智能中文模糊匹配软件

地址和公司名称的自动比对
应用于信用卡审批风险规则

上海协森计算机技术有限公司

2012年12月

公司介绍

上海协森计算机技术有限公司

- ✓公司团队前身，2003年属高阳科技（香港上市公司）国内银行信用卡应用服务事业部。
- ✓2005年1月，团队独立,成立上海协森信息技术有限公司，主要从事银行业应用软件产品研发及项目集成服务。
- ✓2008年12月，更名为上海协森计算机技术有限公司，注册资金增至300万元。
- ✓主要客户：交通银行、中国建设银行、中信银行。
- ✓团队：18人，其中软件工程师15人。

- ✓ 苏信
- ✓ 联系邮箱：suxin_ss@126.com
- ✓ 联系电话：18621996638

SeaSun IT

主题

- 中文模糊匹配技术为信用卡审批带来的好处
- 模糊匹配理论体系
- 中文模糊匹配工具应用体系结构
- 中文地址模糊匹配算法原理
- 企业名称模糊匹配算法原理
- 模糊匹配在信用卡审批业务的应用
- 批处理技术
- 运行平台和性能测试
- 客户化内容和服务
- 本产品的客户案例

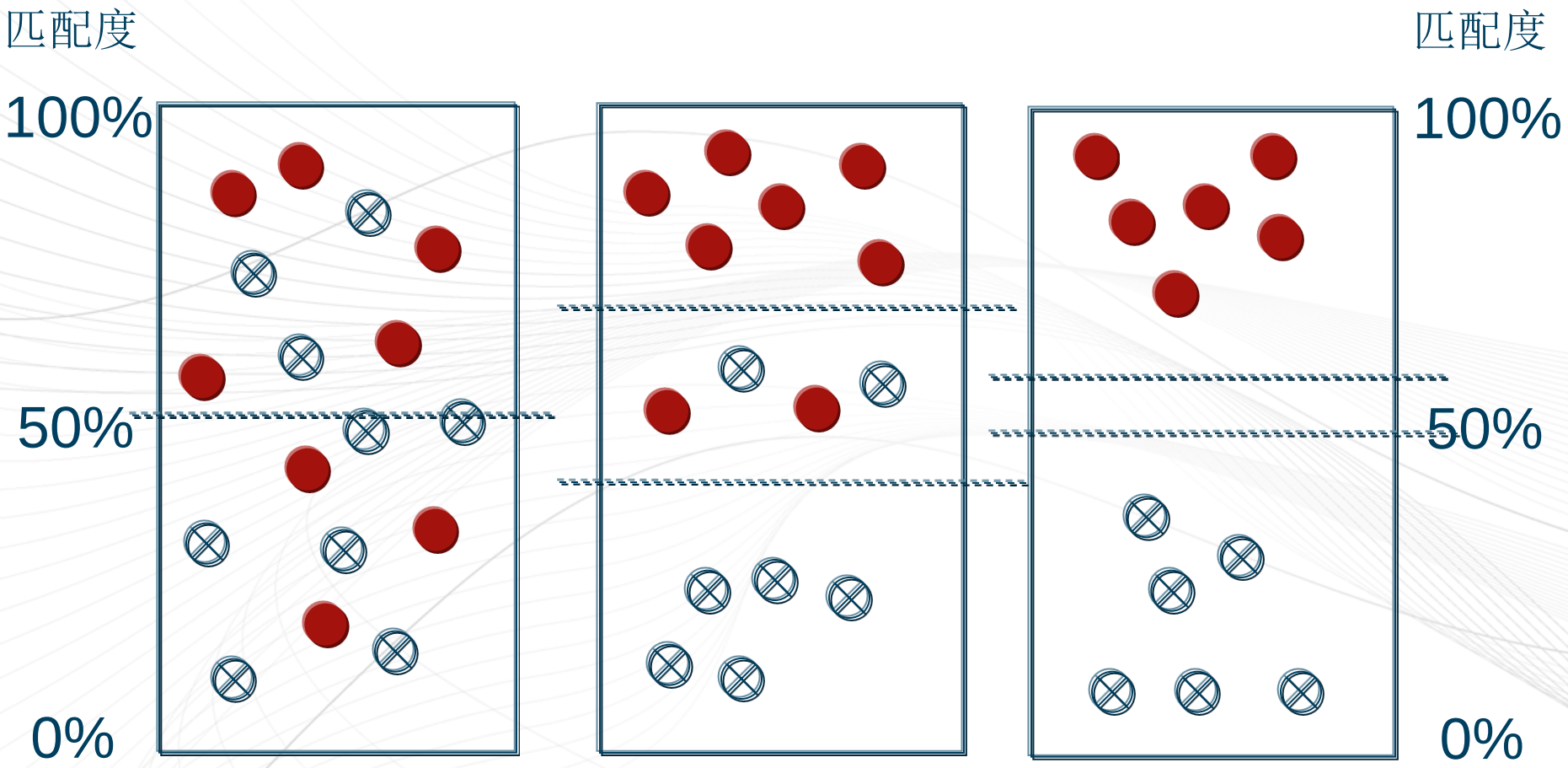
模糊匹配技术为信用卡审批带来的好处

- 信用卡审批流程的工作内容之一：比对客户信息
  - ✓申请表中的客户地址，分别与人行征信系统的客户地址、社保系统的客户地址、与公安局身份证地址等，比对
  - ✓申请表中的客户工作单位，分别与人行征信系统的客户工作单位、社保缴费的工作单位等，比对。
- 现状：地址比对、工作单位比对
  - ✓计算机系统提供准确匹配，约20%
  - ✓人工匹配，约80%。
  - ✓通常，申请表量较大的银行信用卡中心，约需要20人来做人工判断。
- 引入智能中文模糊匹配软件工具
  - ✓计算机系统提供模糊匹配，约80%
  - ✓人工匹配，约20%，减少至原来的1/4。
  - ✓从原来的20人的工作量，减至5人做人工判断。
  - ✓效率提高4倍，节省人工成本每年约100万元。
  - ✓地址分团、单位分团，识别集团欺诈。

