

## 智慧盒子介绍

NOTICE: Proprietary and Confidential

This material is proprietary to DataESP Private Limited. It contains trade secrets and confidential information which is sole property of DataESP. This material shall not be used, reproduced, copied, disclosed, transmitted, in whole or in part, without the express written consent of DataESP Private Ltd.



# 团队介绍



**陈晓朋**  
商业拓展

博士，电子电气工程，  
滑铁卢大学  
本科，电子电气工程，  
新加坡国立大学



**王阳**  
科技研发

博士，工业系统设计，  
滑铁卢大学  
硕士，工业系统工程，  
清华大学  
本科，电子电气工程，  
清华大学



**王坦**  
公司运营

硕士，科技运营和管理，欧洲工商管理学院  
硕士，计算机工程及算法，滑铁卢大学  
本科，计算机软件工程，滑铁卢大学



**陈宁**  
算法研究

博士，计算机软件工程，  
华盛顿大学  
硕士，计算机软件工程，  
复旦大学  
本科，数学，复旦大学



**王延博**  
财务

博士，金融，欧洲工商管理学院  
硕士，计算机工程，麻省理工学院，  
本科，计量金融，新加坡国立大学  
本科，运营管理，新加坡国立大学



**许文辉**

前星展银行主席

前新加坡电信有限公司  
主席

前新加坡航空公司主席

前淡马锡控股公司董事



**黄文华**

Charles & Keith创始人  
人



**陈运琼**

nTan创始人

# 核心竞争力



智慧盒子的数据分析算法，无论是其先进的科技还是巨大的应用范围，均属世界一流



## 专利科技

自主研发所有的科技和分析  
算法

实践证明对于发现各行业的  
**潜在问题**及提供可行性解决  
方案很有效



## 专注人才

核心团队由来自各行各业的  
专家组成

深入分析整个行业纵向跨度



## 专有平台

模块式平台，高灵活性，  
高延展性，使用及其方便  
甚至无需安装，持续供给  
智能化解决方案

与大多数的线上/线下，商业用途的/开放源码的管理  
系统均兼容



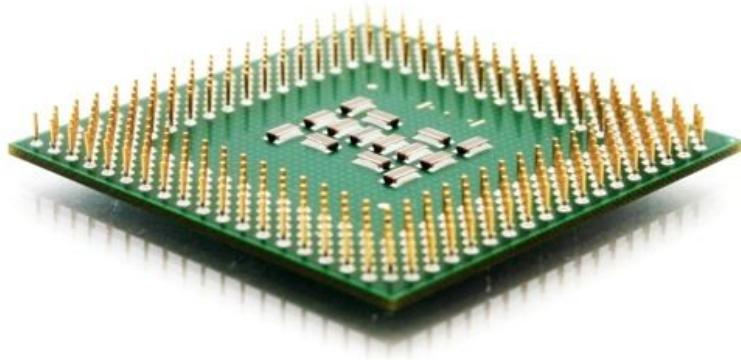
## 专业贯彻

根据您的数据提供全自动化  
的，即时的，处方式的智能  
建议

定期向管理层汇报，使得新  
政策推行能顺利展开

企业可以100%自主使用我  
们的平台服务，无需额外协  
助或咨询服务





# 科学与科技

# 迅速增长的数据：你准备好了吗？



# Big Data is a New Source of Competitive Advantage



*"The rate of return on analytics investments is \$10.66 on every dollar invested. That means, that if you're gathering data and you haven't deployed analytics, you're crazy."* - David O'Connell, principal analyst at Nucleus Research

## Example Applications

- “ Making the right offer to a potential customer at the right time and place ”
- “ Quantify influencers and network of influence connections ”
- “ Be able to offer value-based pricing to B2B customers with greater confidence in outcomes ”
- “ Correlating marketing practices/spend with marketing outcomes ”
- “ Real-time rerouting or trucks and deliveries based on traffic conditions local inventory status costs ”

## Business Performance Advantage

REVENUE 1999-  
2009

(10YR CAGR)  
 Leveraging  
big data

EBITDA 1999-2009  
(10YR CAGR)

# 智能构思: Automate Insight Generation



为了使大数据能够真正成为实用的，容易操作的并能马上产生收益的工具，在企业现有的分析软件基础上，还需要多加一层革命性的研发成果：智能构思层



# 智慧盒子方案流程



智慧盒子能在客户提供的内部数据的基础上，进一步收集相关的外部数据，使得分析结果能够真正被最优化。在我们核心团队的指导和监督下，一些特殊的分析方案还可以通过众筹实现



# DataX-Ray 科技: 从数据到智慧

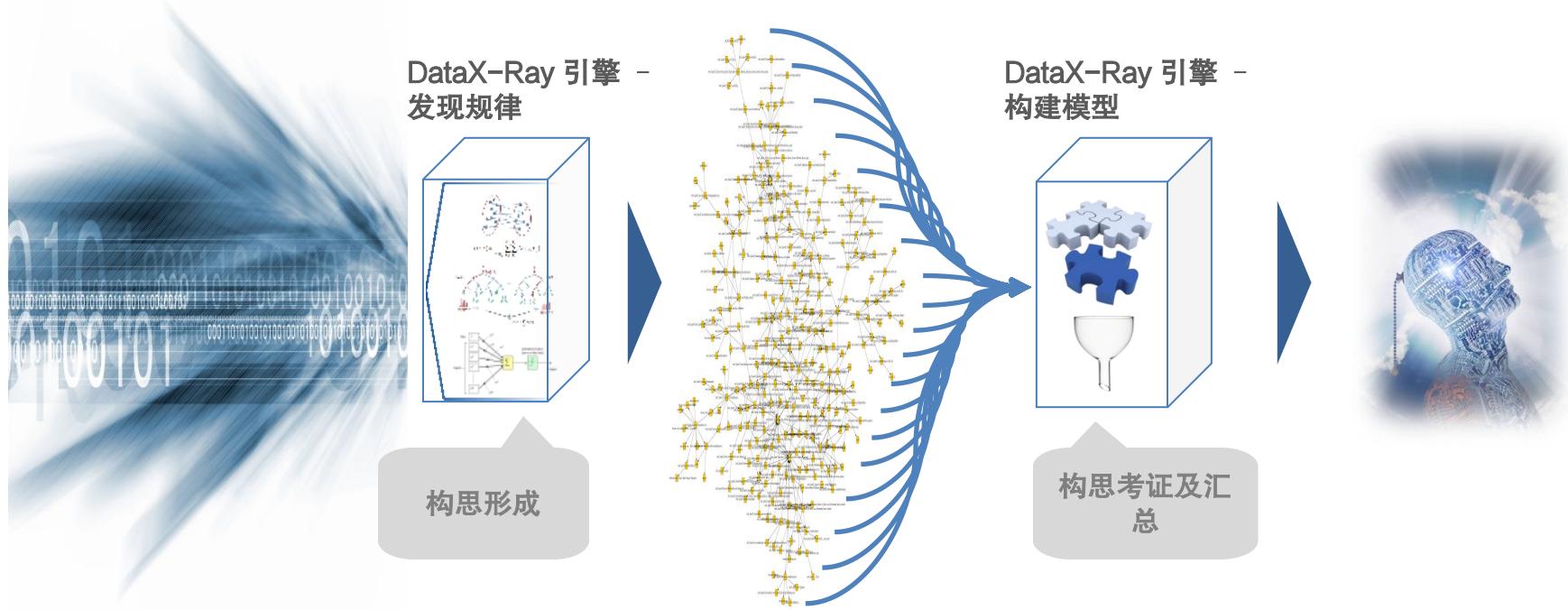


智慧盒子的 DataX-Ray 科技能够在没有行业专家监督的情况下，自动得出极具价值的智慧。其在高阶数据中找出规律的能力更是首屈一指

# 数据吸收消化

规律及构思

实用性智慧



我们的引擎能对任意阶次和大小的数据进行构思、考证规律，并将它们最终汇总成实用性强的智慧。与传统大数据处理相比，新流程不再受限于操作人员甚至数据专家有限的知识储备，不放过任何一种客观存在却无法被人为发现的最优方案，从而发现大数据领域的新大陆。

# 用“凯恩斯主义”的方法服务双面市场



我们的核心团队将在分析市场中的以下“部门”中担任“主管团体”：(a) 将业务难题转化成数据分析问题；(b) 将各类数据分析问题通过众筹的方式分配给合适的大数据方案提供者；(c) 在把最终结果反馈给客户之前，对产出的解决方案进行验证和提升



利用多种前期和后期处理技术，发掘最有价值、最相关的规律

**关联规则挖掘** 挖掘数据和结果之间隐藏的非线性关系

客户变量:

- 年龄
- 收入
- 访问次数
- 是否有宠物
- 是否有车
- 眼睛颜色
- 是否有房
- ...

客户行为:

- 对促销活动的反应
- 是否购买产品
- 给产品和服务的评分
- 花销弹性
- 是否有尝试新产品的倾向
- 是否倾向于赊账消费
- ...

**矩阵因子分解(奇异值分解)** 有效地将成百上千的属性特征归类到各个可操作的小组内



**K-最近邻分类器** 逐步提高分析邻近变量对自身影响的能力



**限制波尔兹曼机** 从不完整的数据中发觉关系并预测结果

**随机梯度下降模型** 通过优化参数选择来严格还原客户行为



**去糟取精** 将不必要的因素从初步构思中去除，从而得到价值和相关性都最高的规律



**成功将噪音从规律中去除**

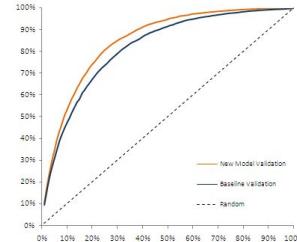


在规律已经找出之后，再用一套完整的模型构造技术来考证，严格保证其相关性和实用性

## 先进的数据分析技术

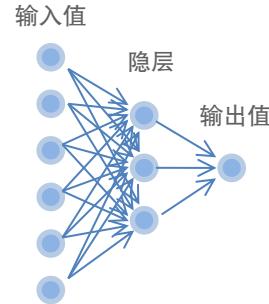
### 逻辑回归算法：

从原始数据中找出一阶关系



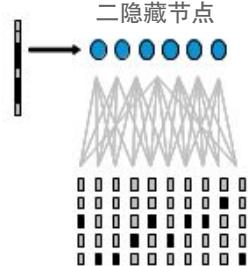
### 神经网络：

擅长从非线性、关系密切且复杂的数据中找出隐藏的规律结构



### 限制波尔兹曼机：

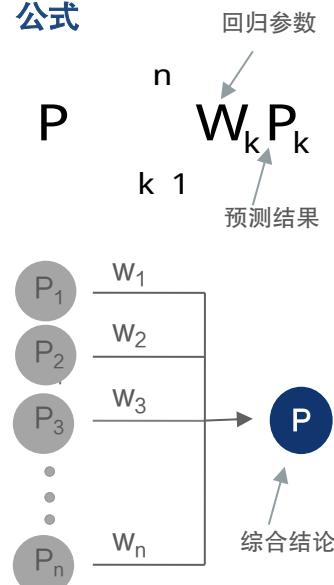
擅长从优质但数量稀少的数据中找出隐藏的规律、结构



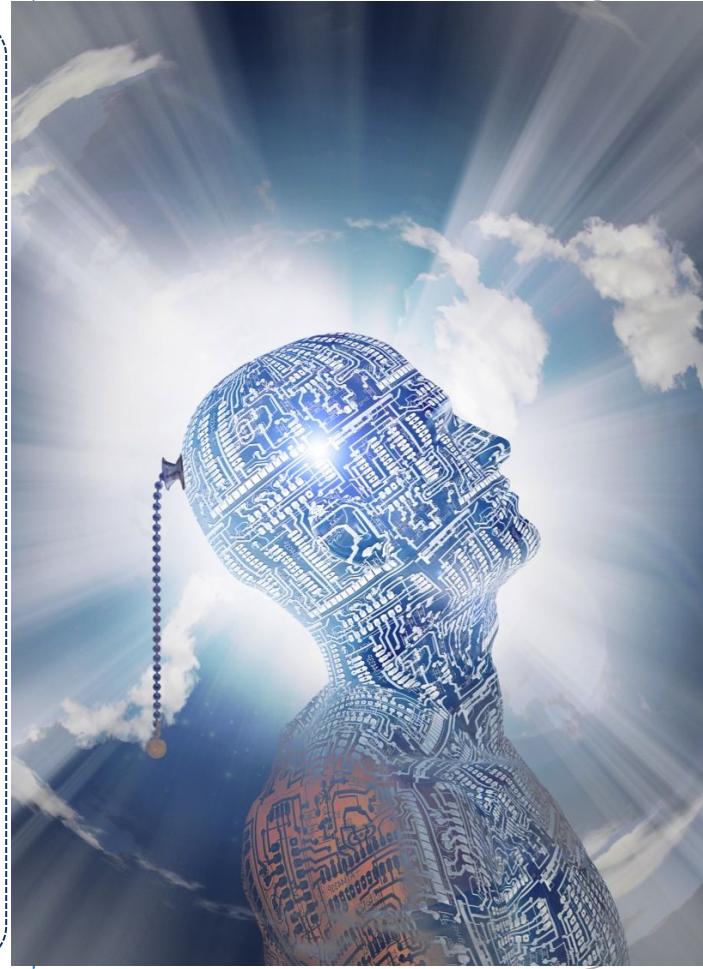
## 综合考证

对用不同分析技术得出的结论进行综合考证，使得最终建立的模型实用性强、更准确

### 公式



## 世界级智慧



# Multiple Segmentation Approaches Were Followed



A number of segmentation approaches were developed to establish accurate benchmark

## Single Variable Segmentation:

Based on Utilization (Av Mileage/year)	3
% missing: 19%	
Based on P11 Values	5
%missing: 2.5%	

## Three Variable Segmentation:

Based on P11 values+ Utilization+ Age on Fleet	60
%missing: 21%	

## Five Variable Segmentation

Based on Fuel Type+ Transmission+ Engine Size+P11 value + Utilization	255
%missing: <1% <sup>1</sup>	

Based on Fuel Type+ Transmission+ + Engine Size+P11 value + Age on Fleet	239
%missing: <1% <sup>1</sup>	

## Six Variable Segmentation

Based on Fuel Type+ Transmission+ #of Doors+ Engine Size+P11 value+ Age on Fleet	529
%missing: <1% <sup>1</sup>	

## Seven Variable Segmentation

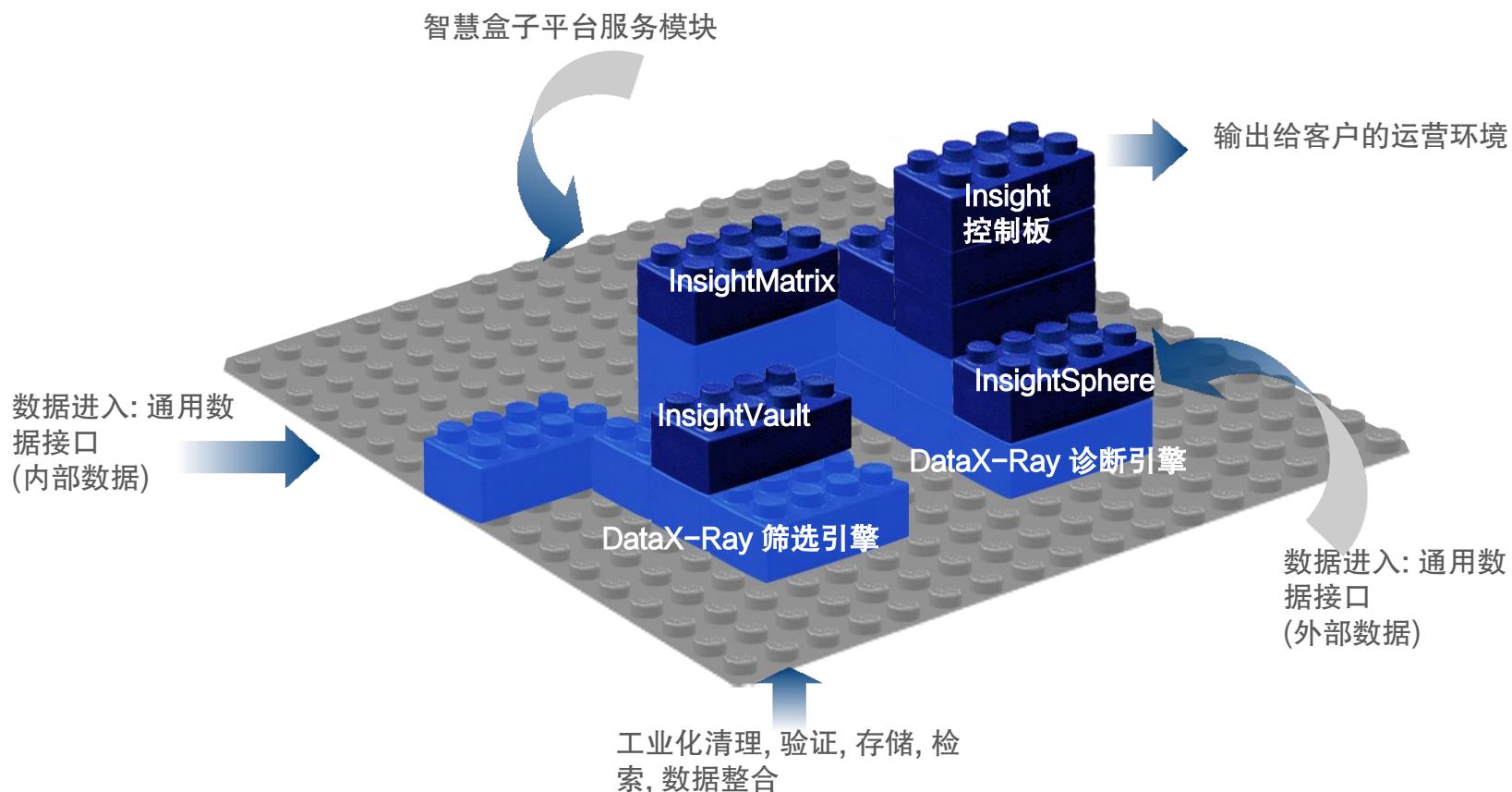
Based on Comm vs. Non-Comm. Utilization + Fuel Type + Transmission + # of Doors + Engine Size + Make Year + 3 Tier P11 Classification	671
%missing: <1% <sup>1</sup>	

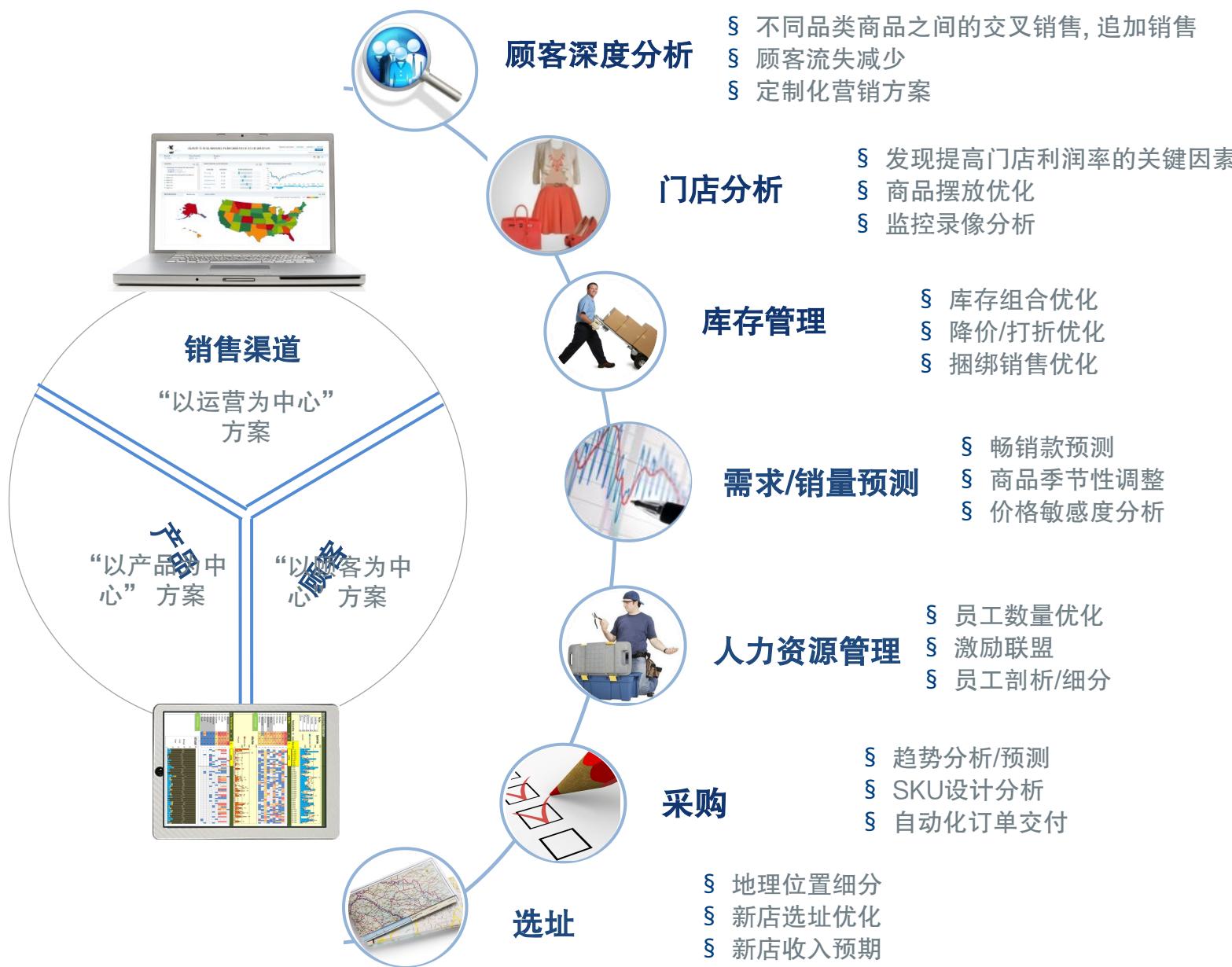
1. For 5, 7 and variable segment, if only one variable is missing then the vehicle is segmented based on other variables. Top 25% segments cover 75-80% vehicles and have greater than 30 vehicles per segment

