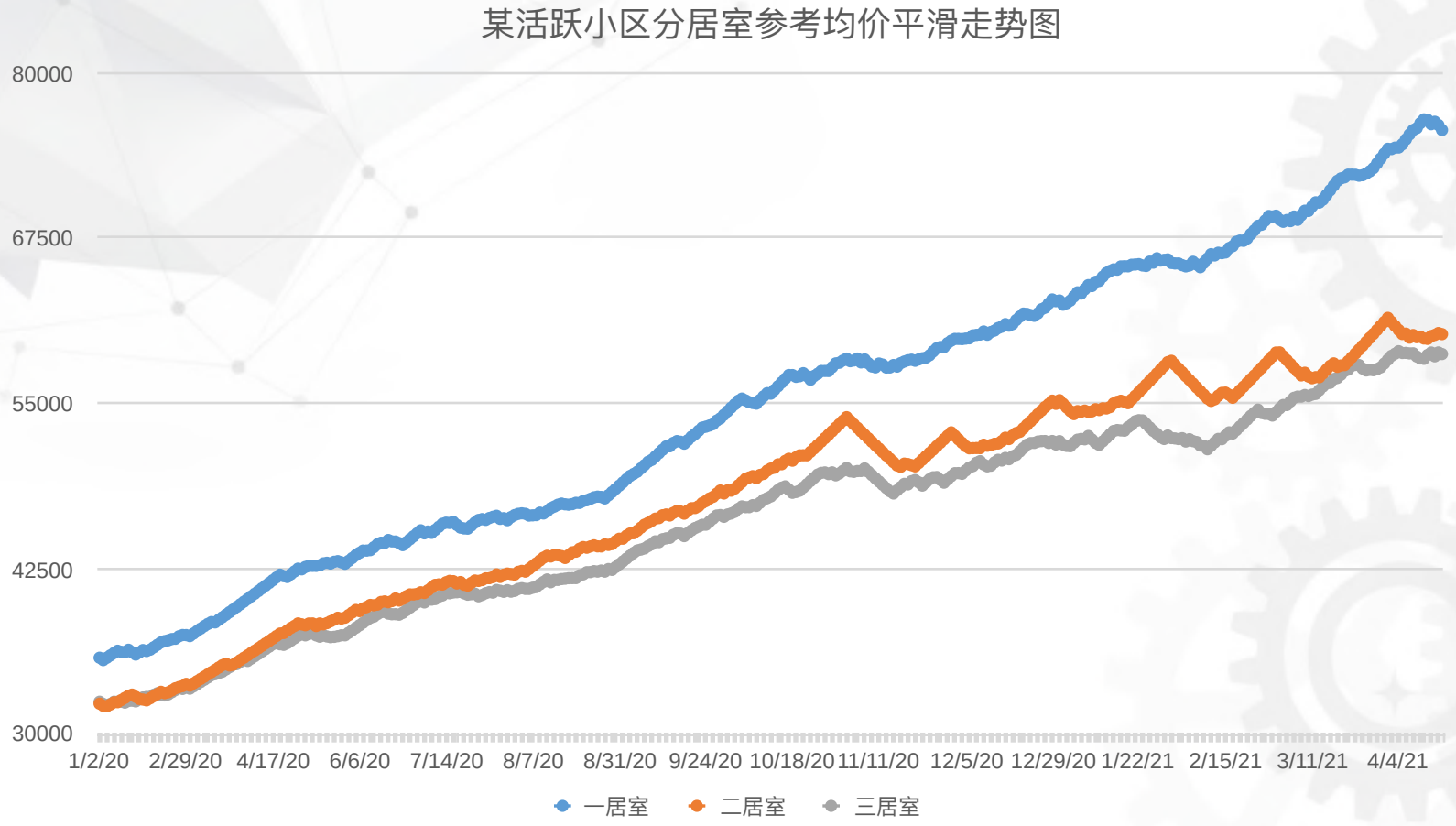
由于数据的稀疏性,很难通过统计的方法去除异常挂牌/成交，每一条成交和挂牌都十分重要

解决方案:

为参考均价添加平滑:当历史数据和新数据发生冲突时，选择相信新数据，但每天只信一点点

等待业务部门复核数据

估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

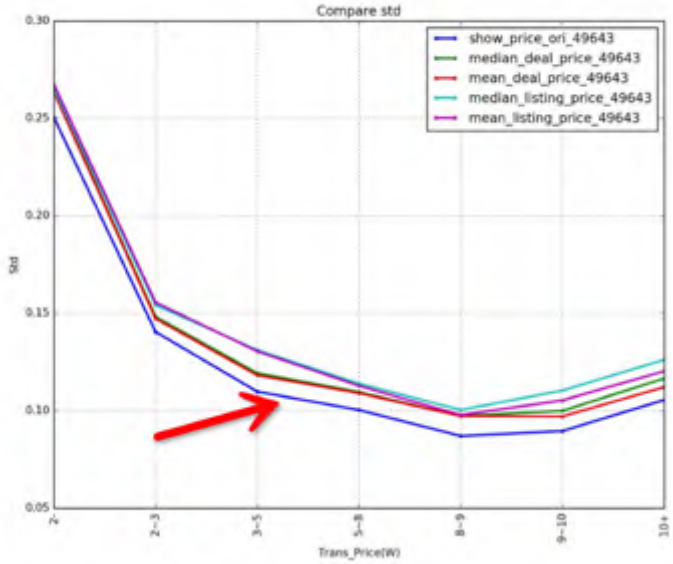


估价系统难点及挑战— 交易数据的稀缺、稀疏和时变

参考均价与其它均价方式对比

方法	参考均价	成交中位数价	成交均价	挂牌中位数价	挂牌均价
准确率	44.25%	39.5%	39.16%	40.42%	39.32%
覆盖率	73367/73367 100%	55551/73367 75.72%	55551/73367 75.72%	61951/73367 84.44%	61951/73367 84.44

多种均价成交比标准差



注: 本数据统计区间2016/07/01 ~ 2016/12/12

估价系统难点及挑战— 物理数据的复杂

估价系统难点及挑战— 物理数据的复杂

诡异的“精装修”

估价系统难点及挑战— 物理数据的复杂

在最开始的尝试中，装修是以“是否精装特征”的形式引入模型中的.....

大跌眼镜，某些城区中，装修的这个特征居然是负向的
同一套房子，精装修居然比非精装修估价要低.....不合常识！

统计样本中，精装修的房子均价确实要比非精装修低，为什么呢?

直接原因: 决定房子价值的最主要因素是地段
市中心成交多普装，偏郊区成交则多精装

根本原因: 数据稀疏，不够稠密，同价位房源之间缺少对比
模型未能捕捉到精装修和房价之间的真正关系

解决方法:
采用规则：按平米数，统一单价，折算成装修费用加进估价中

估价系统难点及挑战— 物理数据的复杂

类别变量

几万种不同户型

估价系统难点及挑战— 物理数据的复杂

常见类型变量处理方法

0 0 1 0 0 0 0 ... 0 0

哑变量
适用于类别种类不多时

户型打分

描述性变量
很难具体量化

估价系统难点及挑战— 物理数据的复杂

数客厅、阳台和卧室
窗户朝向
东 南 北

全南
全北
有南有北
有北无南
有南无北
无北无南

估价系统难点及挑战— 算法选择

数据特点决定训练算法！

估价系统难点及挑战— 算法选择

训练数据特点

✦特征异质化程度较高

- 同时存在类别变量和连续变量，连续变量之间有数量级差异；

✦含有很难清洗的脏数据

- 模型抗噪能力要强

估价系统难点及挑战— 算法选择

训练数据特点

✦可用数据量相对较小，真实成交价同特征之间的对应关系比较复杂
- 需要采用复杂度高的模型，但又要防止过拟合；

✦特征异质化程度较高
- 同时存在类别变量和连续变量，连续变量之间有数量级差异；

✦含有很难清洗的脏数据
- 模型抗噪能力要强

估价系统难点及挑战— 算法选择

Tree Ensemble算法

- Tree算法会自动对特征做交叉
- 基分类器Tree 能够较好的处理差异性特征
- Tree算法的抗噪能力很强

估价系统难点及挑战— 算法选择

Tree Ensemble算法

- Tree-ensemble算法模型VC维（复杂度）可控性强
  可以通过调节树数量、学习率、树深度等参数对模型复杂度进行微调，
  在过拟合和欠拟合之间取得平衡