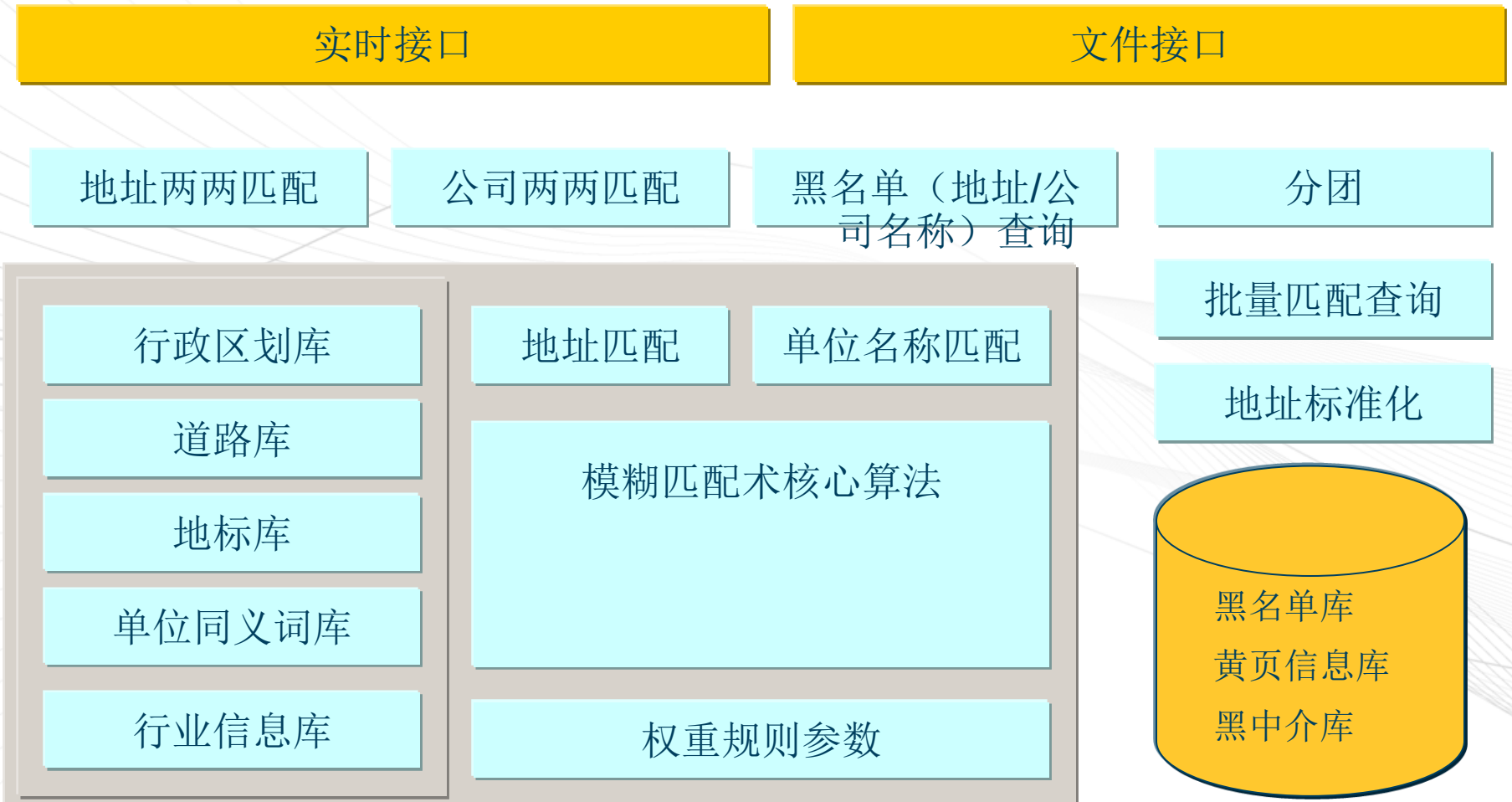
模糊匹配理论体系

- 模糊匹配的算法体系
  - ✓两个事物的匹配度=取值相同的属性的数量加权/属性的总数量加权
  - ✓属性有权重之分
- 模糊匹配算法应用
  - ✓基于关键字的资料检索
    - ➢论文库检索
    - ➢地理信息库检索
  - ✓基于分词的模糊匹配
    - ➢地址比对
    - ➢企业名称对比
- 地址模糊匹配技术
  - ✓地址比对的取值为“是”和“非”，即1和0。
  - ✓理论上，地址模糊匹配的正确率是不可能达到100%。