# 精简的定制化数据分析平台



为了经济而快捷地满足客户的商业需求，我们建立了一个易调控、易延展的，含多种数据分析模块的平台

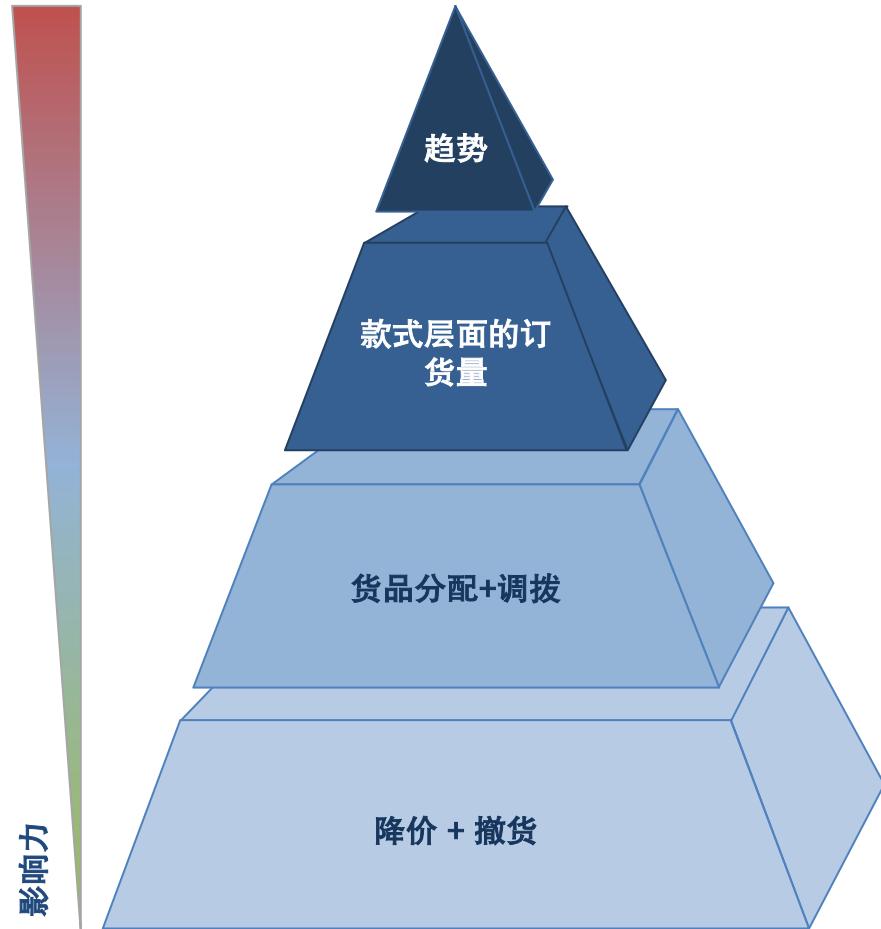




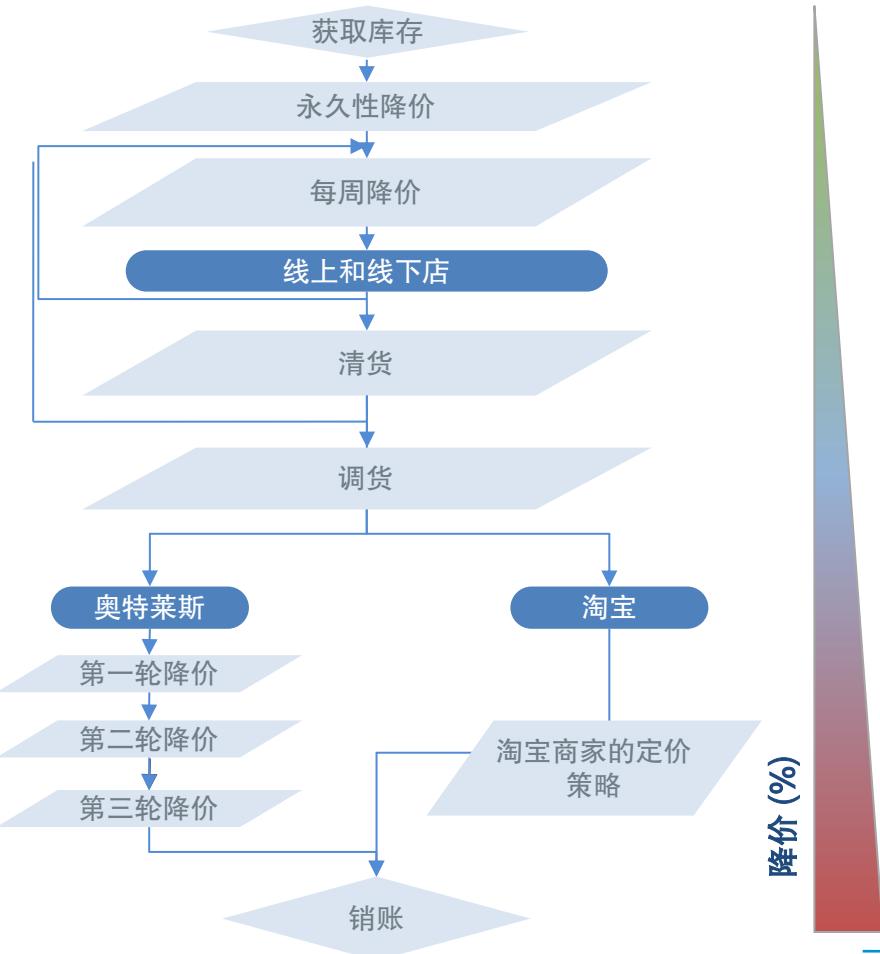
# 运营困难: 库存、定价优化

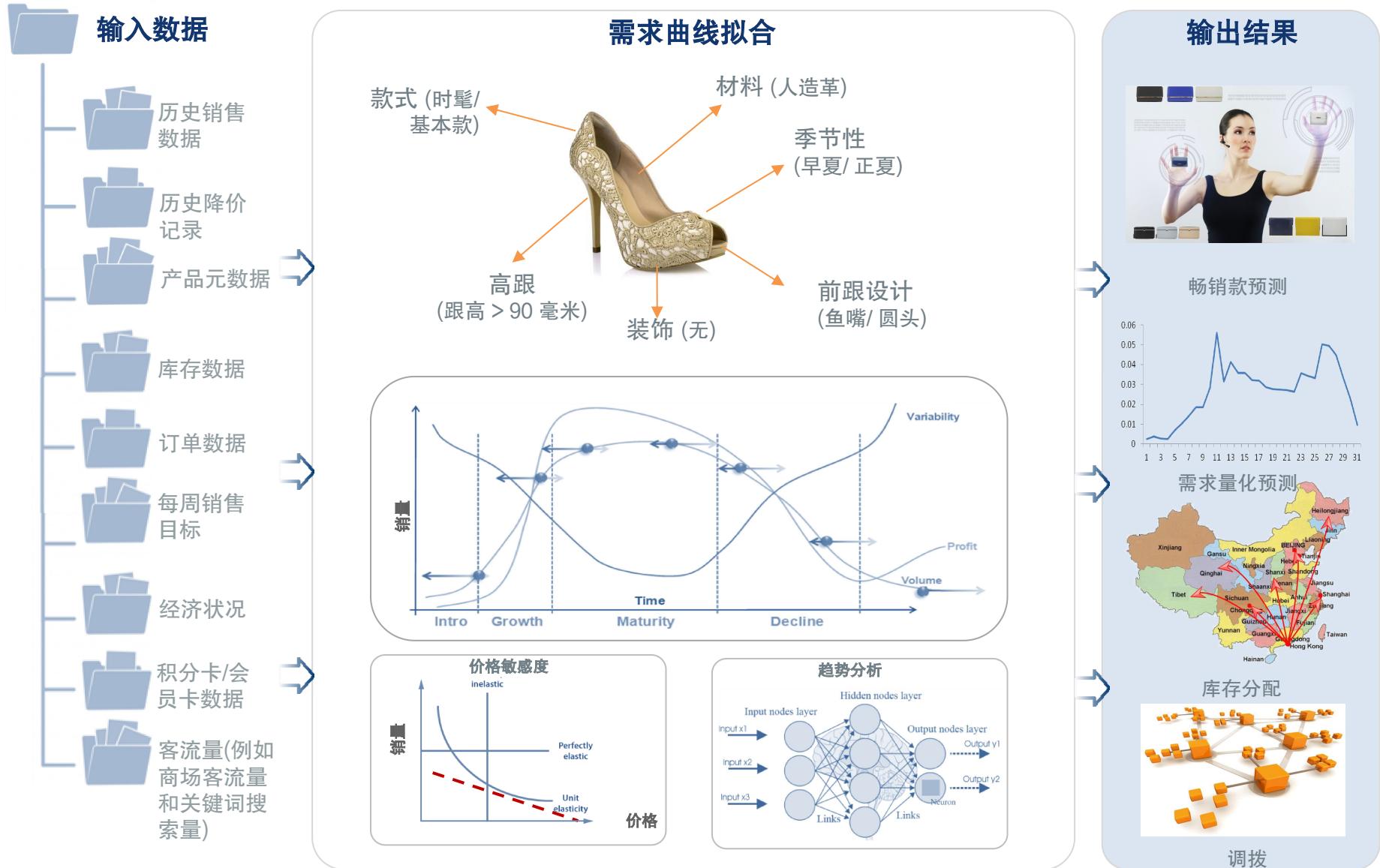
InsightBox 拥有一整套帮助解决零售商普遍面临的运营困难的数据分析方案

## 库存优化



## 定价优化





# 可视化采购



InsightBox的深度学习神经网络能够从众多产品图片/设计中预测出畅销款，从而使得采购经理能够准确而直观地进行趋势分析

## 输入数据

Materials: Embossed PU & Natural PU / Printed Fabric & Natural PU / Natural PU / Vinyl & Patent PU  
Closure: Heel / Garden Open Toe / Sandals  
Height: 105mm  
Age Group: 24 - 41  
Launch: L3-2014-JT  
Theme: Urban Wilderness / Urban Wilderness / Coloured Vinyl



Material: Natural PU & Embossed PU & Mirror Metallic PU  
PU / Patent PU & Embossed PU & Mirror Metallic PU  
Description: Heel / Covered / Stringback  
Height: 90mm  
Age Group: 24 - 41  
Launch: L3-2014-JT  
Type: Fashion Basic  
Theme: Pop, Satin



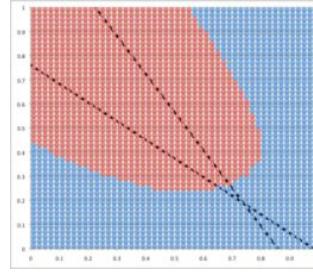
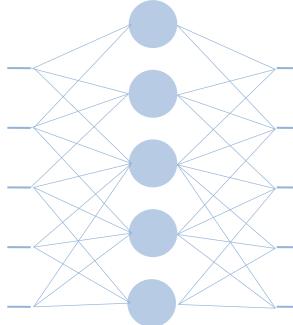
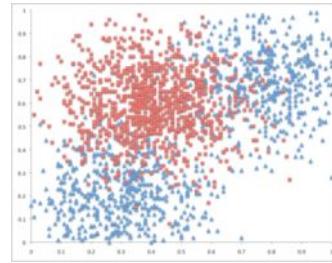
Material: PU  
Class: Trendy  
Handle Type: Single w Strap  
Size: M  
Island: Pop Safari



Material: PU  
Class: Trendy  
Handle Type: Chain/Strap  
Size: M  
Wall: Equilibrium



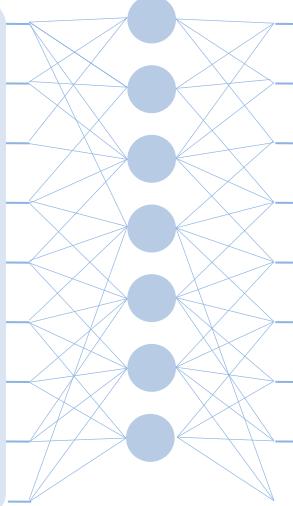
## 深度学习神经网络



$$\hat{R}_{ci} = \sum_{f=1}^N (p_{if} \cdot \sigma(b_f + \sum_{j \text{ rated by } c} w_{score(c,j)} \cdot q_{jf}))$$

广泛大量的产品参数  
(可列举数百个):

- 年龄
- 收入
- 访问次数
- 是否有小孩
- 汽车维护情况
- 是否有房
- 大学专业
- 就业行业
- ...



顾客的行为:

- 对促销活动的敏感度
- 是否购买该商品
- 对该上平/服务的评价
- 花销弹性
- 是否愿意尝试新的产品
- 是否倾向于赊账消费
- 贷款拖欠的倾向
- ...

## 输出结果



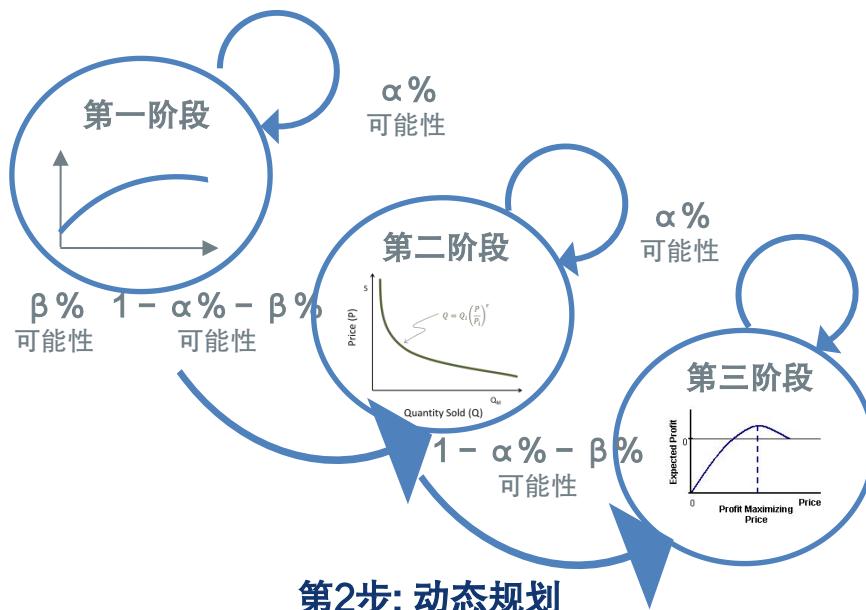
通过预测下星期的需求弹性曲线，对采购经理进行每周促销/降价指导



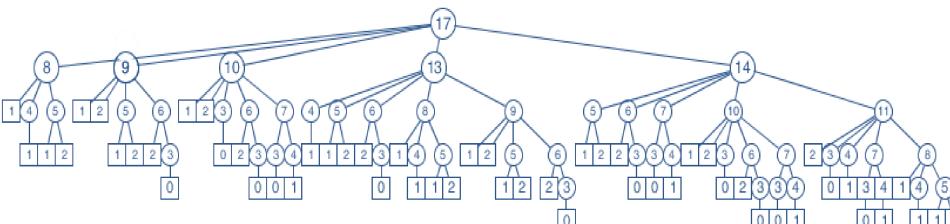
## 输入数据

-  历史销售数据
-  历史降价记录
-  产品元数据
-  库存数据
-  订单数据
-  每周销售目标
-  经济状况
-  积分卡/会员卡数据
-  客流量(例如商场客流量和关键词搜索量)

## 第1步: 马尔科夫决策过程

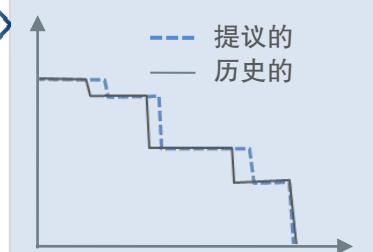
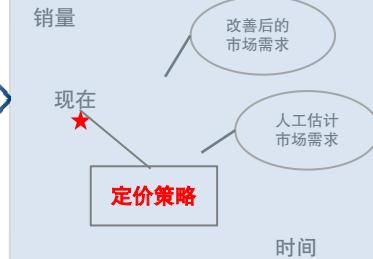


## 第2步: 动态规划



优化一系列活动 (降价) 使得一段时间内的每个商品的回报 ( 销售总额 ) 都最大

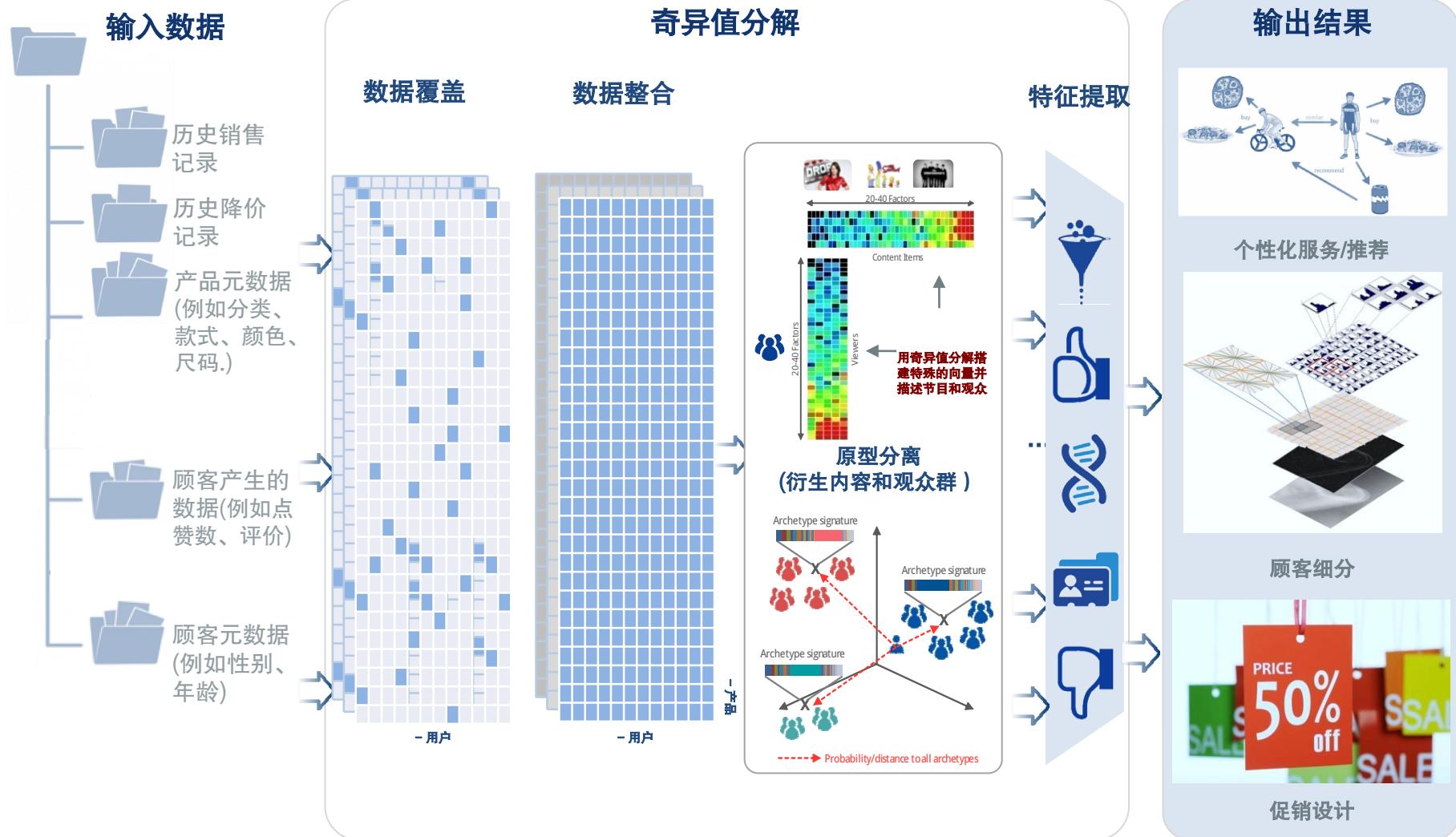
## 输出结果



# 推荐引擎



InsightBox的分布式推荐引擎能帮助线上的零售商开展实时的、针对性的营销





## 输入数据

会员卡/积分卡数据

顾客关系管理数据

外部人口特征数据

历史销售记录

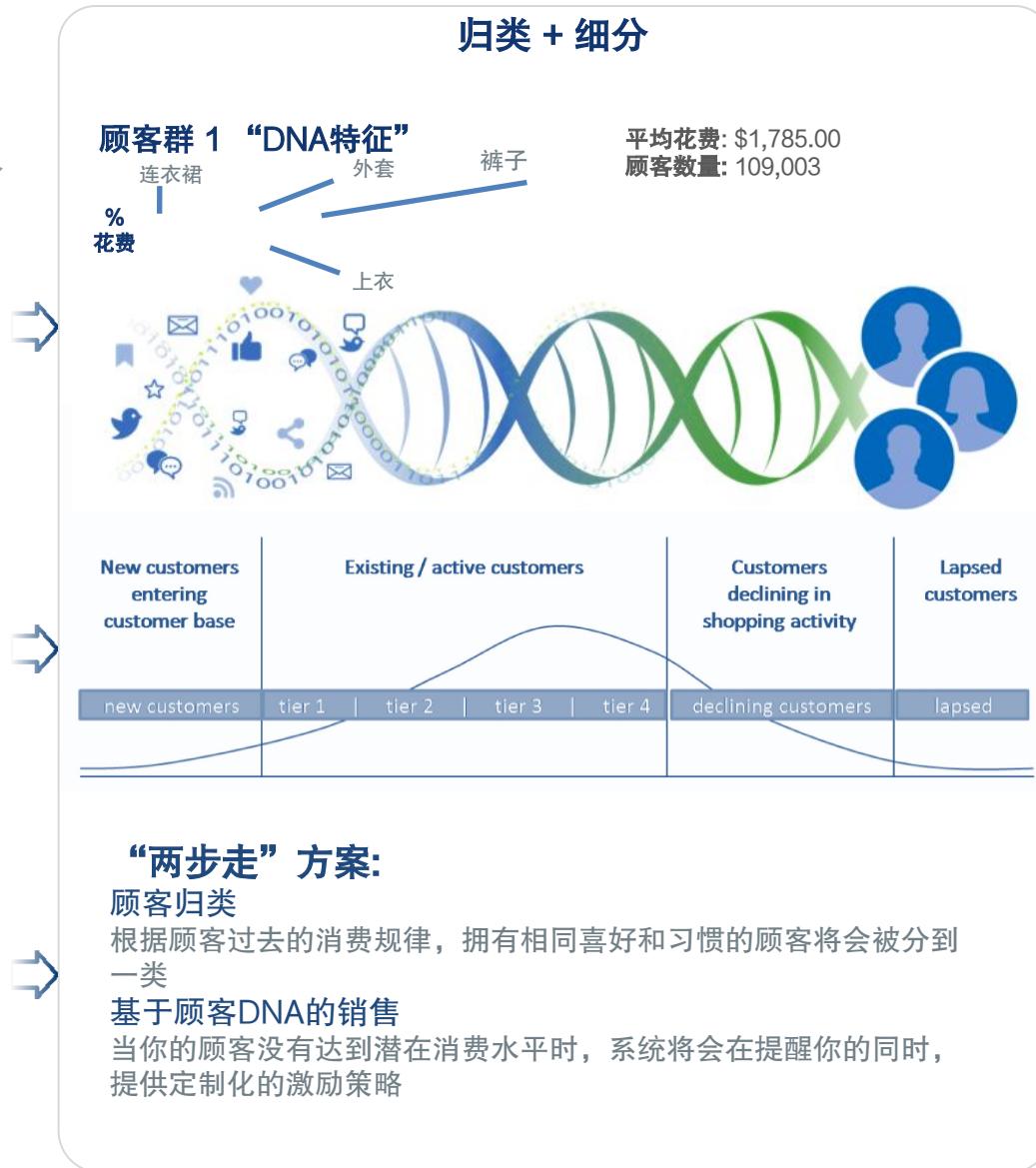
历史营销活动

产品元数据

经济状况

社交网络上的反馈

客流量 (例如商场客流量和关键词搜索量)



