首届阿里数据创新平台大赛

数据格式说明

（一） 购买成交表（r_gmv_alipay）：

- **提取字段：**

买家 ID；商品 ID；卖家 ID；买家信誉值；卖家信誉值；商品所属类目（五级类目属性）；

商品单价区间；物流起点城市；物流终点城市；成交日期；商品 BC 分类；购买数量

- **提取说明：**

类目	对三个类目（女装，电器，3C）每天随机抽取一定量的交易记录
时间范围	2011 年 11 月 1 日至 11 月 30 日，共 30 天
抽样量	希望能在千万量级
抽样比	希望能占到所选类目的 10%以上，说明： 如果对应类目成交太多导致抽样过于稀疏可以选择其类目下的某个**叶子类目**； 如果还是稀疏可以只抽取**钻级以上买家**成交记录； 以及其他能够保证抽样质量的方法。
数据精度	价格如果敏感的话可以只呈现商品单价的一个大致区间，比如 10 元以内，20 元以内等； 地域希望能精确到城市，如果不行精确到省也可以接受。
数据保密措施	卖家与商品 ID 都匿名化； 地域粗粒化；

	交易记录定量占比会控制到比较低的程度，以防止通过类目销量大小来定位到商家； 选择过往的历史记录，保证现有的商品大部分都已经更新或者下架； 以及其他有必要采取的保密措施。

（二） 用户旺旺通信记录

- **提取内容：**

发起对话用户 ID；接受用户 ID；信息发送时间

- **提取说明：**

选取的类目、时间范围以及抽样比例要和之前成交表一样或是对应，最好能对应到相应的用户。

（三） ISPI 细分行业价格指数与物量指数

- **产品地址：**

http://www.aliresearch.com/i_ispi_data/

- **数据地址：**

目前在张文涛的个人表当中，可开放给技术处理人员——

iSPI 总体指数：rl_i_ispi

iSPI 分类指数：rl_i_item_ispi

iSPI 分类占比：rl_i_item_weight

- **提取说明：**

由于是经过处理的指数数据，因此既可以开放整体价格指数，也可以开放与上述对应类

目的细分价格指数。

同时需要开放的是指数数据的方法论，包括指标定义、算法、技术处理方法，以便学者可以参照相同的方法处理外部数据，与此次开放的数据一起进行研究。