模糊匹配算法实现的目标



模糊匹配工具应用体系架构



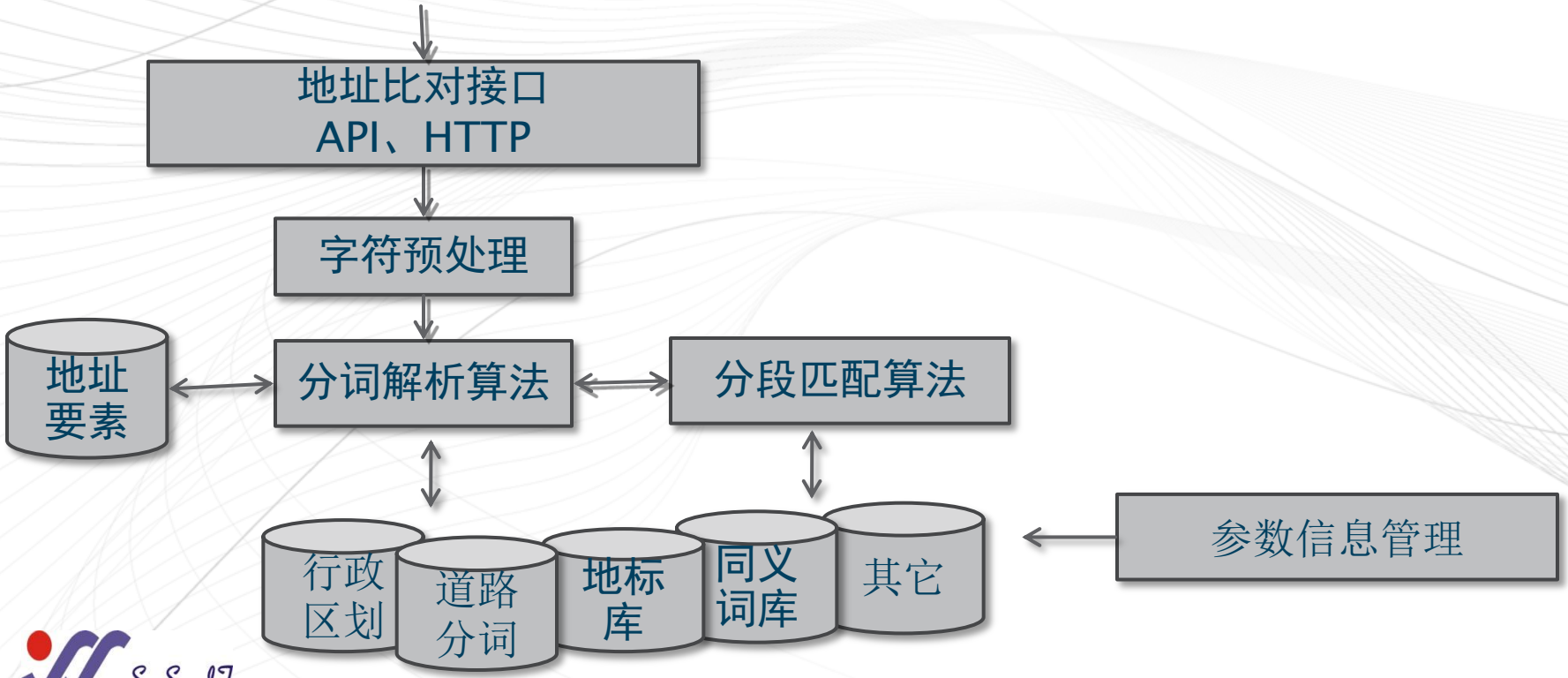
基于分级结构的地址库的分词解析

行政区划	道路	门牌号	地标名称	楼号	层号	室号	其他
广东省广州市天河区	天河路	45号	天伦大厦	2号楼	13楼		

- **行政区划段**
  - 精确匹配，一票否决。
- **道路段+门牌号、地标名称**
  - 道路段在道路库中找到，则精确匹配，否则，模糊匹配
  - 地标名称与道路门牌的含义是重复的，今后建立地标库就可做到精确匹配。
- **楼号+层号+室号**
  - 模糊匹配
- **规则与权重**
  - 利用规则和权重的组合算法，最终得到两地址对比的匹配度。

地址模糊匹配技术实现

- 基于分词的地址匹配技术
  - ✓ 通过构造地址名称的分级体系，形成地址分词库。
  - ✓ 匹配处理的过程是，解析分词，匹配分词，得出匹配结果。
  - ✓ 简繁体字的转换
  - ✓ 汉字数字与阿拉伯数字转换
  - ✓ 采用这种方法，易于维护地址分词库，并构建一个可自学习的地址库，不断丰富完善，提高匹配范围与匹配精度。



地址知识库

- 行政区划库
  - ✓建立了三级行政区划库
    - ➢23个省、5个自治区、4个直辖市、2个特别行政区
    - ➢283个地级市
    - ➢374个县级市、1636个县、852个市辖区
  - ✓根据业务需要，行政区划可扩充到四级
    - ➢镇、乡、村
- 道路库
  - ✓近100万条道路，并包含其对应行政区划的信息
  - ✓可升级的道路库
  - ✓自学功能
- 同义词知识库机制
  - ✓地标与详细地址的对应关系
  - ✓同一地点的两种写法的对应关系

主要算法机制

- 过滤特殊字符
- 繁体转换为简体
- 地址分词
- 解析行政规划段
- 地标知识库机制，地标名分词从字典中搜索道路门牌同义词并替换为标准名称
- 解析道路段
- 去除噪声词
- 号码段(大写/中文)数字格式转换解析门牌号码段
- 分段权重参数
- 一票否决参数