# 防止顾客流失



InsightBox的防止顾客流失的解决方案用离散存活分析等技术，在早期就识别出可能流失的顾客，并自动给出挽救措施



## 输入数据



历史销售记录



历史降价记录



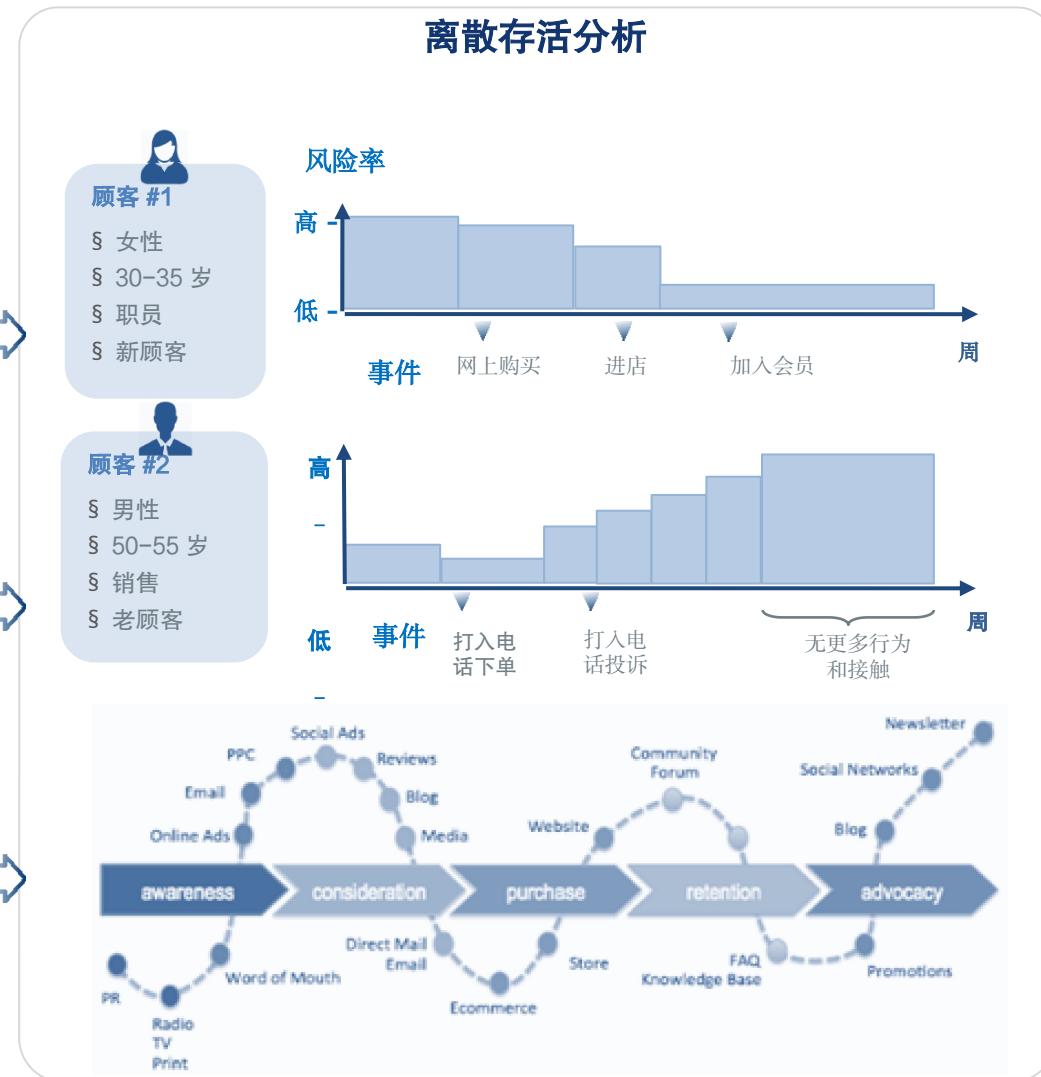
产品元数据(例如淘宝信息、外部数据)



顾客产生的数据(例如点赞数、评价)



顾客元数据(例如内部数据和外部数据.)





# 解决方案

我们的各种方案可根据不同的商业问题和企业情况被灵活搭配使用

	垂直行业	关键客户	机会	解决方案
1	专卖店	玩具连锁店	为定价提供一个系统全面的方法	商店分类，定价优化，测试监测和管理
2	高端零售	奢侈品店	通过深度了解顾客来提高销量和利润	顾客细分，推荐引擎
3	餐厅	多品牌休闲餐饮集团	通过提高服务和研究社交网络潮流来提高销售额	客服ROI，社交媒体分析
4	便利店/燃油	提供燃油和其他设施的加油站	通过对整个价值链进行先进算法分析来发现机会	客流量增加器，货品品类优化，免费商品，竞争分析
5	企业对企业	给企业餐饮等其他服务设施提供者	设计科学的客服和顾客交流的方式	推荐引擎，员工表现分析
6	顾客直销	杂货店零售和外卖服务	深度了解顾客从而在付款时给出个性化推荐	顾客细分，推荐引擎
7	售货目录零售	礼品收藏品直营者	用分析的方法使各渠道的销量增加	推荐引擎，促销和媒体优化

## 顾客细分



## 商店分类



## 商品品类优化



## 社交媒体分析



## 顾客忠诚度



## 定价优化



## 客流量增大器



## 促销与媒体宣传优化



## 推荐引擎



## 竞争情况分析

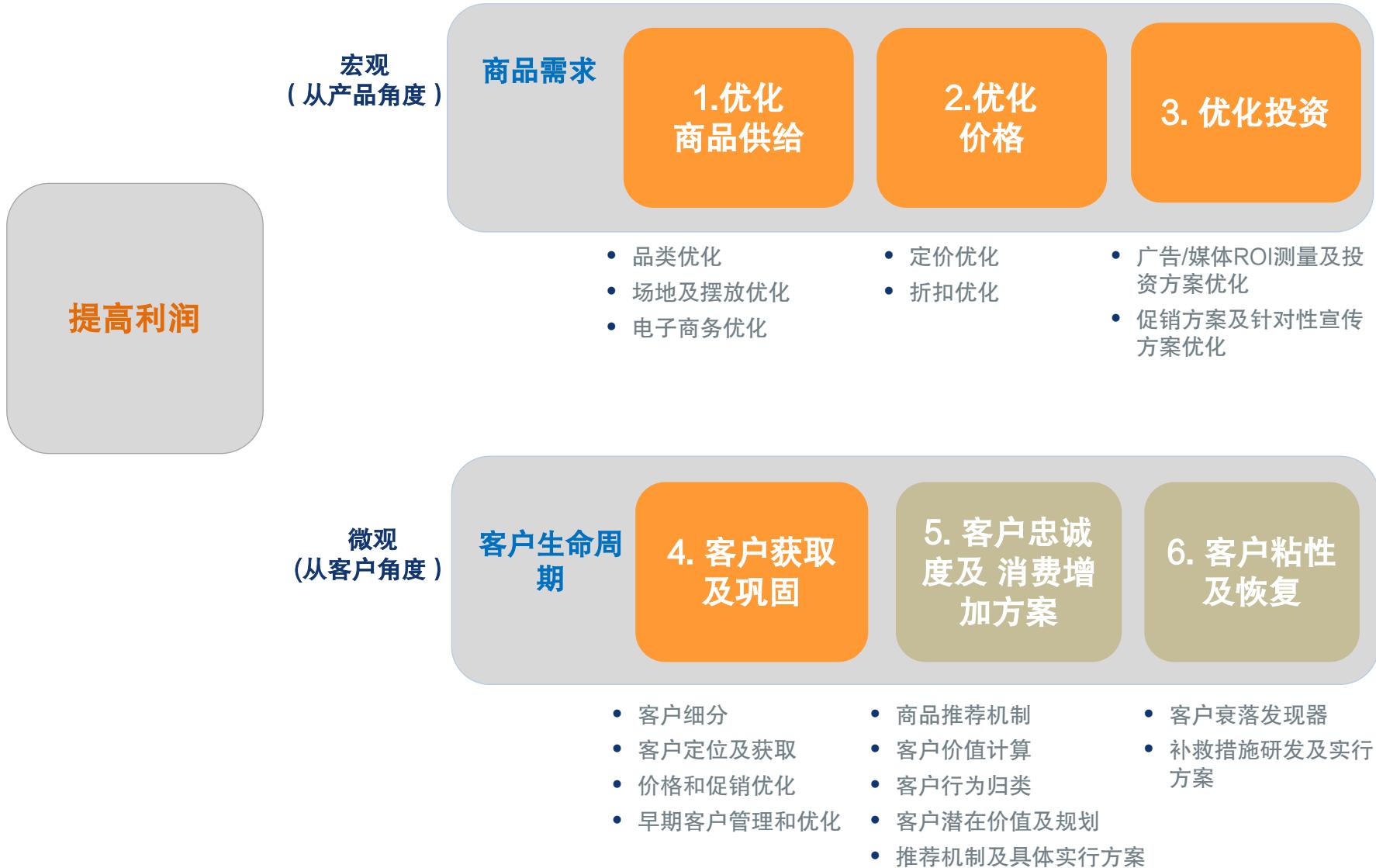


## 免费商品配套



## 测试监测和管理





## 营销战略

### 影响客户

#### 7. 品牌拥护及管理

#### 8. 协同营销

#### 9. 客户需求及分类

- 通过社交网络持和情绪分析续测量品牌的净推荐值以及品牌认知
- 精细测量市场竞争(从每个用户，地区)
- 分析锁定并使净推荐值增加的人事和项目
- 鉴别品牌拥护者，推行为者和影响者的行为
- 设计营销方案使其积极推动并宣传品牌
- 长期、实时、全面发现新生的客户需求和行为
- 发现新的客户类群及商品供应

### 深度解析

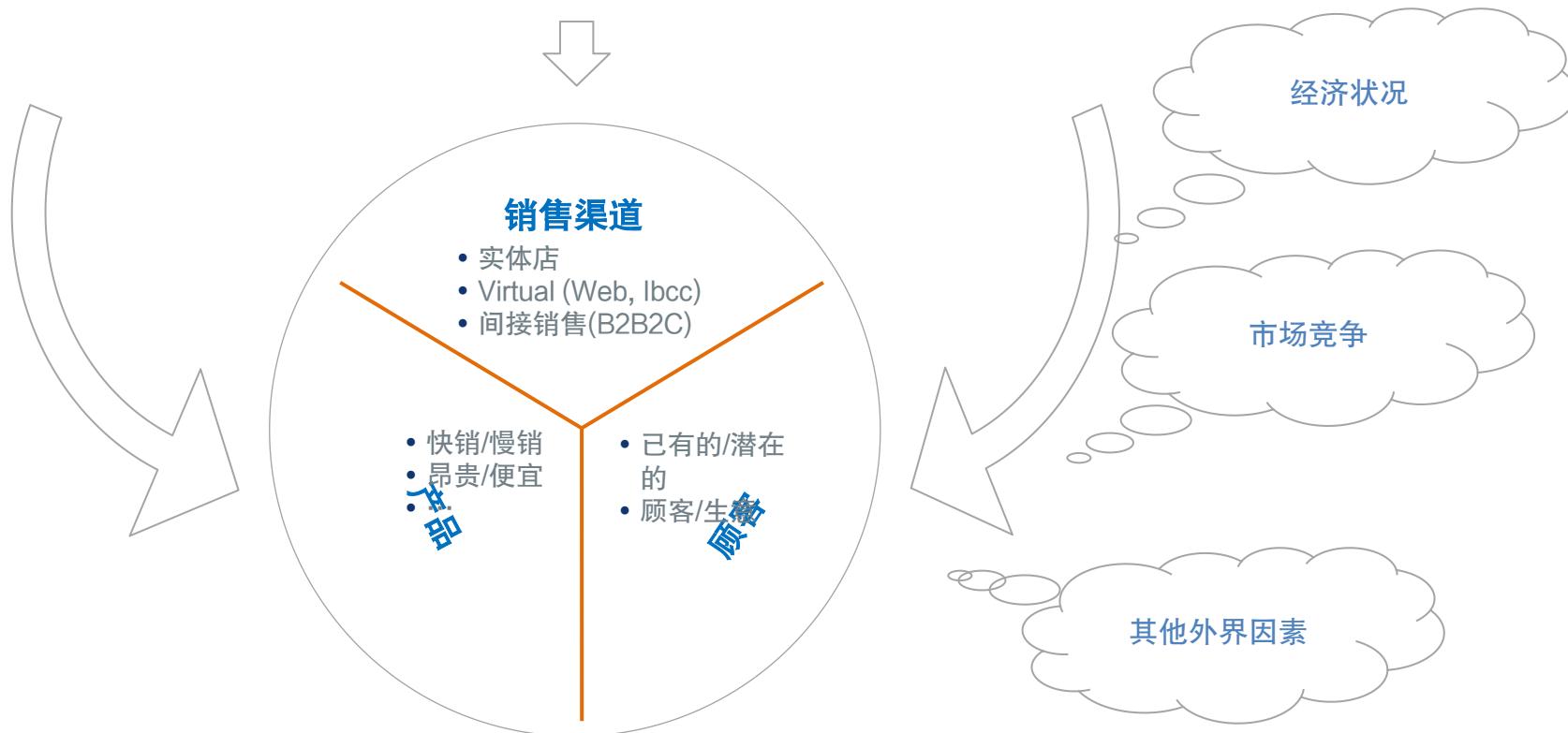
#### 10. 供需智慧

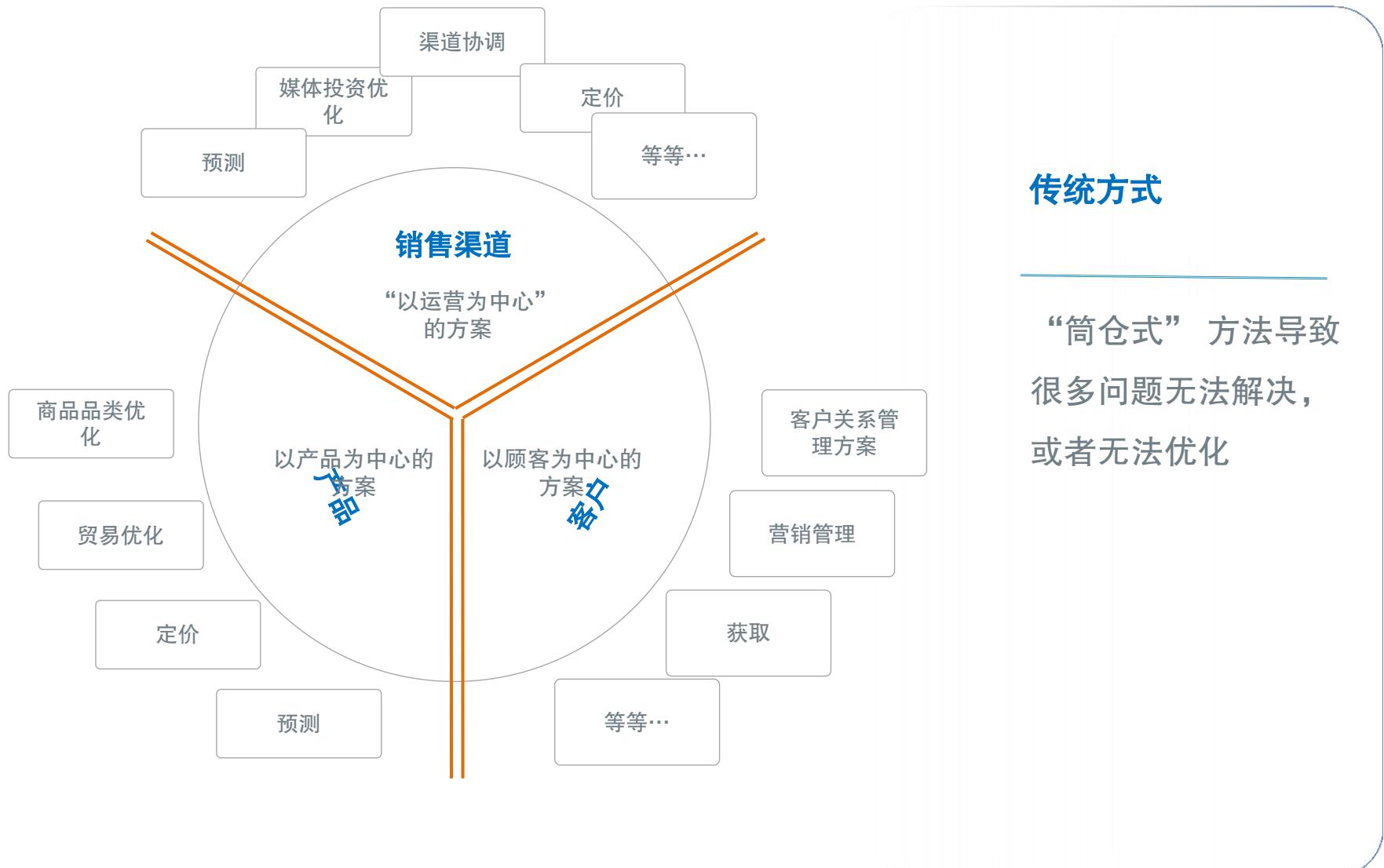
#### 11. 市场竞争及合作伙伴智慧

#### 12. 推动创新

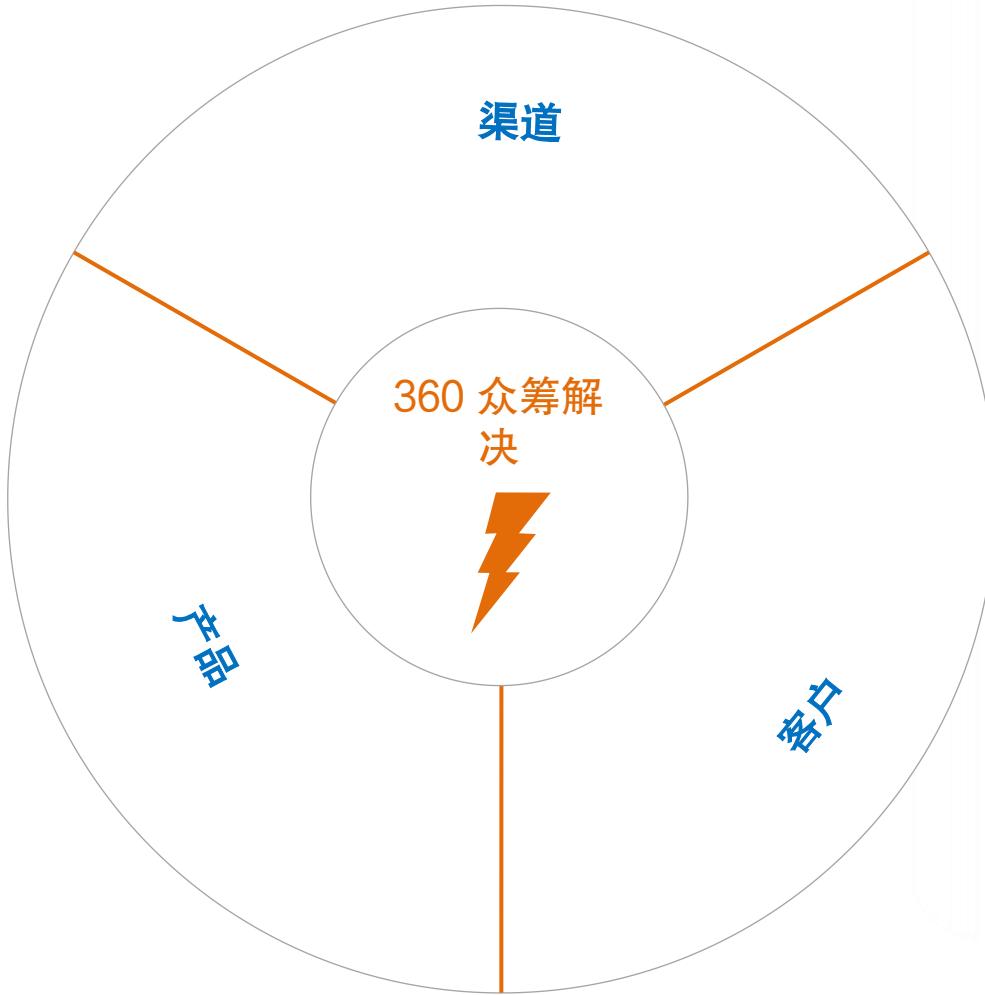
- 锁定及量化使消费者需求增加的因素和指示器
- 通过确立特定话题和职能范围持续锁定竞争者及合作伙伴的动态，倡议及表现
- 长期、实时、全面发现新生的客户需求
- 量化客户需求和趋势

## 需求杠杆





# 我们的优势



DATAESP

营销方案市场竞争力= 360  
CROWD CONNECT

“在不假设任何客户/产品/渠道观点的情况下，充分借助大数据和先进算法，来增加企业价值”

Every one of our solutions leverages “Wire frames” belonging to each sector

# 数据生态系统



收集的各种各样的内部和外部数据将会运用于各家餐厅；此数据生态系统能反映所有其商业系统中的主要问题

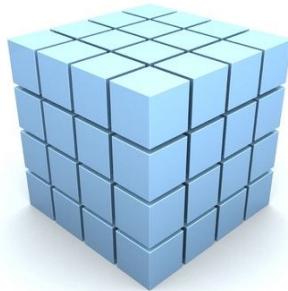


# 专有科技: The Insight Cube, 您商务上的“核磁共振”



我们的Power 和 Precision Insight cubes 使得经常隐藏在公司内部可被潜在节约的支出可视化

## Insight Cube



### The Power Cube

构建完整的企业内部花销  
– 使得企业内的花销透明度增加接近一倍，从而捕捉所有的机会



### The Precision Cube

提供极为精细，并且可向下细分的方案类别 – 整合外部数据和内部数据，在迄今为止仍然不透明的领域内捕捉机会

## Insight Cube 科技的主要特点

- § **完整:** 从各资源收集完整的花销数据：各种系统，应付账款，采购卡，差旅和娱乐等
- § **准确:** 在每条消费记录层面上进行分类，而非应付账款总合的层面
- § **方便使用:** 用户可以从个人电脑上通过视觉效果丰富的界面，简单地点击，来读取结果
- § **灵活:** 可被根据用户需求被动态细分的报告
- § **细致:** 可细分到发货单-发货层面
- § **可执行的:** 使得企业能够针对性地快速反应，执行战略并且快速完整地发现机会