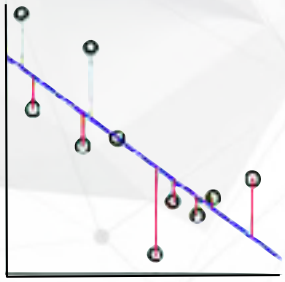
- Tree算法会自动对特征做交叉

- 基分类器Tree 能够较好的处理差异性特征

- Tree算法的抗噪能力很强

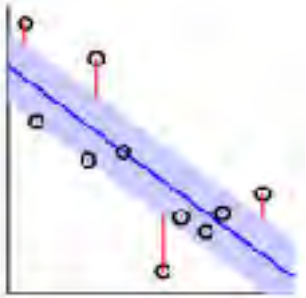
估价系统难点及挑战— 算法选择

模型复杂度控制



多元线性回归

加速: 加变量
刹车: 正则化



SVR

加速: 加变量，升维
刹车: 正则化

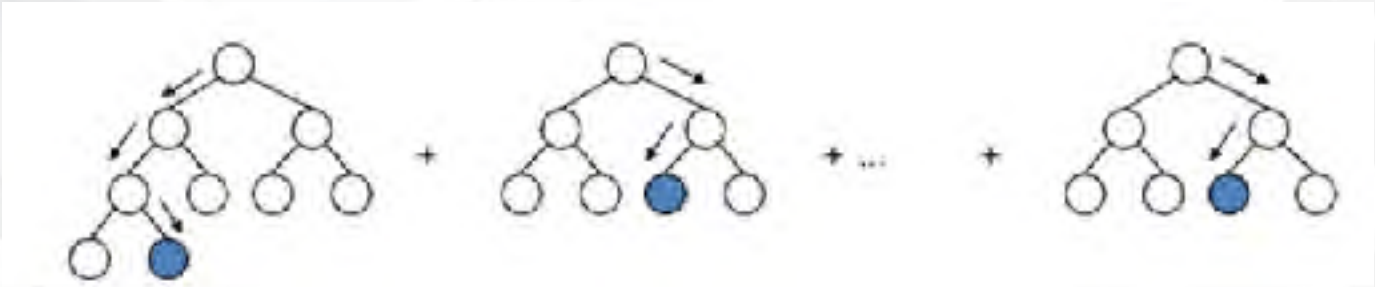
神经网络

加速: 加变量，加隐层
刹车: 正则化，dropout....

对模型复杂度的可控性较差，无法微调！
在数据量小的时候，容易失控！

估价系统难点及挑战— 算法选择

模型复杂度控制



Tree Ensemble算法

- 复杂度可加性
  通过增加Tree的方式，可以方便快捷的增大模型复杂度

- 复杂度可控性
  通过限定增加的Tree的深度、叶子结点分裂条件、学习率等参数，
  能够有效调节每次增加的复杂度

即使样本量较小，也可以采用逐步逼近的方式，使用较高复杂度的模型，
在过拟合和欠拟合之间取得平衡！

估价系统难点及挑战— 算法选择

几种Tree Ensemble算法效果对比

算法	准确率	平均误差
scikit-learn GBDT	81%	4.1%
scikit-learn RandomForest	75%	4.5%
XGBoost	79.5%	4.16%

内容概要

- 为什么要做估价
- 估价系统现状
- 估价系统总体设计
- 估价系统难点及解决方案
- 总结

总结

- 领域知识至关重要！领域知识至关重要！领域知识至关重要！
- 在数据稀疏的情况下，很难用统计的方法去除异常点，特征平滑能够救你！
- 机器学习模型强依赖数据，数据稀疏时，可能学习到违反常识的“知识”！需要领域知识进行修正！
- 数据稀缺，特征之间的差异性大，交叉关系复杂时，模型复杂度可控性是关键，Tree Ensemble算法是首选！

欢迎大家使用链家房屋估价

Q & A

THANKS

SequeMedia 盛拓传媒 ITPUB ChinaUnix