企业名称模糊匹配算法原理

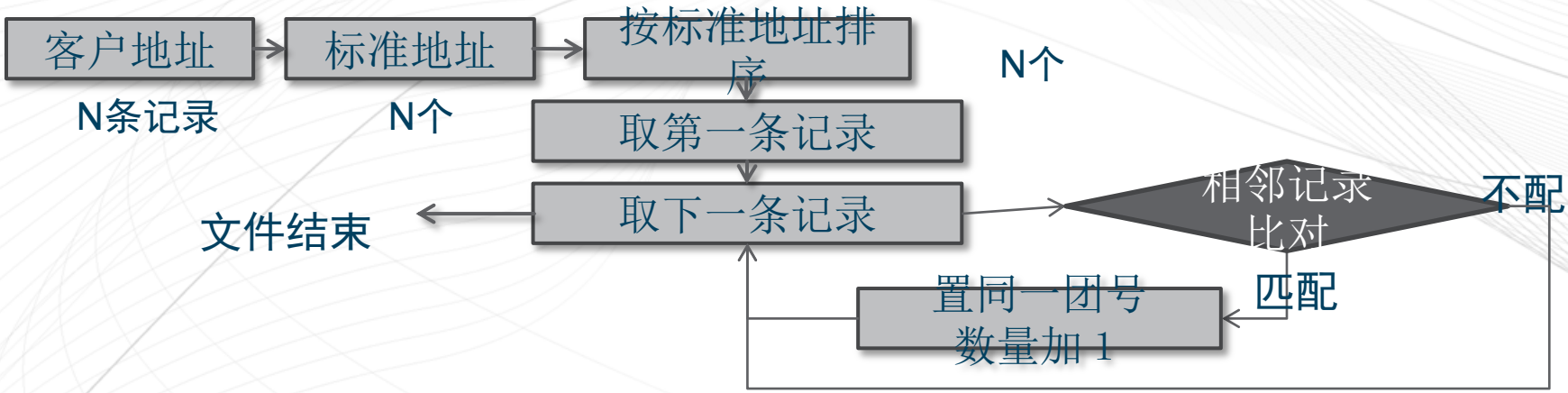
- 基于分词的模糊匹配算法
  - ✓企业标准名称，分段模糊匹配
    - ➢行政区+企业字号+行业属性+企业属性+分支机构+部门
    - ➢分段权重
    - ➢一票否决
- 同义词及知识库机制
  - ➢ 知名企业的同义词，2000个企业同义词库
  - ➢ 行业库
  - ➢ 关键词库

中文模糊匹配在信用卡审批业务中的应用

- 申请表信息比对
  - ✓申请表，分别与人行报告信息、公安部信息和社保信息，比对。
  - ✓申请表与历史申请表的信息的比对
  - ✓申请表与人行报告历史库信息的比对
  - ✓申请表与欺诈历史库信息的比对
  - ✓单笔黑名单查询
- 批处理应用
  - ✓批量黑名单查询
  - ✓分团/插团功能，防集团欺诈
  - ✓批量查询

批处理技术

- 分团性能优化算法
  - ✓地址分团时，标准化地址并排序后，街道相同的地址都会排到相邻位置，这样，分团的效率就提高很多。如下流程图。
  - ✓公司名称分团时，标准化后，按行政区分批进行分团。
- 黑名单（黑中介）、黄页、历史编码库性能优化算法
  - ✓原信息库内容，标准化、排序、分区。
  - ✓待查记录直接定位到对应的区块中查询，提高效率。



运行平台及性能测试

- 硬件环境
  - ✓CPU：Intel 或AMD 8核 3GHz以上
  - ✓内存：16G以上
  - ✓硬盘：320G以上
- 支持两种软件环境
  - ✓Windows Server 2008 企业版R2, 64位操作系统，Microsoft .Net Framework 4.0、IIS 7
  - ✓UNIX平台（Linux），DB2数据库，J2EE体系

- 实时接口
  - ✓ 单笔响应时间，0.16秒，
  - ✓ 20笔并发，最长响应时间0.38秒，
  - ✓ 50笔并发，最长响应时间0.55秒。
- 地址单对匹配
  - ✓ 10万，31秒
  - ✓ 100万，9分钟
  - ✓ 500万，51分钟
  - ✓ 1000万，1小时37分钟
- 公司单对匹配：
  - ✓ 10万，32秒
  - ✓ 100万，9分钟
  - ✓ 500万，55分钟
  - ✓ 1000万，1小时43分钟

- 地址分团匹配
  - ✓ 10万，35秒
  - ✓ 100万，10分钟
  - ✓ 500万，1小时1分钟
- 公司分团匹配
  - ✓ 10万，38秒
  - ✓ 100万，11分钟
  - ✓ 500万，1小时12分钟
- 黑名单导入
  - ✓ 10万，5秒
  - ✓ 100万，2分5秒
- 黑名单匹配（比较6个字段，1个单位名称，2个地址，3个电话）
  - ✓ 100条查100万条黑名单，4分钟
  - ✓ 1000条查100万条黑名单，17分钟

客户化内容和服务

- 算法偏好的调整
- 两两匹配实时接口
- 两两匹配文件接口
- 数据库模糊匹配查询
- 分团功能

本产品的客户案例

- 交通银行信用卡中心
- 建设银行风险管理部和信用卡中心
- 中信银行信用卡中心

谢谢！