Power 和 Precision Cubes 能被快速开发 – 一般在收到数据之后四周内即可完成。一旦开发好，因为内部算法能自动准确地给新数据分类，此科技可被反复使用。

# 图示: Precision Cube的力量

我们的Precision Cube能深度洞察迄今为止还不透明的类别。除了运用先进的科技，我们的采购专家还通晓如何设计优质的数据模块并获得整合最重要的数据。

## PRECISION CUBE

- 5 因为我们的科技能快速整合并结构化各式数据，所以我们能够非常快速地联合相关的外部数据，使得迄今为止仍然不透明的领域更通透（我们会例行获取客户的高度细分的花销记录）。
- 5 Precision Cubes 还能帮助我们快速给不明确的花费分类，例如广告代理费和临时工，使得我们之后能够对非单品类的花费采购进行完整地优化。未来的花费也将被准确自动地分类。

## 范例: 餐饮模块

供货商

Supplier 145,612,369 of 145,612,369		
	Spend (USD)	Count
USF	125,776,766	62,608
Sysco	19,835,603	9,316

分销商

Distributor Name 140,003,165 of 145,612,369		
	Spend (USD)	Count
UNASSIGNED	26,192,988	18,739
NULL	25,512,597	9,038
US FOODSERVICE	11,946,737	9,437
STARBUCKS COFFEE CO.	4,608,403	142
GEORGIA PACIFIC COMMERCIAL	3,421,293	208

标注

Label 126,280,974 of 145,612,369		
	Spend (USD)	Count
STARBUCKS	7,689,926	304
PACKER	5,369,946	3,818
GLNW FRMS	4,384,463	276
PATUXENT	4,240,874	1,012
ECOLAB	2,724,938	514
STARWOOD	2,282,791	24

装箱数

Case Pack 25,148,489 of 145,612,369		
	Spend (USD)	Count
4/1 GA	3,539,035	2,303
4/5 LB	3,293,196	940
15 LB	3,126,951	357
10 LB	3,012,441	2,140

计量单位

Measuring Unit 127,552,280 of 145,612,369		
	Spend (USD)	Count
LB	44,317,842	19,845
OZ	35,273,791	15,137
EA	14,692,810	11,855
GA	6,984,671	4,248
LBA	6,828,522	2,411

分销商品  
编号

Distributor SKU # 3,643,334 of 145,612,369		
	Spend (USD)	Count
7404882	1,037,439	2
Z59217	748,625	1
4404893	630,738	2
425140	625,868	1

花费类别

Revised Spend Category 145,612,369 of 145,612,369		
	Spend (USD)	Count
FOOD & BEVERAGE	128,212,805	59,771
EQUIPMENT & SUPPLIES	14,453,843	12,237
CHEMICALS	2,909,555	997
TBD - MISC.	36,166	181

单品描述

SKU Description 7,434,531 of 145,612,369		
	Spend (USD)	Count
TISSUE, TLT 2 PLY EMBS WHT	1,037,439	2
COFFEE, GRND HOUSE BLND FOIL	967,705	2
EGG, LIQ WHL W/ CTRC ACID	656,809	3
WATER, SPRG PLST	846,625	1
COFFEE, GRND DECAF HOUSE BLND	750,387	10
TISSUE, FACI ANGEL SOFT WHT	630,738	2
COFFEE, GRND AFRCN KITAMU FOIL	625,868	1
WATER, NTRL SPRG	600,865	1
COFFEE, GRND FOIL PK LTNT	585,027	3
SAUSAGE, LNK PORK RAW 2 Z FZN	533,268	1

平均价格  
(每箱)

Avg Price (Per Case) 25,986,899 of 145,612,369		
	Spend (USD)	Count
(blank)	19,835,603	9,316
\$74.40	1,006,641	6
\$97.65	859,515	8
\$27.15	750,778	11
\$41.88	629,386	4
\$27.51	617,828	14

运输量

Quantity Shipped 15,399,446 of 145,612,369		
	Spend (USD)	Count
1	6,111,925	12,948
2	2,956,812	7,121
6	2,236,653	3,154
4	2,112,101	3,839
12	1,981,955	1,815

花费类别

Spend Category 145,612,369 of 145,612,369		
	Spend (USD)	Count
BEVERAGE	18,988,112	2
POULTRY	10,065,361	2
DAIRY	9,158,703	1
BEEF	8,526,707	2
FOAM, FOIL, PAPER & PLASTIC	6,571,044	3

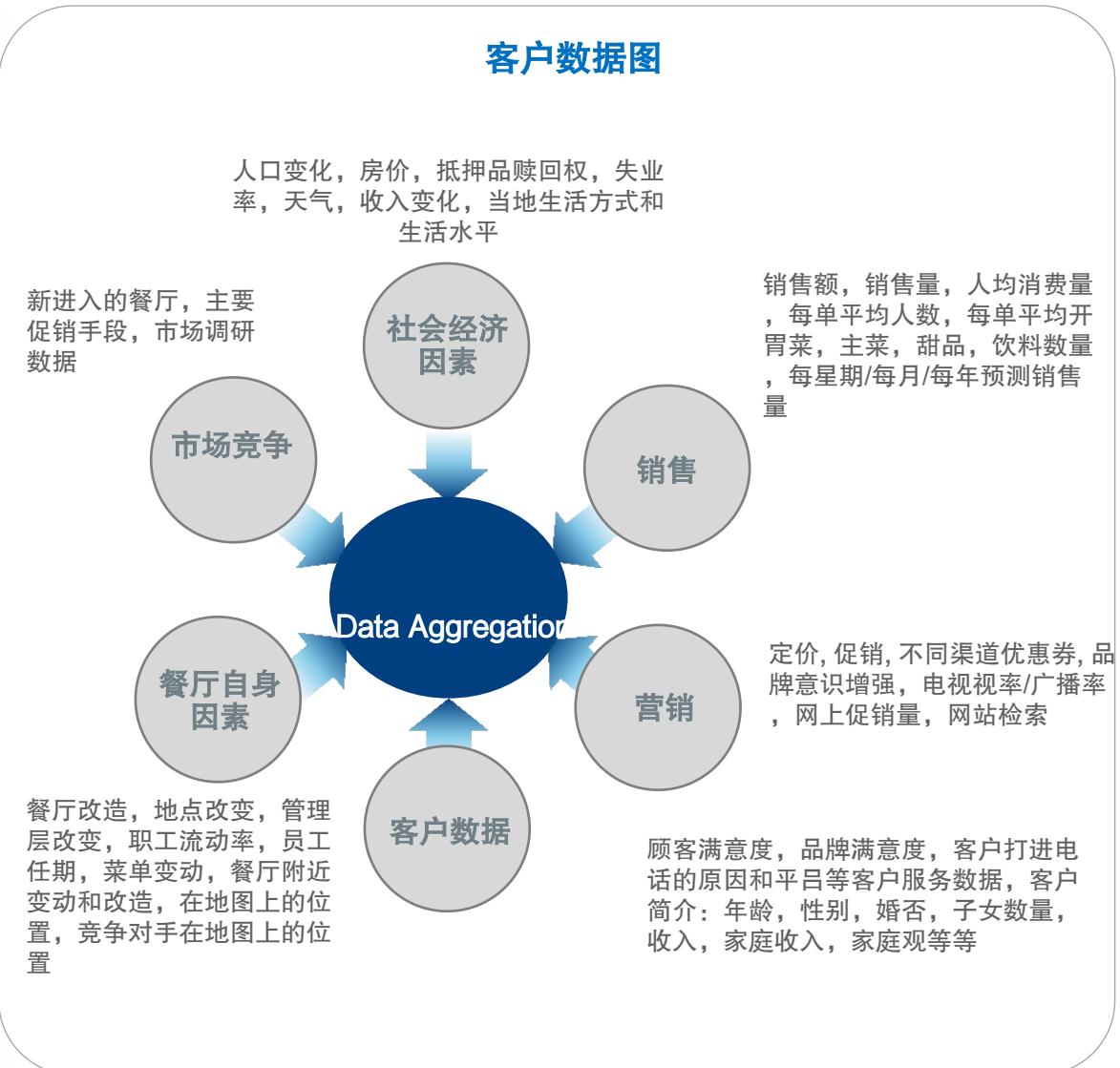
## 背景

- 虽然很多年表现优异，近期销售额却落后于竞争对手之后。
- 装修店面或传统降价营销手段均无效
- 需要找出最核心的客流量推动力并实现8%的年销售量增长

## 传统

- 复杂的数据整合
- 数据来源各式各样，尤其外部数据，数量大且增长快
- 不同应用领域的数据难以整合并很好地理解
- 从全国范围内搜集来的数据，如果需要在不同的地区被很好地应用，需要被很好地细分和解读
- 不同行业之间的需求没有被很好地理解。对于有限的数据还只能采取从上往下的分析方法，因此很多需求无法被完整地理解
- 对于一直在改变的需求推动力和市场状况需要非常快速地做出反应

## 客户数据图



## 变量确定

- 原始变量: 有2300个被锁定和测试过; 包括不同地区和地点, 价格变动, 广告支出的增长和下降等等
- 转化变量: 客户对自身所拥有的财富的感知(从每周标准普尔指数得来)等
- 使用的方法: 相关性分析, 决策树分析, 资料搜集分析, 神经网络, 奇异值分析, 变量聚类, 验证性模型, 矩阵因子分解

## 信号侦察

- 运用保留样本法来缩小变量选择的范围并验证通过模型所得到的结果、优先考虑关键需求
- 关注马上可变的变量, 抓住机会

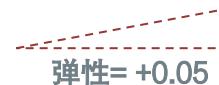
## 细分的洞见例子: 未提示知名度客流量的影响

未提示知名度

开始全国广告

%改变对客流量的影响

%未提示知名度的改变对客流量的影响



示例

## 从各个模型整合出的需求表

### SIGNAL HUB建立

- 建立让系统能够对市场环境快速反应的机制
- 搭建可扩展的并且灵活的结构和开放的环境，使得系统能够持续地改善
- 持续提供监测，智能提高等方面的支持，并帮助增加新的维度

Domain	Driver	Key Variables / Insights	Size of Impact	Direction of Impact
Marketing	Pricing	Pricing has a strong negative impact on traffic., especially in areas with higher competitive intensity	Large	⬇️
	Advertising	TV Ad Stock investment is ROI positive. Ad impact is stronger in newer markets than older markets	Medium	⬆️
	Promotions	Current set of promotions are not very effective in driving incremental traffic	Small	➡️
Structural	Competition	Competitive presence in the trade area negatively impacts performance	Medium	⬇️
	Demographics	Percent of children under the age of twelve in trade area increases performance	Small	➡️
	Economics	Unemployment is a significant driver of performance and recent trends have helped performance	Medium	⬇️
Customer Experience	Store Layout	Larger format stores provide a 30% lift.	Large	⬆️
	Service	Friendliness, Greeter positively impact traffic	Small	⬆️
	Variety	Stock-outs have a negative impact on traffic	Medium	⬇️
Staff	Labor	Employee Engagement positively impacts traffic	Medium	⬆️
	Turnover	Turnover has a small impact on performance	Small	➡️
Capital Investments	Remodeling	New branding updates are not ROI positive	Small	⬆️
	New Locations	High variation in new store performance observed	Large	⬇️

# 案例分析1 - 连锁餐厅Signal Hub (4 of 4): 财政影响和控制面板



此分析发现了价值等同于三千万美元息税前利润的机会，并且提供了一个能够监测到市场趋势和机会的互动控制面板

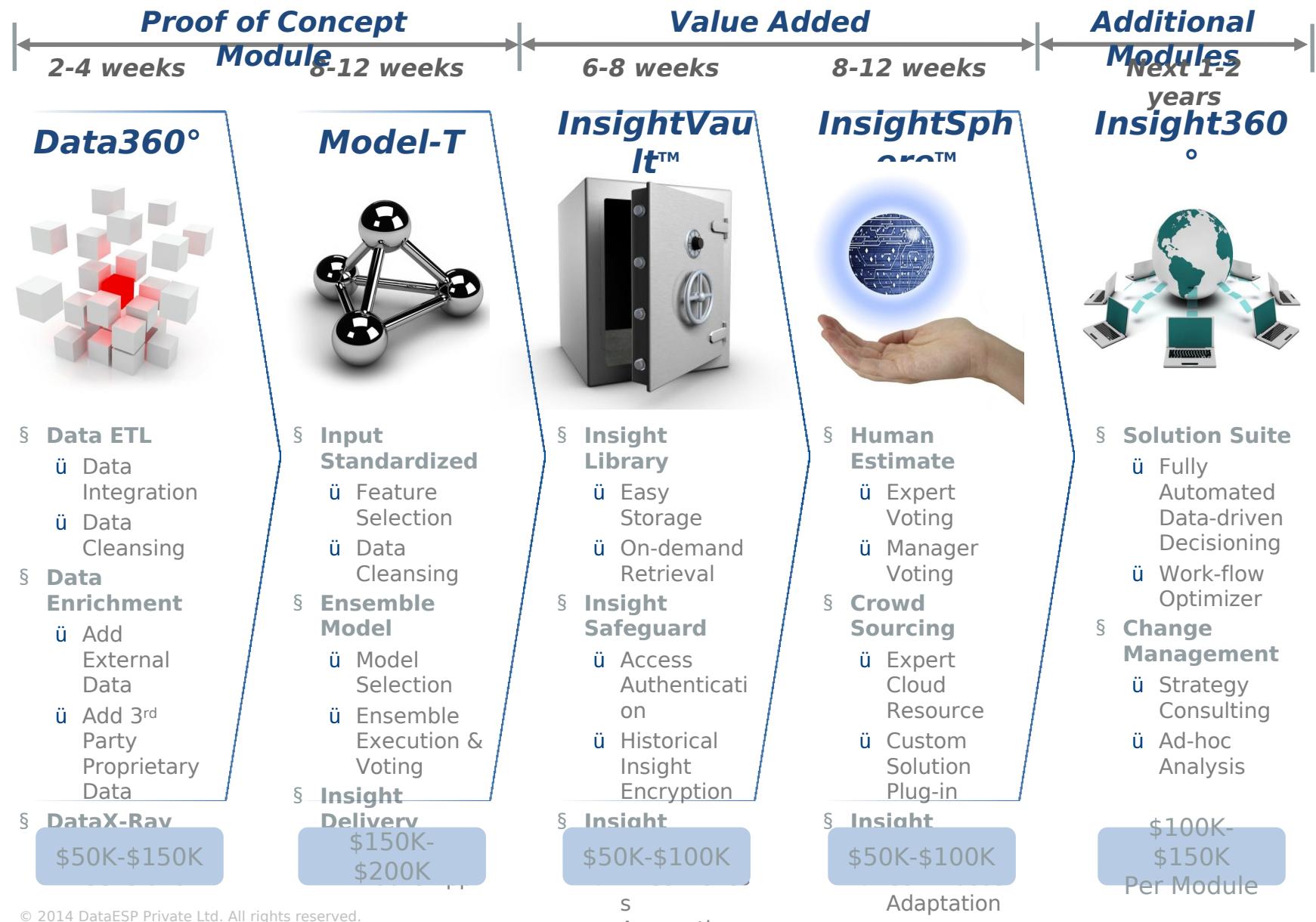
The screenshot displays a complex business intelligence dashboard titled "Pricing > Tier5". The interface includes:

- Performance by Driver:** A bar chart showing the impact of various drivers on same-store sales. Key data points include:
  - Pricing: -3.77%
  - Show All: +1.31%
  - Tier 1: +0.18%
  - Tier 3: -0.81%
  - Tier 4: -2.24%
  - Tier 5: -2.63%
- Performance Over Time:** A dual-axis chart showing Sales Per Restaurant (blue bars) and Avg. Entree Price (orange line) from March 2010 to November 2011, with projections through November 2012.
- ALERTS:** A sidebar listing several alerts:
  - Underperforming Promotions (Promotion A, Promotion B)
  - New Competitors
  - Alert 3, Alert 4, Alert 5, Alert 6
- MY RESTAURANTS:** A heatmap and data table view showing restaurant locations across North America, with a specific callout for Oregon.
- Unemployment Heat map:** A map of the United States and Canada showing unemployment rates by state/province.

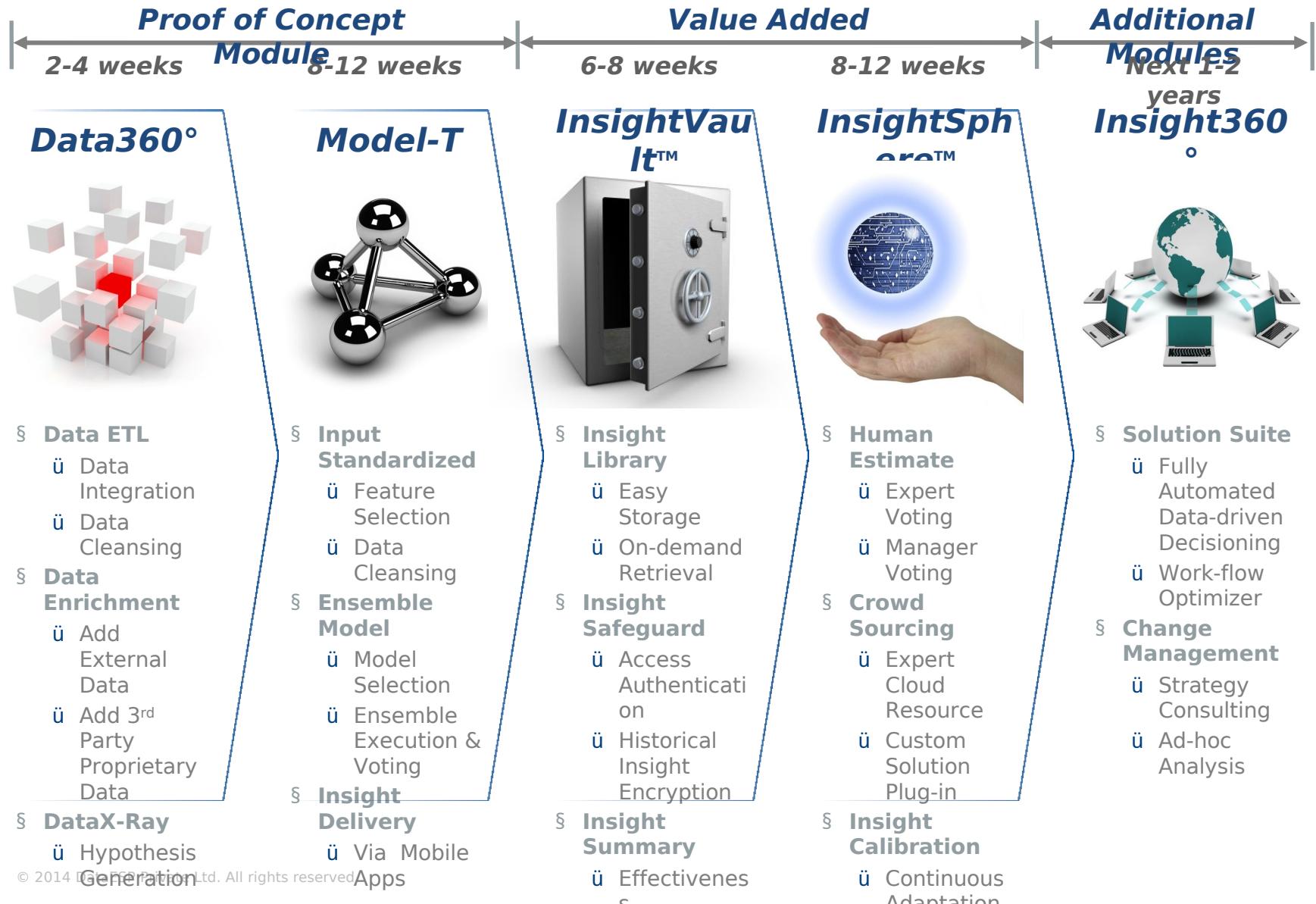
Three callout boxes highlight features:

- 左侧气泡:** 推动力分析能够找出业务范围内的绩效以及财政影。可以选取这个面板上的参数来放大进行查看。
- 中间气泡:** 从长期的绩效来看，我们提供的需求驱动力指标和绩效之间有明显的线性关系。
- 右侧气泡:** 对于表现不好的领域，会有警示，并会给予由数据分析得出的参考建议，使得业务效能得到提高。
- 底部气泡:** 用户可以向下展开并分析各个地区，甚至各个餐厅的绩效，并同时看到多个推动指标的状况。

# Roadmap for Adoption of InsightBox Analytics



# Roadmap for Adoption of InsightBox Analytics





# 案例分析

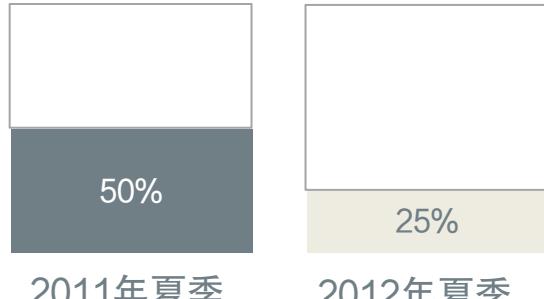


# 市场需求预测

# 满足动态市场需求的快速补货方案

需求预测方案可以帮助提高盈利, 增加市场占有率, 推行更好的库存管理, 并降低运营风险

## 畅销产品缺货程度



2011 销售/净利润损失: \$1.35亿  
/9.5千万美元

2012 销售/净利润额外增长: \$6千万  
/4千万美元

## 库存指标

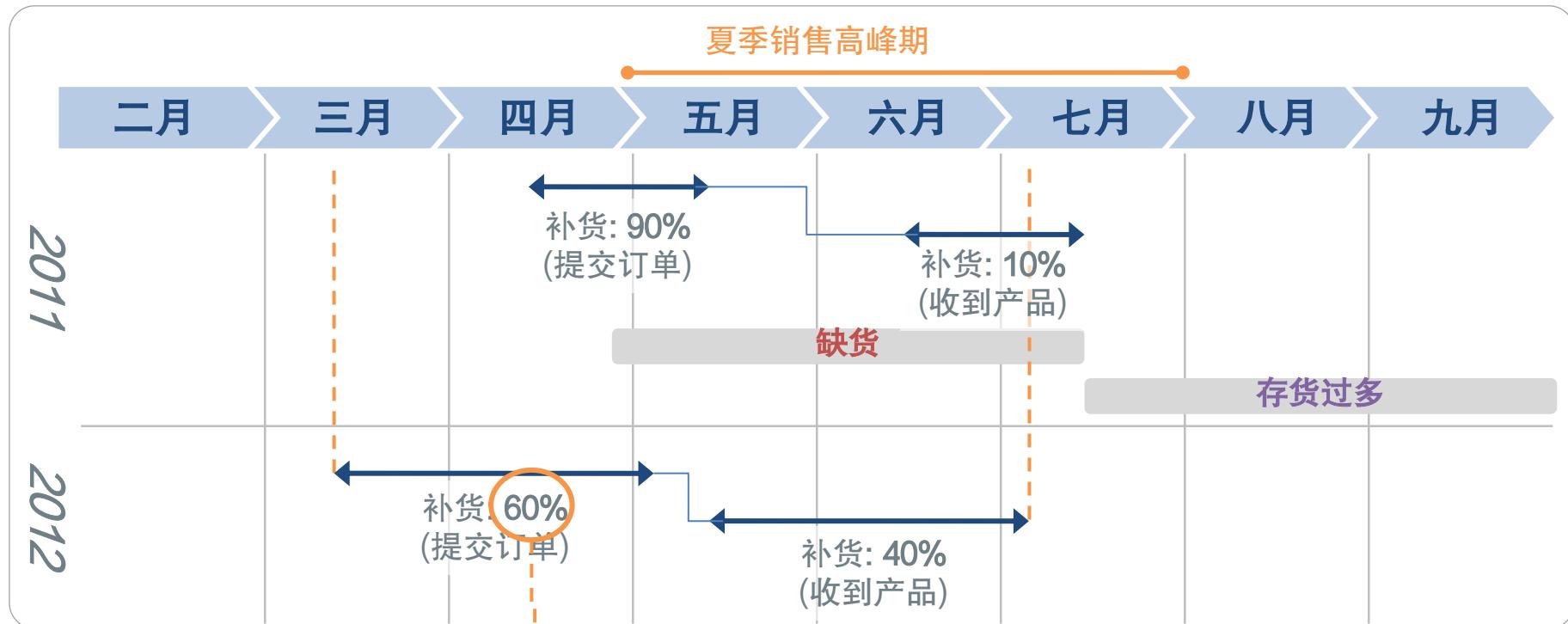
2011 – 28%存货积压; 大部分为滞销商品 – 利润损失: \$2.5千万美元

2012 – 存货降低20% – 增长利润: \$1.8千万美元

# 快速补货模型



对动态市场快速做出反应的能力是成功建立快速补货模型的基础。通过对SKU历史销售数据精准分析和对季度销售业绩的持续追踪，您将能很快应用我们的新模型



降低初始订单以减少存货过多的风险

- 提前订货/补货的计划能及时对市场需求做出反应，避免畅销产品过早脱销

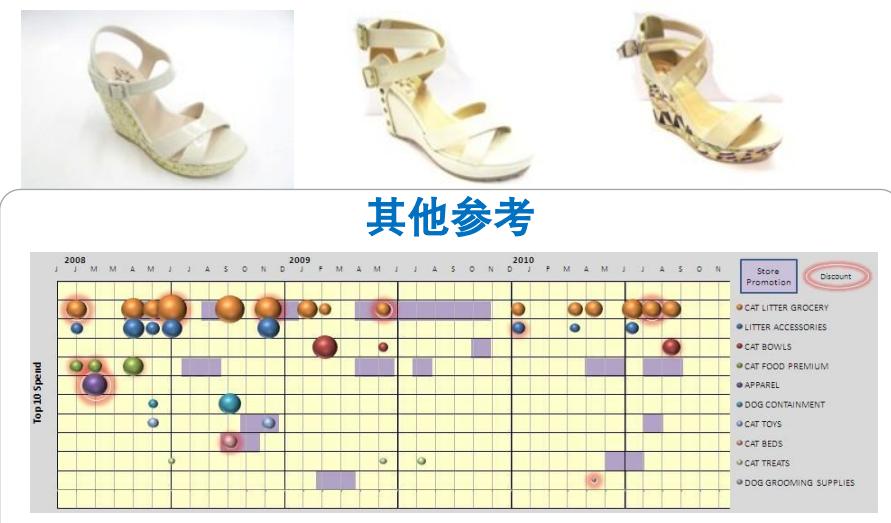
# 先进算法驱动的需求预测方案

通过分析历史销售数据，定义出5种不同的需求曲线。对于每个新产品，通过KNN邻算法可以找出有用信息最多的模型和相关的SKU参数

## 1 分析历史销售数据并定义主要的需求曲线种类

### 五种不同的需求曲线

## 2 把新的模型与最合适的需求类型和相关参数关联

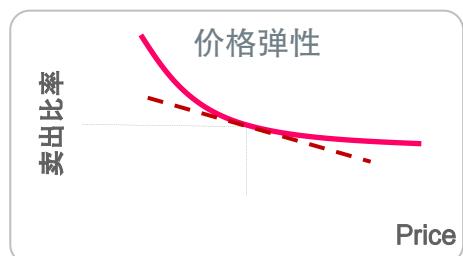


## 先进算法驱动的需求预测 (接上页)

# 3

### 需求的价格敏感性:

- 折扣对需求影响的模型化
- 平均价格弹性为：5折销售将提高销量4倍

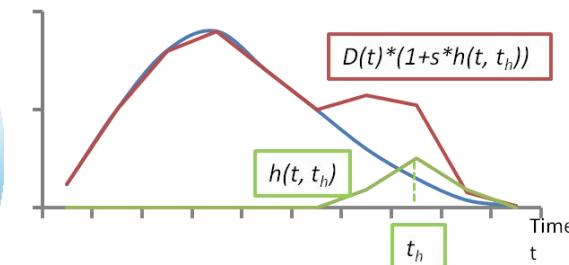


$$SR \quad \left( \frac{p_0}{p} \right)$$

# 4

### 节假日效应: 节假日需求高峰

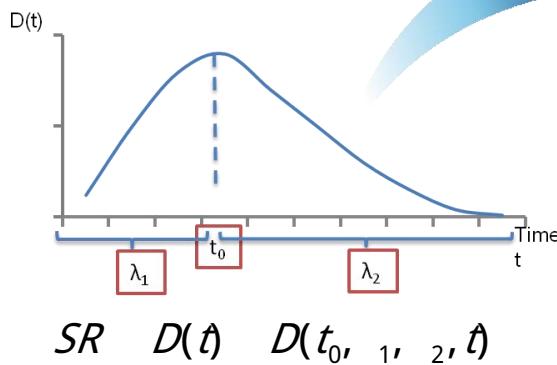
- 五一劳动节: 高于平日80%
- 母亲节: 高于平日80%
- 端午节: 高于平日25%



# 5

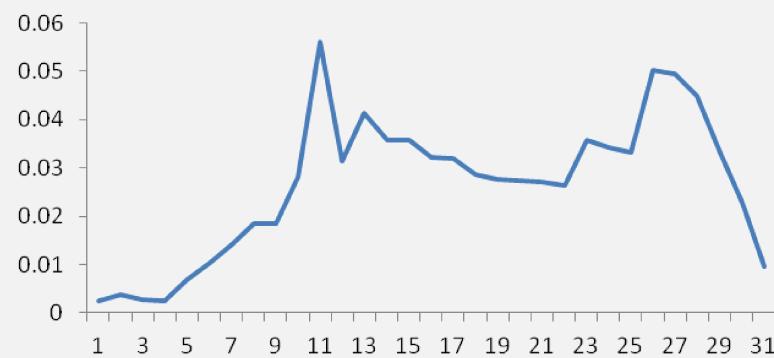
### 其他因素:

- 流行趋势
- 气候
- 等等.



# 6

### 预测的销量



# “大数据” 优化定价



当销售季节开始后，用相关的工具每周对销售跟踪和对定价优化，以达到毛利润的最大化和/或寂寞库存最小化

## 销售& 需求跟踪(每周)

销售额

现在

定价策略?

修正后的市场需求

最初预测的市场需求

Time



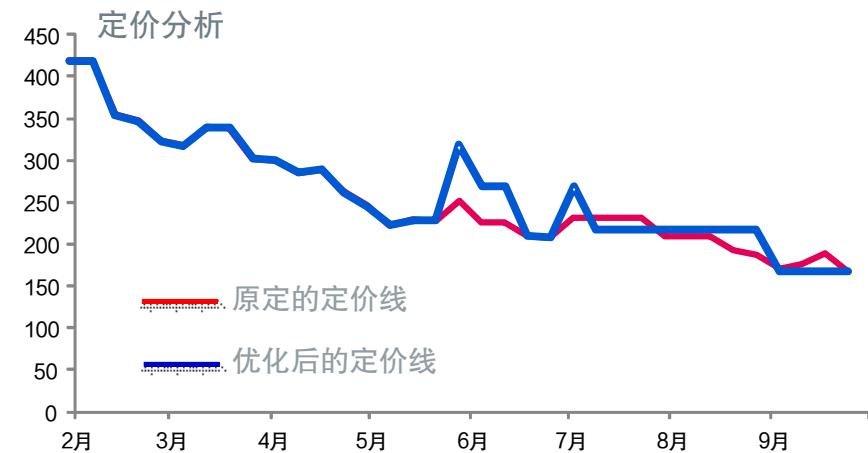
123019715

	需求	存货
最初	155,000	175,000
修正后	185,000	175,000

通过跟踪销售，需求和存货，这个产品被认为有断货脱销的危险

## 优化定价

定价分析



定价线	销售 (双)	毛利润	存货 (双)
原有	175,000	29,480	0
优化后	175,000	39,529	0

定价优化工具建议提升价格以增大毛利润

# 需求预测



采购组准确地把握未来需求——更早、更快、更准地找出未来畅销产品，并使得客户得以对需求及定价定出灵活的策略

## “大数据”提高了市场能见度 – 需求预测

夏季度 2011 vs. 2012

2011

2012

季度前  
(基于历史数据)

- 找出了**80%** 的畅销款
- 找出了**9** 款误判的畅销款
- 需求预测的准确率为 **50%**

- 找出了**90%** 的畅销款
- 没有被误判的畅销款
- 需求预测的准确率为 **65%**

季度中  
(基于实际销量)

- 畅销款在5月被确认（劳动节后）
- 劳动节后，需求预测的准确率为**80%**

- 畅销款在3月中旬被确认；此时需求预测的准确率为 **80%**
- 劳动节后，需求预测的准确率为 **90%**

# 减少断货和脱销



## 提高利润的机会

利用先进的分析算法迅速检测并减少断货的情况，从而提高利润

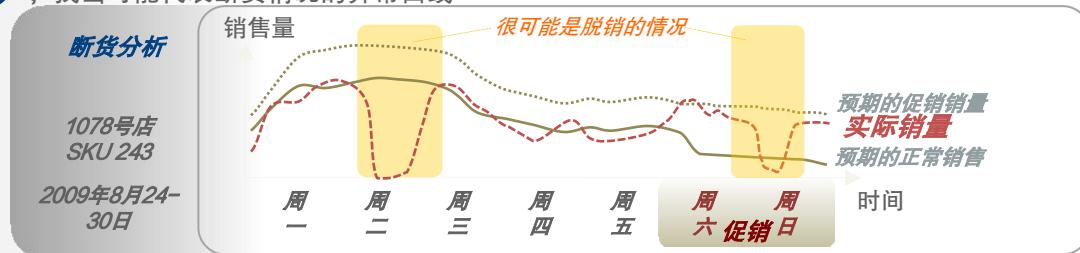
§ 国际范维内的零售行业的平均断货率约为 8%

## 顾客对断货的反应



1

利用先进的分析算法为每个店的每个SKU建立每天的销售量底线，并与实际销量对比，找出可能代表断货情况的异常曲线



2

分析根本原因，理解断货现象



### 根本原因分布

上游部门原因  
货物在后台，并未上架

订货错误、无法对需求做出准确预测

全球零售商

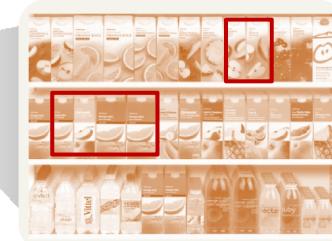
3

研究并寻找可行的建议来减少未来的断货情况

#### 例子

根本原因	负责人	推荐采取的行动
不足的货架空间	总部	<ul style="list-style-type: none"> <li>§ 优化店铺平面结构</li> <li>§ 检查商品品类</li> </ul>
订单效率低	店面	<ul style="list-style-type: none"> <li>§ 调整每个商品的订单数量和流程</li> </ul>
促销计划与店铺的订货情况冲突	总部	<ul style="list-style-type: none"> <li>§ 改进与商店之间的交流</li> <li>§ 在促销期调整商店货架</li> </ul>

### 平面图样例



# 其他采购和定价策略的有用发现

此“大数据”方案还发现了其他一些可以进一步提高客户采购和定价策略的方法

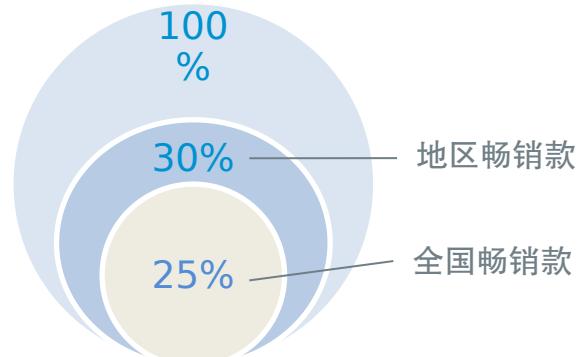
## 1

### 畅销产品的重要性

前30%的畅销产品创造了65%的应收，即90%的利润

前25% 的畅销产品均在全国范围内畅销，剩下5% 为地区范围内的畅销产品

#### 畅销产品



## 2

### 不同地区不同定价

不同地区往往会有不同的价格敏感度

在不同地域采取不同定价的方案可以提高销量和利润

#### 价格敏感度

## 3

### 节假日拉大销量

节假日的影响可被量化，并被用来更大程度地提高销售额：

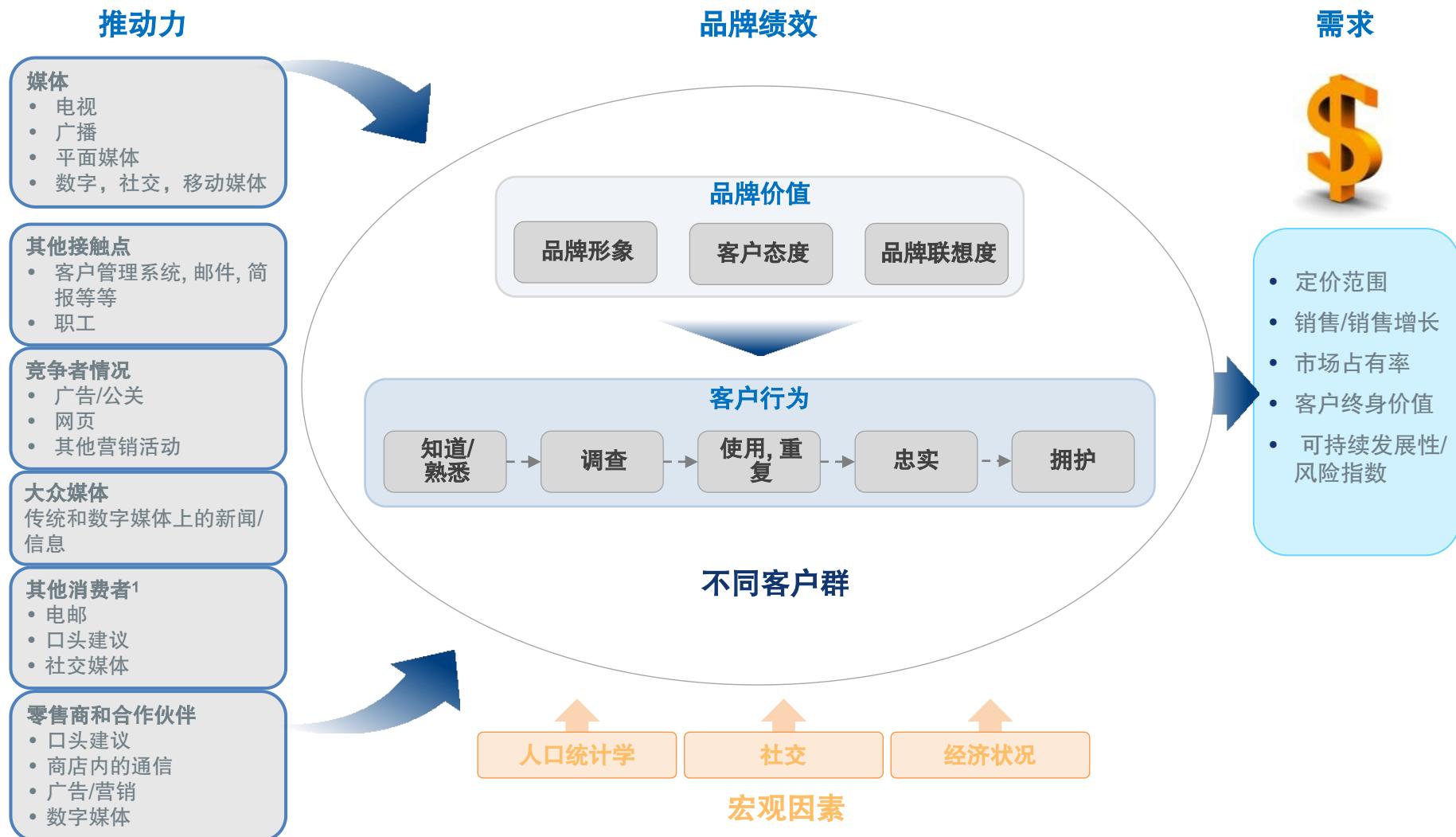
五一劳动节: +80%  
母亲节: +80%  
端午节: +25%

#### 需求曲线



# 确定可能的需求推动力

我们先找到所有可能影响整个需求生态系统的因素，再用先进的分析方法来测量其影响力



<sup>1) 朋友, 博主, 社交媒体用户等</sup>

# 关联客户业绩与需求增长点



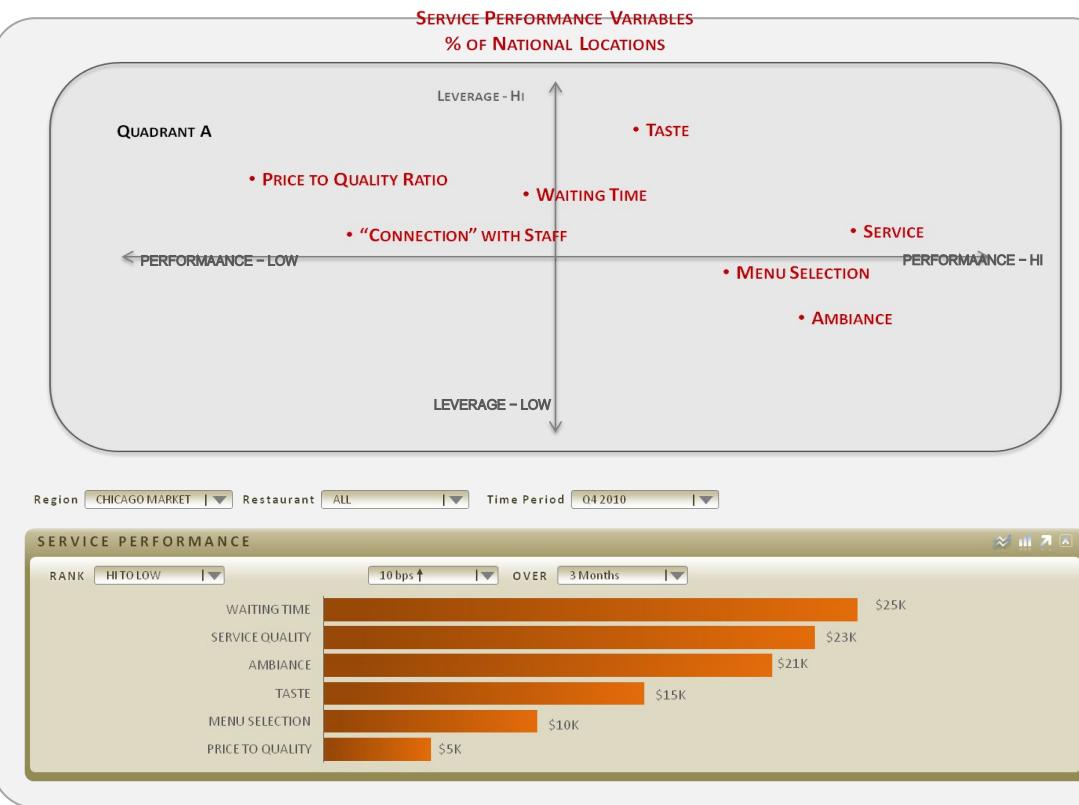
## 挑战

- 主要的连锁餐厅需要快速的，准确的并且能够迅速执行的，关于需求推动力的洞见，从而能够实现尝试了多种方法却难以达到的9%的年增长速度
- 需要容易使用的集成控制板来对需求驱动力和市场状况做出迅速反应；还需具备开放的结构使得能够对其进行持续的改良

## 方法

- 结合机器/人工智能来监测来自餐饮业生态系统的信号，使其能够被用来提高客户业务绩效：
- 有结构地集合所有相关的可获得的数据
- 同时由下而上及由上而下地确立并给关键测量指标分类
- 应用先进的描述性和指令性的分析来衡量绩效并给出针对性建议
- 在系统的监测，评估和改良等方面提供长期支持

## 分析图式



## 影响

- 从复杂的信号中把对客户销售额增长最有益的因素确立出来：改变定价策略，增加广告投入，加大员工培训和增强顾客忠诚度
- 提供战略决策上的操作版，从而监测市场动向和机会
- 其价值达到三千万美元（息税前收入）

# 人力+机器智能的方法

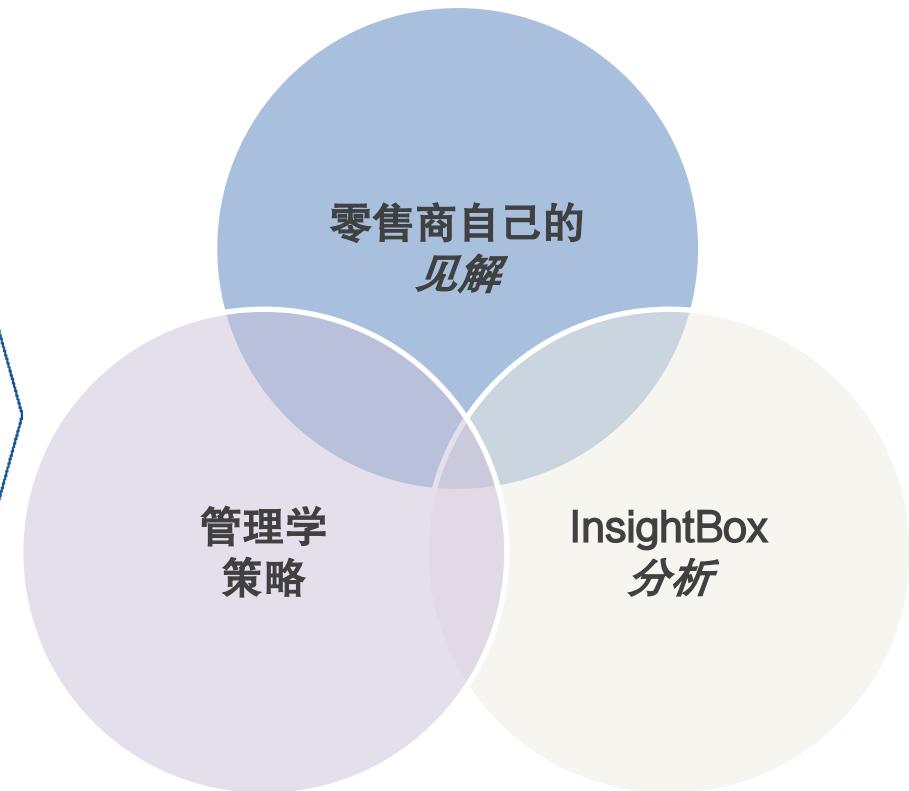


通过结合零售商的见解和InsightBox的机器学习的方法，公司对欲求的预测的准确性能高达90%，从而促使销售和利润进一步的提高

## 大数据带来的机会

仅凭**人类**是**无法**找到上述规律的 – 这需要电脑快速地存储，读取，处理并检测大量高维度的数据

**机器**不能自发地懂得这些规律；它们必须由人来“教育”。因此，机器学习方面的专业知识是充分激发“大数据”巨大能量的核心要求





# 消费者 360°

# 再创Netflix的成功

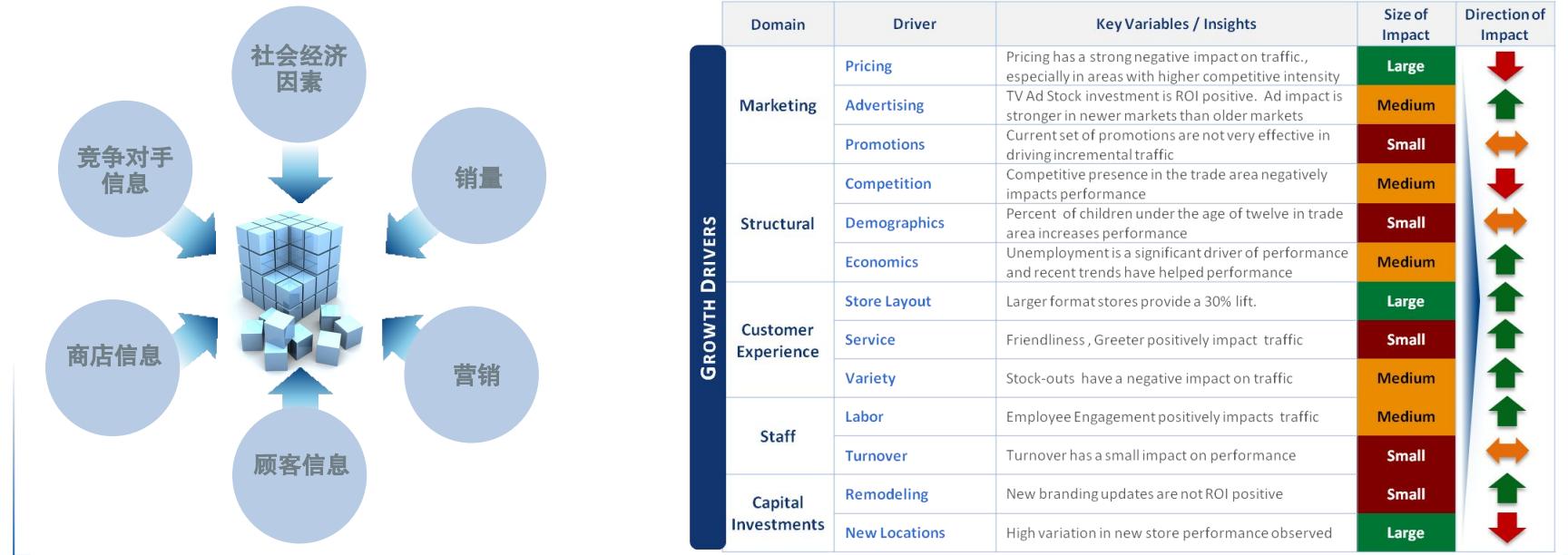
通过InsightBox's 客户360, 你能够马上准确地定位潜在客户, 做到给客户分类并且根据不同人对不同产品做出的潜在反应来给每个客户的价值打分



# 顾客360° 纵视图



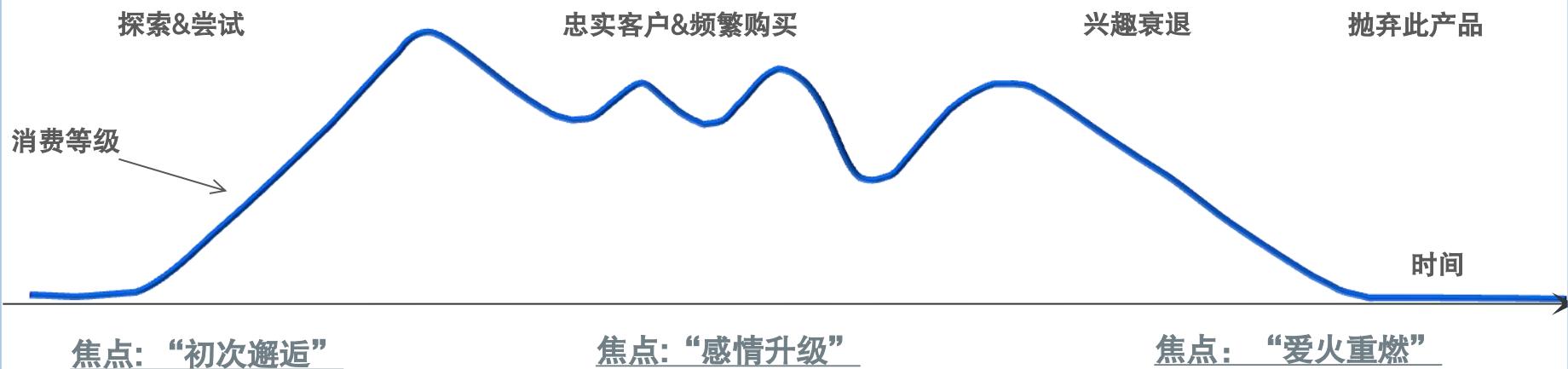
我们第一步是整合来自内部和外部的各式有用的信息，提供给您顾客的全方位视图，再在此基础上建模



# 最大化客户生命周期的价值

我们提出的促进利润增长的方案能帮助零售商在客户生命周期的每一个点来最大化与客户的关系，从而促进顾客的形成购买习惯并且进一步的增加利润

## 我们在顾客生命周期中不同时期的侧重点



### 焦点：“初次邂逅”

定义购买倾向高的人群

发现顾客的购买动机，然后适当修改产品，是它更吸引顾客

§ 奖励和巩固客户对产品的持续关注

### 焦点：“感情升级”

§ 判断顾客最大的购买力

§ 判断顾客的未来购买，购买率和价格敏感性

§ 创建并部署一个个人化的销售循环来提高顾客的参与性和利润

### 焦点：“爱火重燃”

§ 判断标志一个SKU/分类和子类购买率下降的关键信号（比如说，陡峭的下降曲线一般意味着顾客的兴趣衰退/抛弃此产品）

§ 采取定制的，有针对性的手段来预先组织兴趣减退或重新吸引已经不感兴趣的顾客

# “倾听” 顾客的数据生态系统

我们第一步是整合来自内部和外部的各式有用的信息，提供给您顾客的全方位视图，再在此基础上建模

## 内部数据

我们第一步做的是从POS、优惠券兑换、网站访问情况、客服和手机应用使用情况等内部数据中形成成百上千的规律



形成顾客最初DNA...



...然后与外部数据进行匹配

## 外部数据

将宏观经济状况、人口特征、天气、客流量、社交网络等外部数据巧妙地注入内部数据中，使其内容更为意义丰富



# 各式各样的数据来源

我们会将各式各样来源的数据整合，包括结构化数据和非结构化数据，来呈现每个客户的完整视图

## 交易及客户资料 数据

- 购买历史
- 人口统计特点

## 会议数据

- 出席率
- 时间地点日期
- 决策

## 网上足迹

浏览次数，浏览时间…

- 网页个性化
- 热门分类
- 市场
- 下载的菜单

## 社交网络足迹

- WW 社区活动
- 情感
  - 脸书Like / 链接数据

## 活动数据

- 体重
- 目标
- 输入的食物
- 点数 / 进度
- 锻炼情况

## 推广营销

- 电邮宣传
- 邮件直投营销
- 拨出电话

## 客服中心数据

- 服务事件
- 问题及查询

## 移动足迹

- m.ww.com: 浏览次数，浏览时间…
- 手机/iPad应用下载

## 数据生态系统

我们希望帮助您巧妙地使用数据流中所有的规律，从而在解决因为数据不断增长而带来的挑战

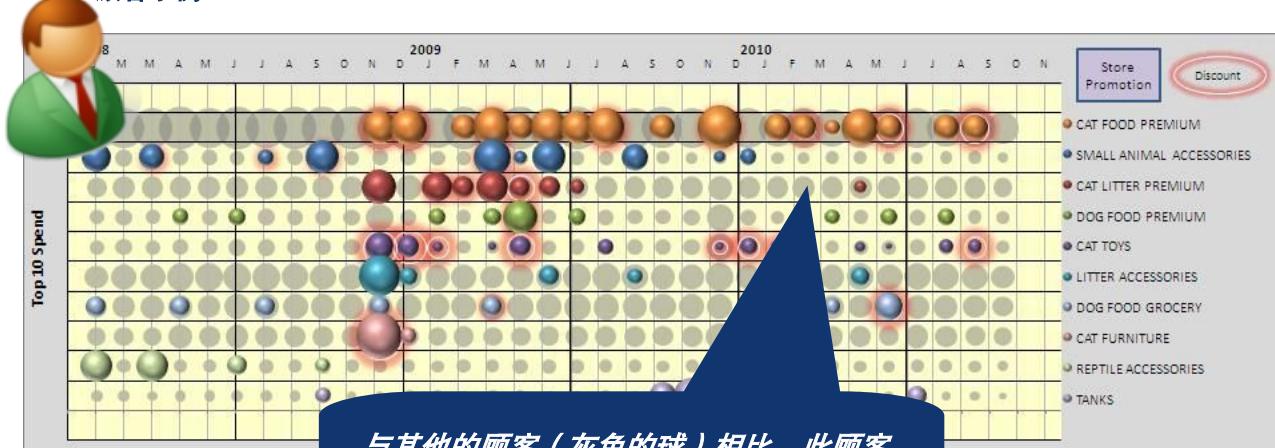
# 创建顾客的DNA

通过了解顾客过去的消费习惯等因素，我们可以把顾客分组来学习他们的消费模式，从而找到刺激消费欲望的方法，增强与顾客的交流与互动，减少顾客流失

## 第19组的“签名”



## 顾客示例



与其他的顾客（灰色的球）相比，此顾客（上色的球）购买了优质猫砂。

潜在消费能力

消费者现在的花费

# 高级聚类



我们使用客户，店铺，产品和供应商的DNA来准确的分类顾客，同一组顾客将采取相同的对待方法



## 机器学习技术

- SVD – 奇异值分解
- ANN -- 自动压缩神经网络
- K最近邻算法
- K聚类算法
- 混合高斯模型和最大期望算法



72个顾客群



24个商店群



30个产品群



191个供货商群

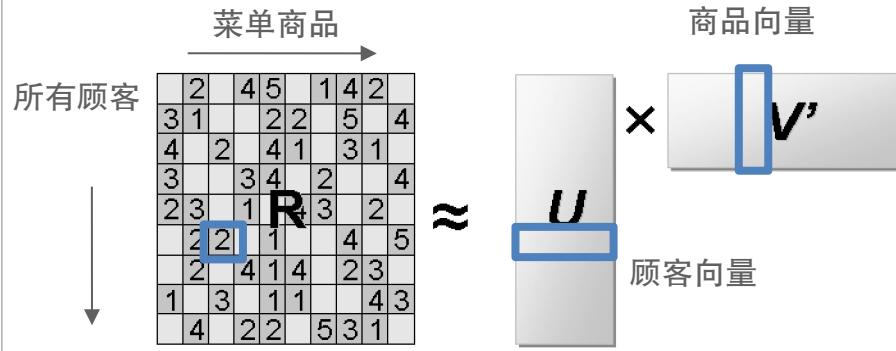
## 测量相似度

- 余弦距离
- 马氏距离
- 欧几里得距离

# 模式生成：范例

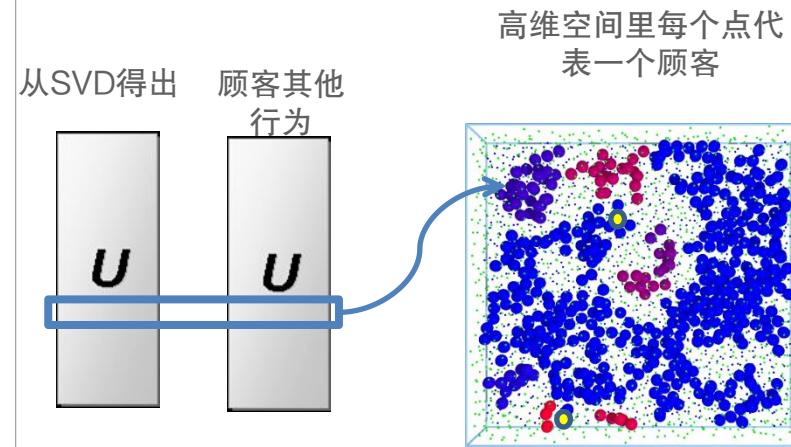
利用奇异值分解将多个中介（例如顾客）关联到多个目标（例如产品）

## 利用SVD 来捕捉目的特征



- SVD 是一个隐型变量线性模型
- 为每一个顾客和产品创建一个向量，它们的积体现了顾客和产品的关系
- 通过不同的特征和行为对顾客进行分类

## 利用K聚类算法来归纳模式分组

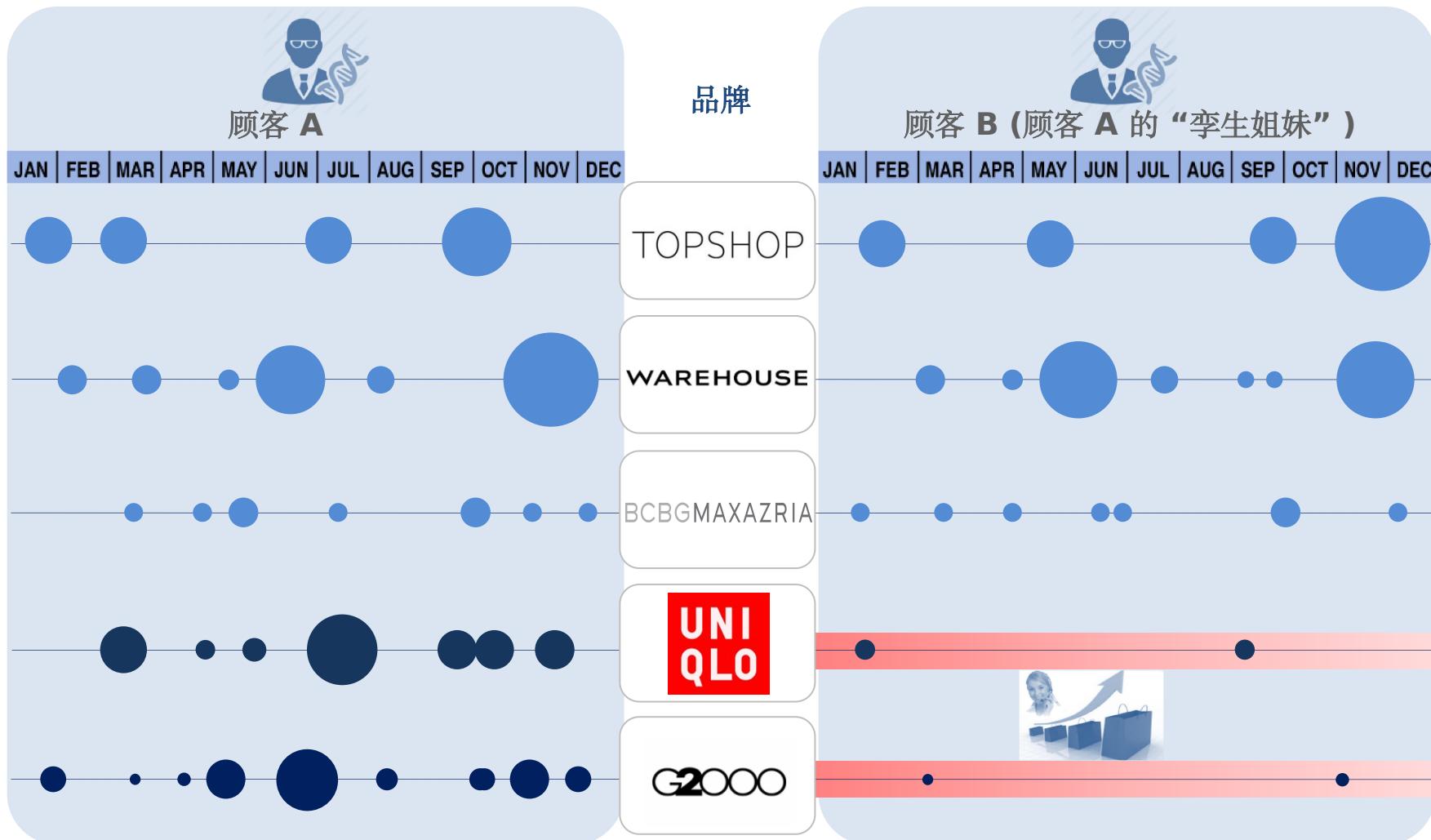


- 每个顾客用高维空间里的一个点表示
- 行为类似的顾客的点之间的距离较近
- 聚类算法归纳出有代表性的类组

# “孪生姐妹” 模型 - 同侪影响



我们的推荐引擎能通过分析顾客的相对购买时间、消费潜力和实际花费来精准刺激客户消费



# 阶段性方案管理

通过分析每个顾客的喜好倾向，系统可以设计出一个具有针对性的预测方案，并且很快的根据顾客的反应做出调整

## 第一份促销方案

- 根据客户历史
- 深度分析

### 方案 1

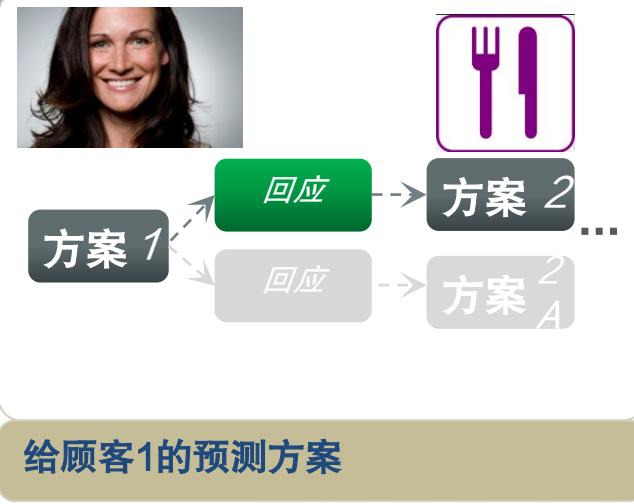


## 接下来的促销方案

根据：

- 预测的顾客反应以及
- 客户对之前促销的真实反应

### 顾客1



## 给全部客户的促销方案

Customer 3

Customer 2

顾客1



# 对历史促销活动的分析



通过分析历史促销记录，找到最有效的促销渠道

**顾客的敏感性**  
按不同渠道分类

顾客反应热度    低    高    预测的    实际的



# 客户的营销方案优化



我们对客户的实时数据流进行整合分析，提取最优实用信息，用以客户的营销方案优化

**CUSTOMER**

**CUSTOMER**

**CUSTOMER**

**客户 1**

**前四位最需商品预测**

**潜在客户: Tony Tan 男士**

**已知信息**

- § 信用卡花费: \$265000
- § 年龄: 35
- § 子女: 2
- § 已婚
- § 住址: Buona Vista
- § 爱好: 读书, 旅行 (来源: Infobase)

**家用器皿**

**户外设备**

**健康养生**

**汽车维护**

**营销方案优化**

客户 # 1: 可选方案	销售渠道成本 (\$/顾客回复)	商品价值 (\$/顾客回复)	方案价值 (\$/顾客回复)
<b>Facebook</b> (商品 1, 销售渠道 1)	\$5	\$100	\$95
电邮营销 (商品2, 销售渠道 2)	\$3	\$50	\$47
<b>客户 # 2</b>			
<b>Facebook</b> (商品 1, 销售渠道 1)	\$10	\$40	\$30
电邮营销 (商品2, 销售渠道 2)	\$3	\$80	\$77

**优化权衡因素**

客户角度 方案价值 	方案角度 预算 	方案角度 总量 
	<b>Facebook</b>	电邮营销
客户 #1	\$95	\$47
客户 #2	\$30	\$77

## 营销花费的分析

### 几种典型的营销渠道

看促销广告

点击网络广告条

收到邮件广告

看到电视直销广告



传统营销还无法优化的问题：

- § 每种渠道的投资/回报比例?
- § 每种渠道对消费者的影响力?

通过进行复杂的数据分析，我们可以使用针对性的营销渠道来有效刺激顾客的消费。

### 每个人的刺激物的纵向分析



对每个渠道的具体贡献进行更加精确的计算

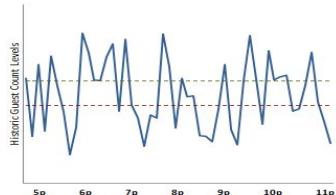
# 减少排队时间

一个优化的人力管理流程可以更好的利用期望的需求和需求的历史波动性，从而提高利润，降低员工成本

期望的时段  
AWGC



历史客流量波动分布

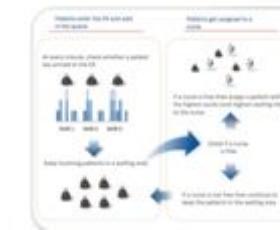


## 商业要求

- § 目标等待时间
- § 服务人员最少化

## 等待时间模拟

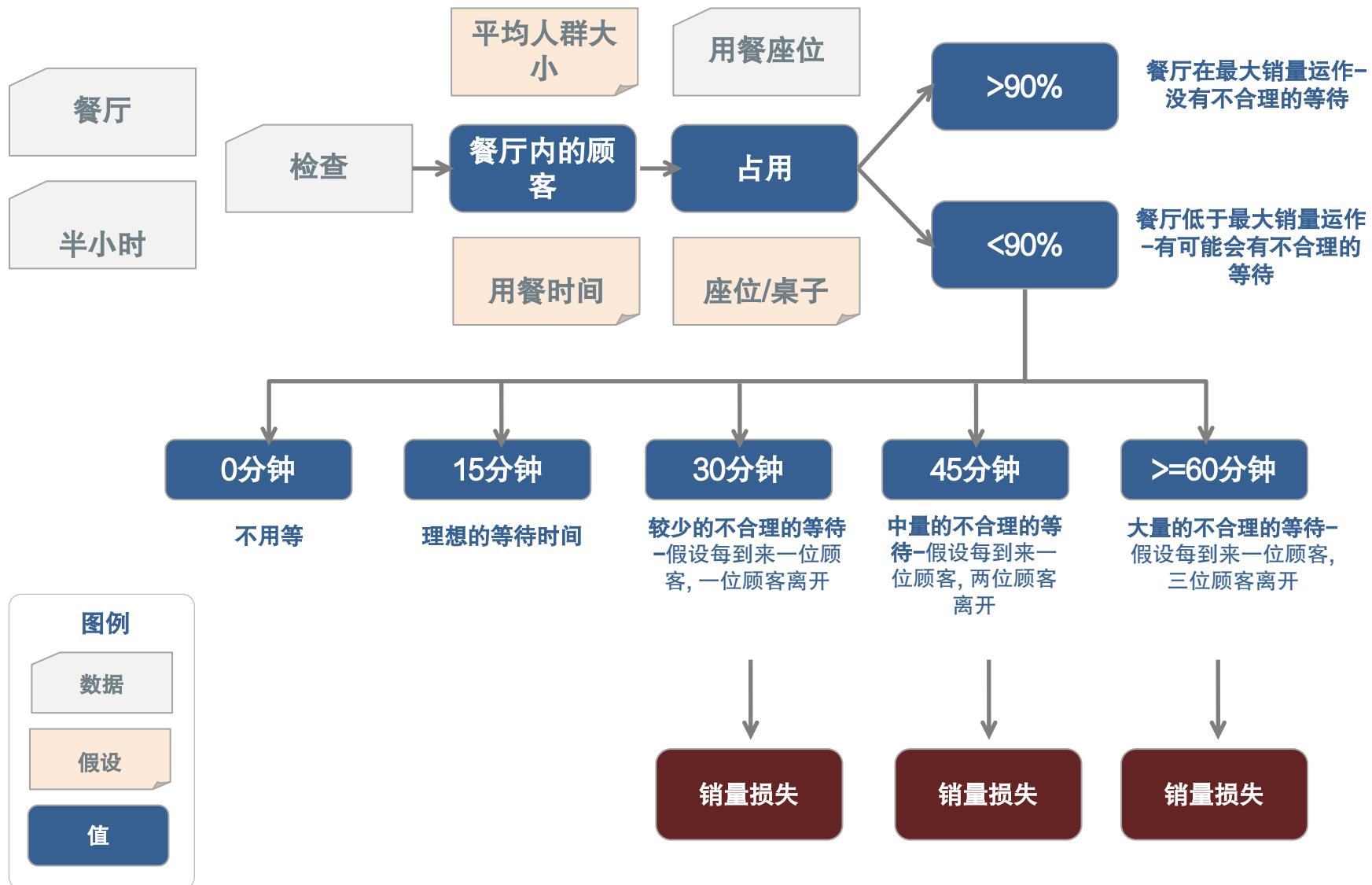
员工数量等级= X



根据历史的模式和波动性来预测顾客到访模式

优化的等待时间

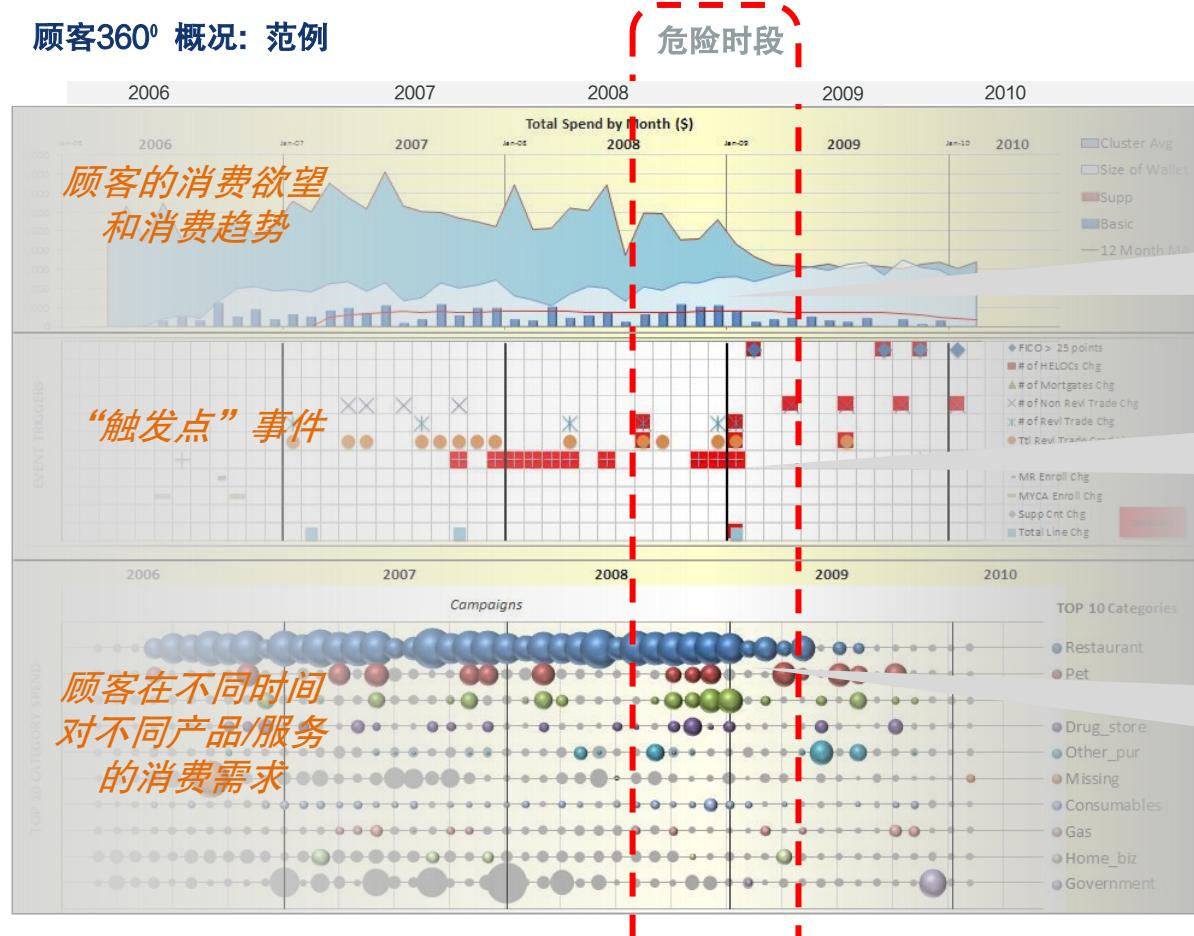
# 优化等待时间 - 机会的大小



# “兴趣衰退”的信号不易被发现

我们的“惯性模型”可以根据消费习惯的微小变化给出不同等级的警示

## 顾客360° 概况：范例



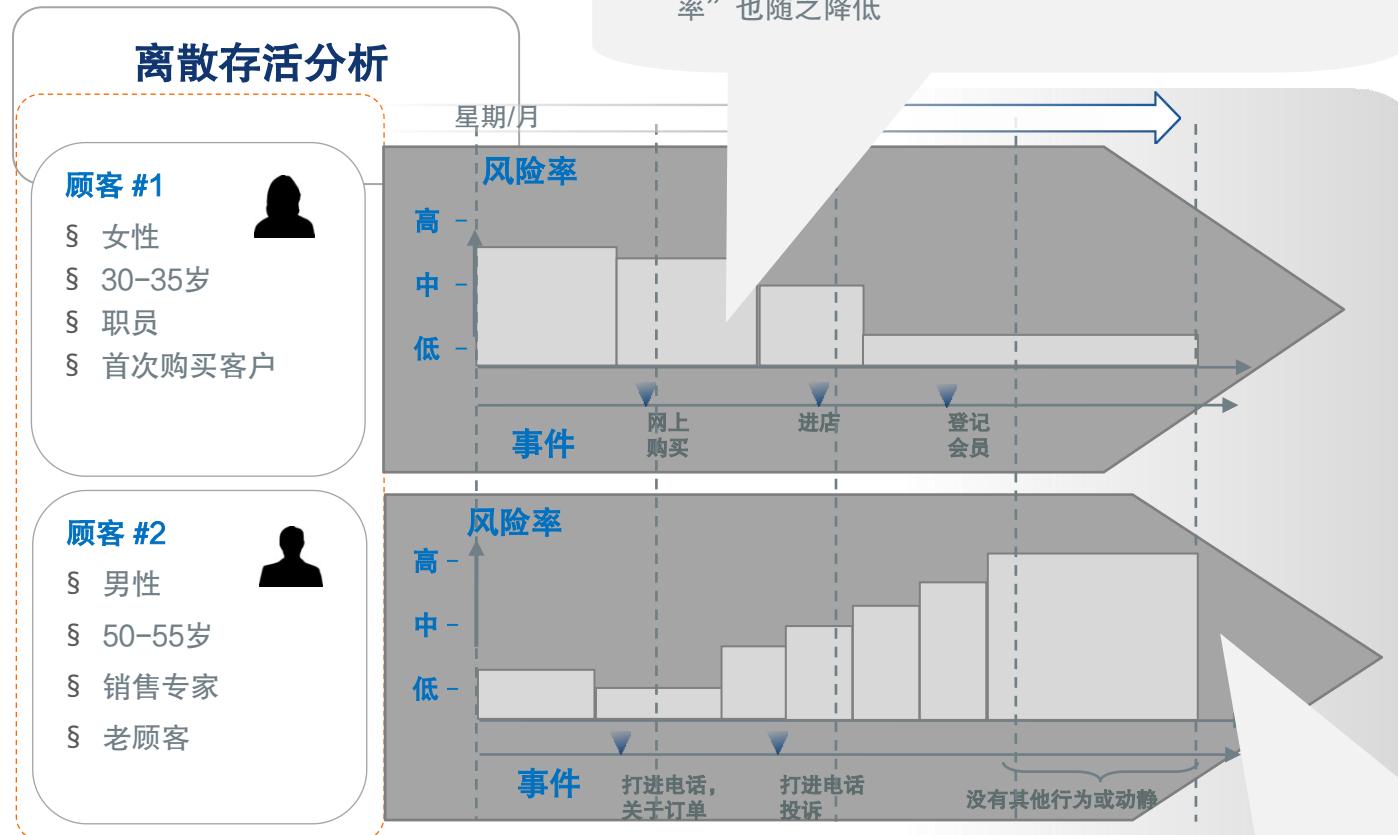
**外部因素**  
变化率:  
• 购买力  
• 顾客花费

**触发物**  
变化率:  
• 利率增加  
• 信用额度减少

**消费细节**  
变化率:  
• 餐厅种类

# 构建早期监测客户流失的模型

根据网页登陆，进店访问，电话和投诉等数据，结合“离散存活分析”的概念，得出每个客户流失的可能性（“风险率”）



- § 顾客的个人信息表明，最初的“风险率”中等偏上
- § 之后一系列事件表明该顾客的行为逐渐乐观，“风险率”也随之降低

## 结果

- 该公司找到了多出平常一倍的可能流失的顾客
- 我们发现的20%最有风险的顾客占了将要流失客户的36%

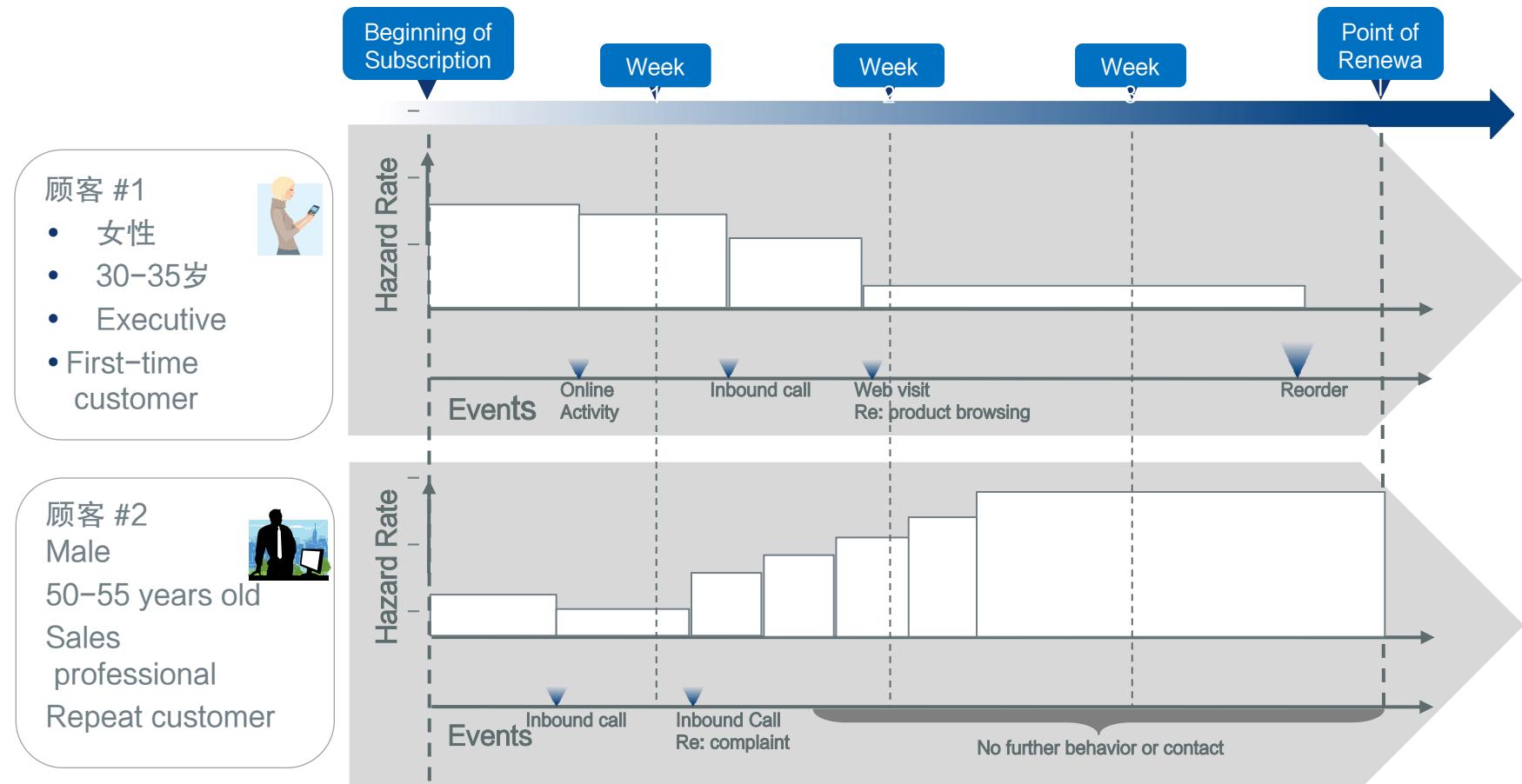
- 给顾客演示产品以及顾客登记的情况表明，一开始的“风险率”很低
- 投诉因没有被及时处理使得“风险率”逐步增加增加



# 解决方案: 有限的交易数据来预测顾客流失



因为客户仅运营一个订阅模式而且他的每个顾客一般也只续订三次, 根据传统的交易记录判断客户流失非常困难



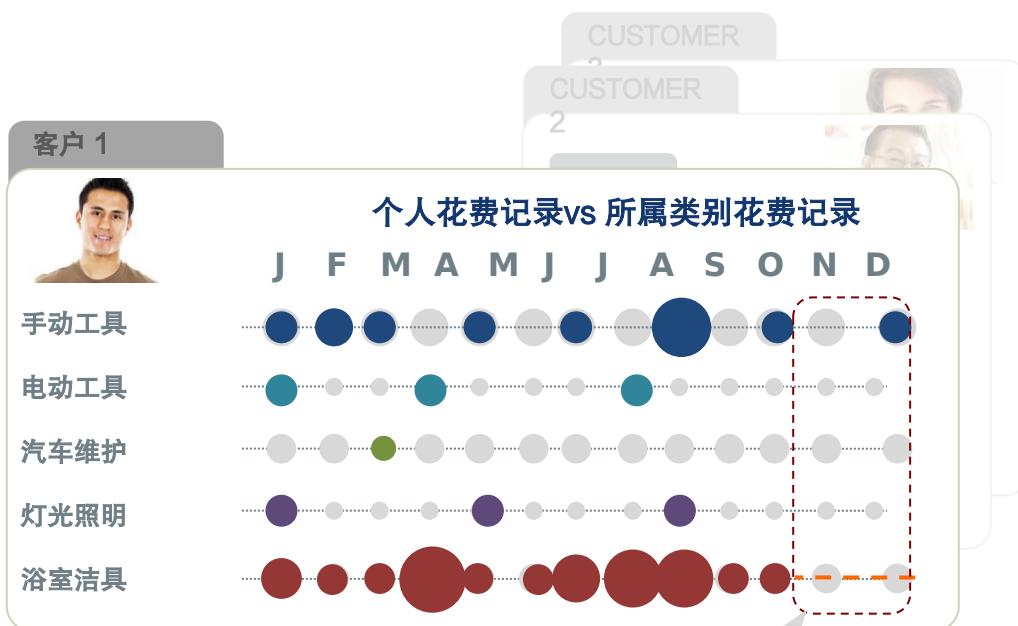
Hazard Rate – the odds of not renewing –  
changes each time new behavior (or lack thereof) is available

# 客户流失管理

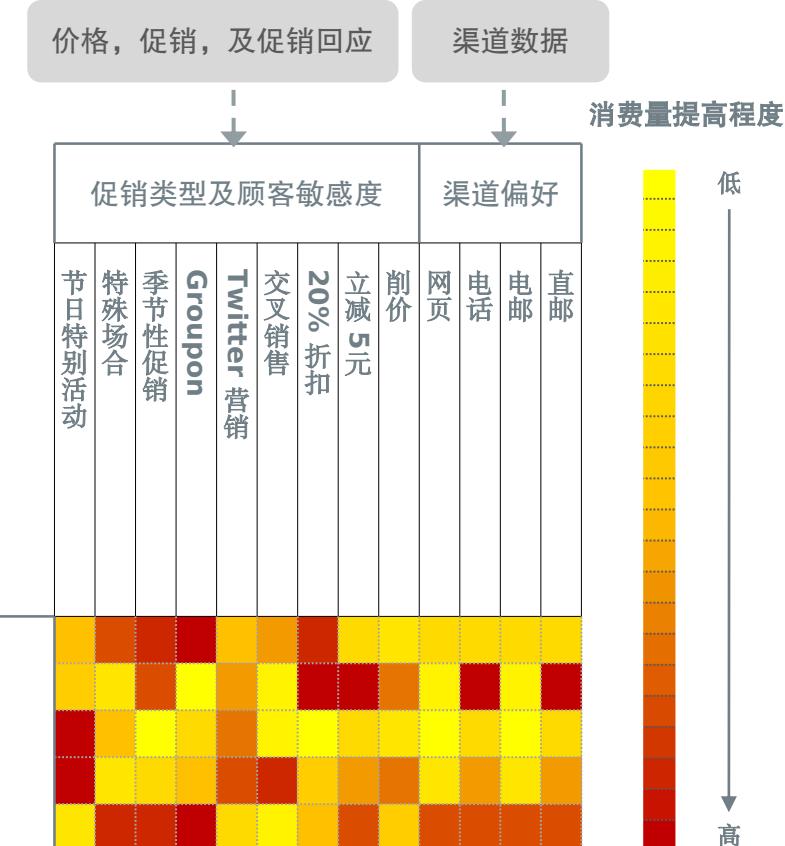


理解顾客对不同刺激方案的反应是重新赢回顾客的关键。以下范例描述了如何运用我们的系统来分析顾客对不同刺激方案的反响度，从而防止顾客流失。

## 防止客户流失方案



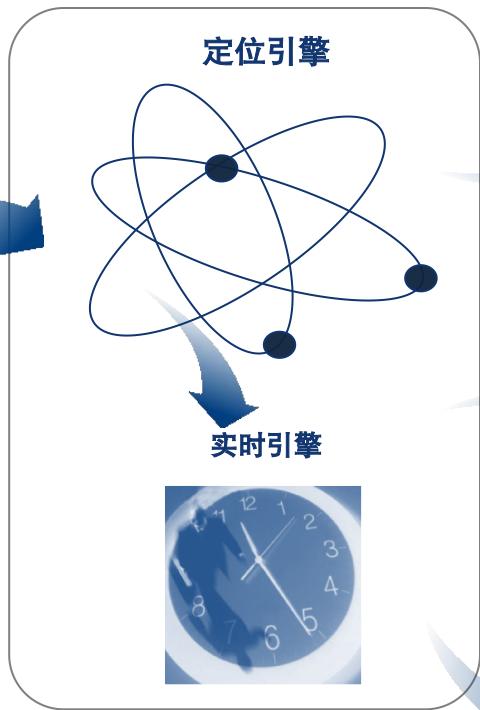
## 营销方案优化



# 基于位置的智能



InsightBox的定位引擎可以生成每位顾客的“移动DNA” -企业可以根据顾客的实时动态提供相应的提案



处理百万计的数据: 收集方法:

- 顾客允许采集的移动信息
- 信用卡交易/认证系统



示例, 快餐店偏好  
(McDonalds Vs. Pilot)

- 实时捕捉顾客的位置
- 提取地点特征
- 和顾客的Insight数据相结合
- 提取社会, 人口, 商业方面的信号
- 通过多元运算为每个用户创建DNA



# 示例一目标顾客：“商务旅行者”

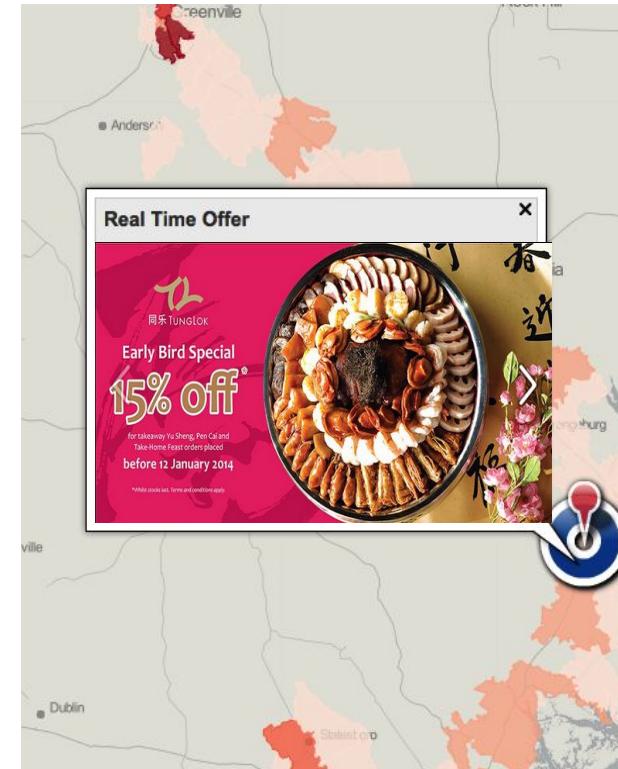
目标：根据商务人士的旅行地点，时间，和对产品/服务的需求，提出相关的旅行配套方案。



顾客是一位商业顾问，每周在两个城市之间往返



在旅行超过350英里时，该商业顾问会收到与他的时间地点相关的提案



提案示例：午餐/晚餐优惠券 – 适用于长途飞行结束或入住酒店后

# 汇总



通过整合外部的数据资源，我们可以进一步充实现有的数据库，估算出每个顾客的潜在消费能力；从而通过推荐引擎给出方案、观测这个方案的执行效果、反复学习并改良模型

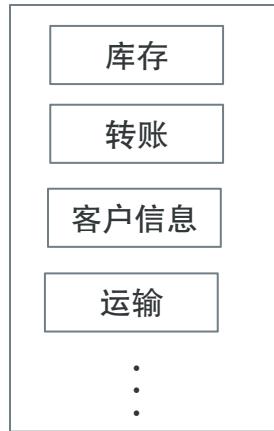


# Insight-as-a-Service 平台



针对时尚零售行业的客户，我们专门研发了基于云系统的分析方式，能同时准确识别市场的动态变化趋势与静态消费规律

## 源数据



## 32节点 Hadoop 类别



用户界面

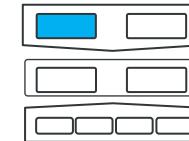
- 数据大小
  - ~ 1000万次转账/月
  - ~ 300万活跃用户
  - 4年历史记录
  - ~ 200 GB 数据总量

- 时间要求
  - 每日处理时间需要控制在2小时以内
  - 用户界面指令的反应速度：毫秒级

- 设备要求
  - 32 Node cluster + 1 master node
  - Ubuntu 14.04
  - Dual Xeon X5550 @2.67 GHz, 64 GB 存储
  - 2x500GB data drives per node

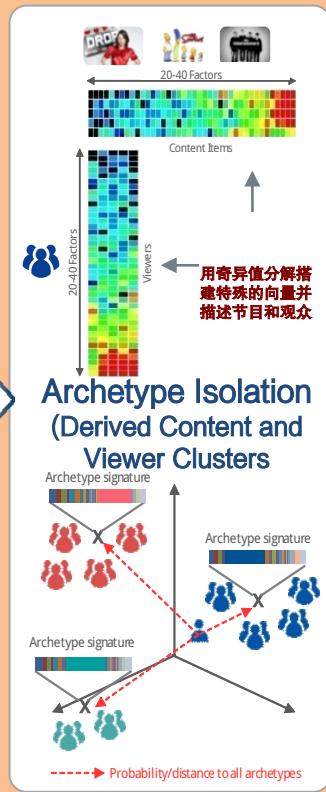
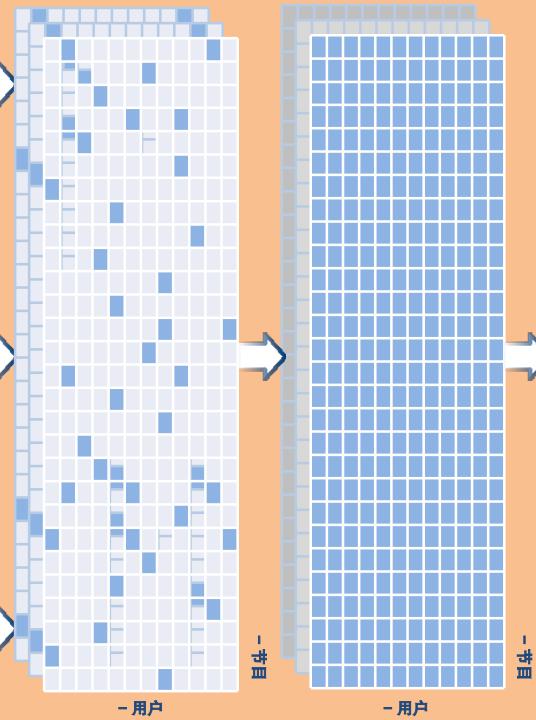
TX: 转账  
 HDFS: Hadoop 文件分类系统  
 MSSQL: Microsoft SQL 服务器

# 人口特征预测: 方法概述 (1/2)



## DataX-ray Synapse™

多重数据重叠      多种数据融合      数据降维      特征提取

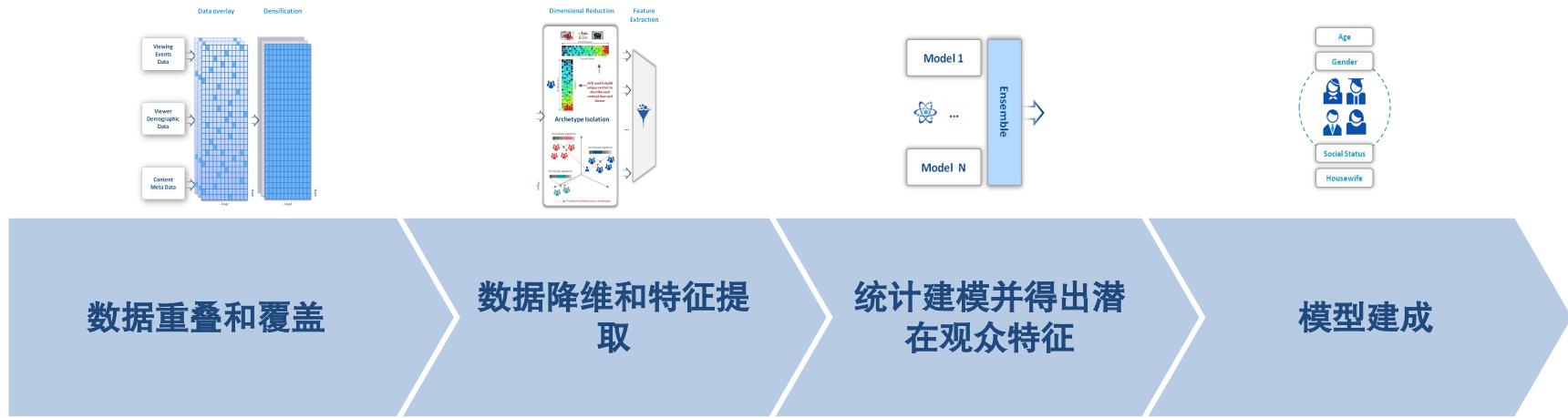
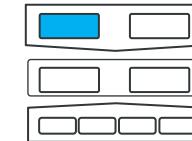


## DataX-ray 运用

统计建模      潜在观众特点



# 人口特征预测: 方法概述 (2/2)



- 观众的观看习惯是一个稀疏矩阵。系统可以把这个矩阵和推荐引擎相融合，从而预测出观众很可能喜欢看的节目
- 观众-内容矩阵是由实际被观看的内容和预测将会被观看的内容共同组成的
- 具体使用到的数据包括实际观看事件记录、观众人口特征数据和内容元数据
- “观众-内容”的密集矩阵的特征会被降维之后提取出来，使得最后每个观众的矢量仅包含最相关的特点
- 在已降维的矩阵的基础上开展一系列监督学习和无监督学习，最终分条列点指出不同类型的潜在观众的不同特征
- 最终建成的模型可以用于对未知特征的观众进行预测，从而推测出他们可能具备的特点

# 分析至上的餐饮工程

Insightbox可以提供更多的有创意的方案来使您的生意保持活力



## 创新

- § 菜单项目分类(设置套餐) 可以轻松增加销售
- t 购物篮分析

- § 通过优化菜品名称来提高品牌效应
- t 夏纳遇上北海道

- § 通过优化菜单更新的频率来吸引更多顾客

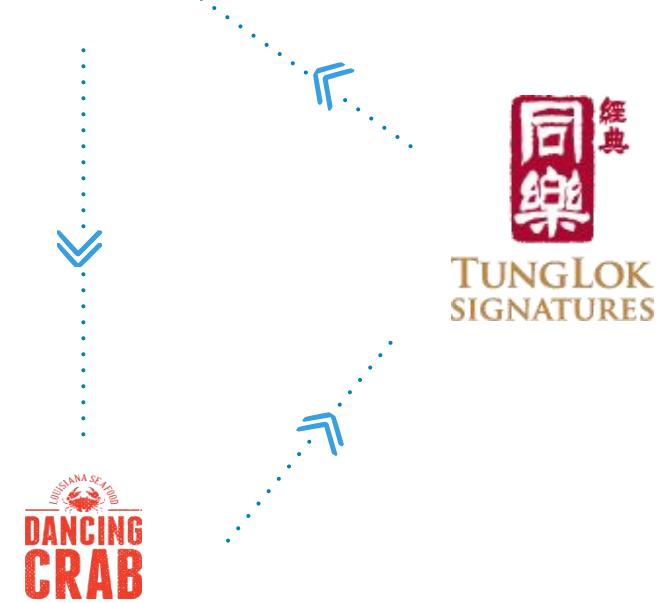
# 算法驱动的追加/交叉销售方案

定制优惠方案，鼓励顾客访问您的其他店铺，从而提高销售

## Traffic shaping



- § 新店开张
- § 根据顾客位置制定的优惠政策
- § 销售提升



# 鼓励顾客频繁光顾



通过仔细设计的优惠券和品牌推广计划，可以大幅提高客户的忠诚度



光顾同乐



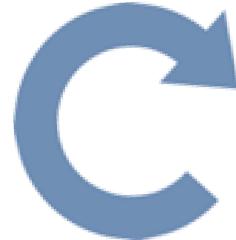
优惠券返还机制



兑换同乐的赠品



再次光顾



把最好的同乐产品  
带回家





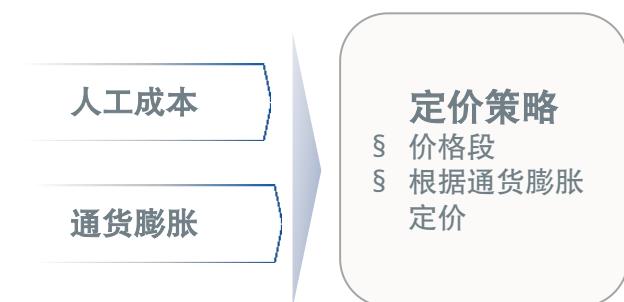
# 定价

# 定价策略



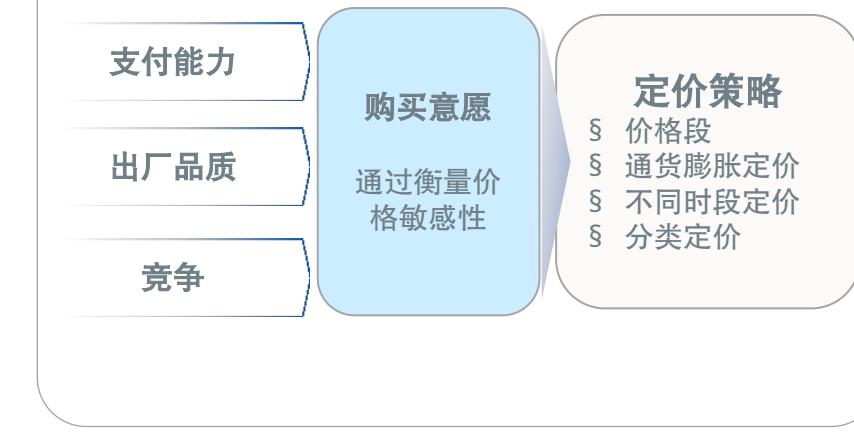
## 过去的市场动力学

### 历史的定价方法



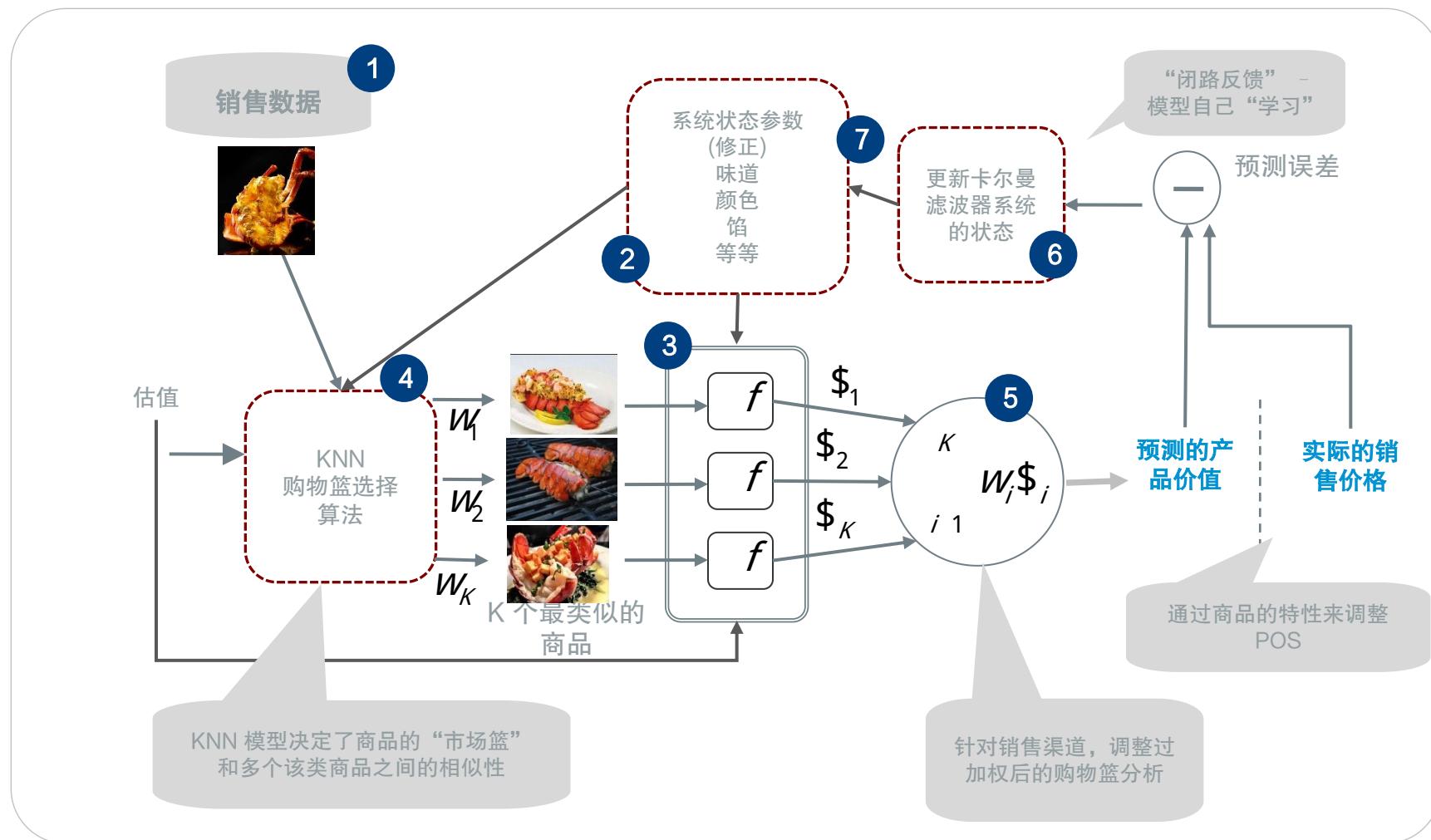
## 现在的市场动力学

### 推荐的定价方法



# 基于数据分析的定价

通过引导性的分析技术（例如卡尔曼滤波法），我们的定价模型可以结合大量的市场交易数据与人类的智慧来更准确的预测市场标价



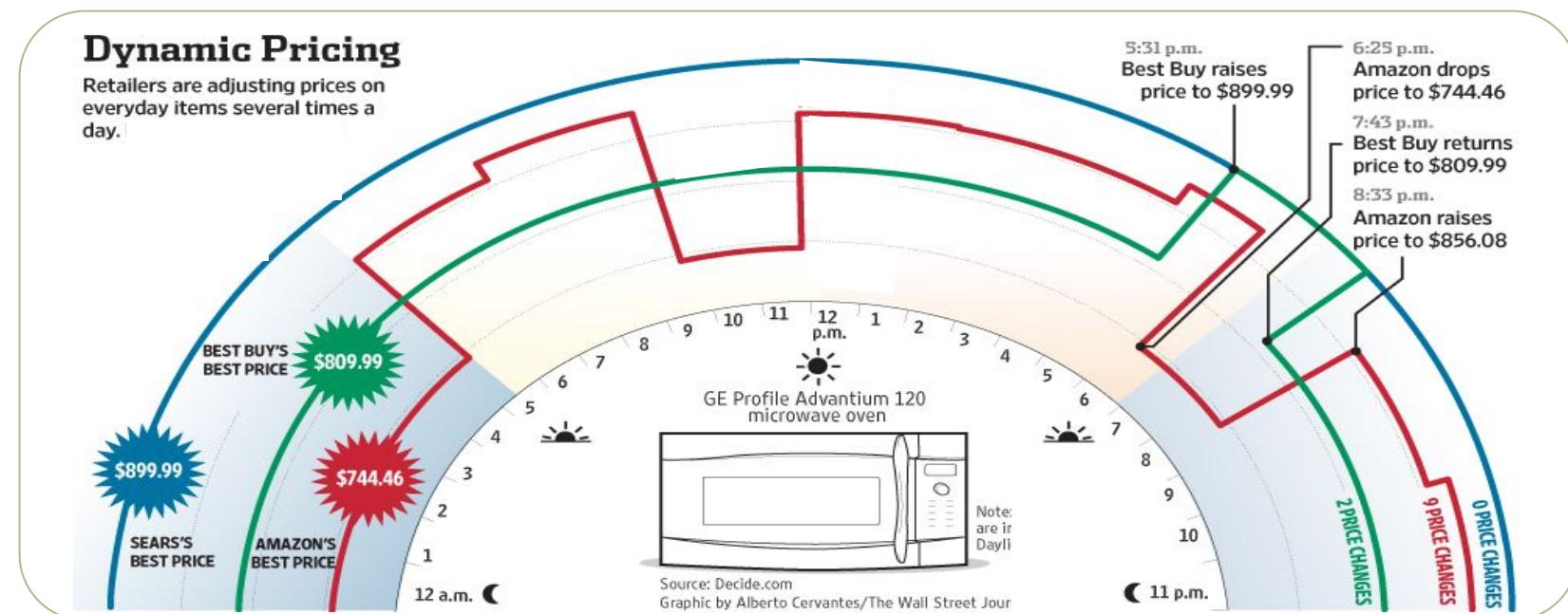
# 动态定价

InsightBox的动态定价模块能帮助客户获得和亚马逊、百思买等行业领导者所具备的零售价优惠能力



被监督的  
特性

产品名称,  
图片  
促销  
价格  
评分



# 我们的智能系统驾驭网络



我们的智能系统拥有三个模块，能实时整合大量的线上数据，将其转化为可被立即落实的方案，帮助我们的客户做出最具效益的商业决策

## 我们的智能系统

§ **辨别** 数据中的最优资源

§ **抓住** 数据中的最重要信息

§ **测算** 最重要的价格走势，营销策略，与顾客情绪信息

§ **盈利** 于由数据分析得出的最佳策略

### 定价与推广模块



- § 监测竞争对手所有产品的实时价格，促销方法，产品通告与优惠券发放情况
- § 数据实现纵向对比：每小时走势，每日走势，或根据客户需求选取时间节点间隔
- § 价格临界值警告系统
- § 监测竞争对手的促销行为并自动发出提醒，实现快速回应
- § 可被立即落实的推荐方案：计算定价水平，优化促销方案，推进销售额增长
- § 产品与客户ID实现一一对应，让每一次数据分析更快速、更准确

### 附加模块

### 顾客情绪模块



### 市场化信息模块



# 智能系统: 定价与推广模块

我们定价与推广模块能够提供最智慧的价格策略，配以可被立即落实的执行方案，能准确抓住任何一个足以直接改变你心中价格底线的机会



## 结构性数据: 竞争对手网站

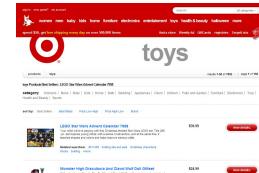
难点: 同时对多家竞争对手的所有商品进行价格、促销模式等实时追踪与比对



ToysRus®



Walmart®



TARGET



amazon.com®

## 定价与推广模块 功能:

### 采集数据

§ 同时采集竞争对手多种形式的数据，做时间上的纵向对比：

- 产品名称，图案，型号
- 价格，促销，折扣，运输
- 用户回馈：评价与评分

§ 添加多种信息源来实现关系构建：新闻，天气，库存

### 商品分析

§ 将产品特性、图案等数据通过 InsightSphere 与对应产品绑定分析

## 定价与推广模块 优点

### 基于智能推荐系统的定价与推广模式

§ 根据对手的定价与促销方法推荐具体定价、推广渠道，实现快速响应

§ 产品与客户ID实现一一对应，让每一次数据分析更快速、更准确，能准确得知产品所有销售信息

### Enhance Comp Shop & War Room

§ 通过智能网络辅助价格比对，使关键数据的采集更迅速，更便捷

# 定价与推广模块：深入分析 智能系统对竞争对手的有效识别

我们的模块通过SKU层级数据分析（竞争对手的产品，价格，促销方法等）来提出优化方案，从而增加客户TRU在定价与推广方式上的竞争优势

## 数据提取区域事例



The screenshot shows the Amazon.com website for the 'Toys & Games' category. It highlights several key areas:

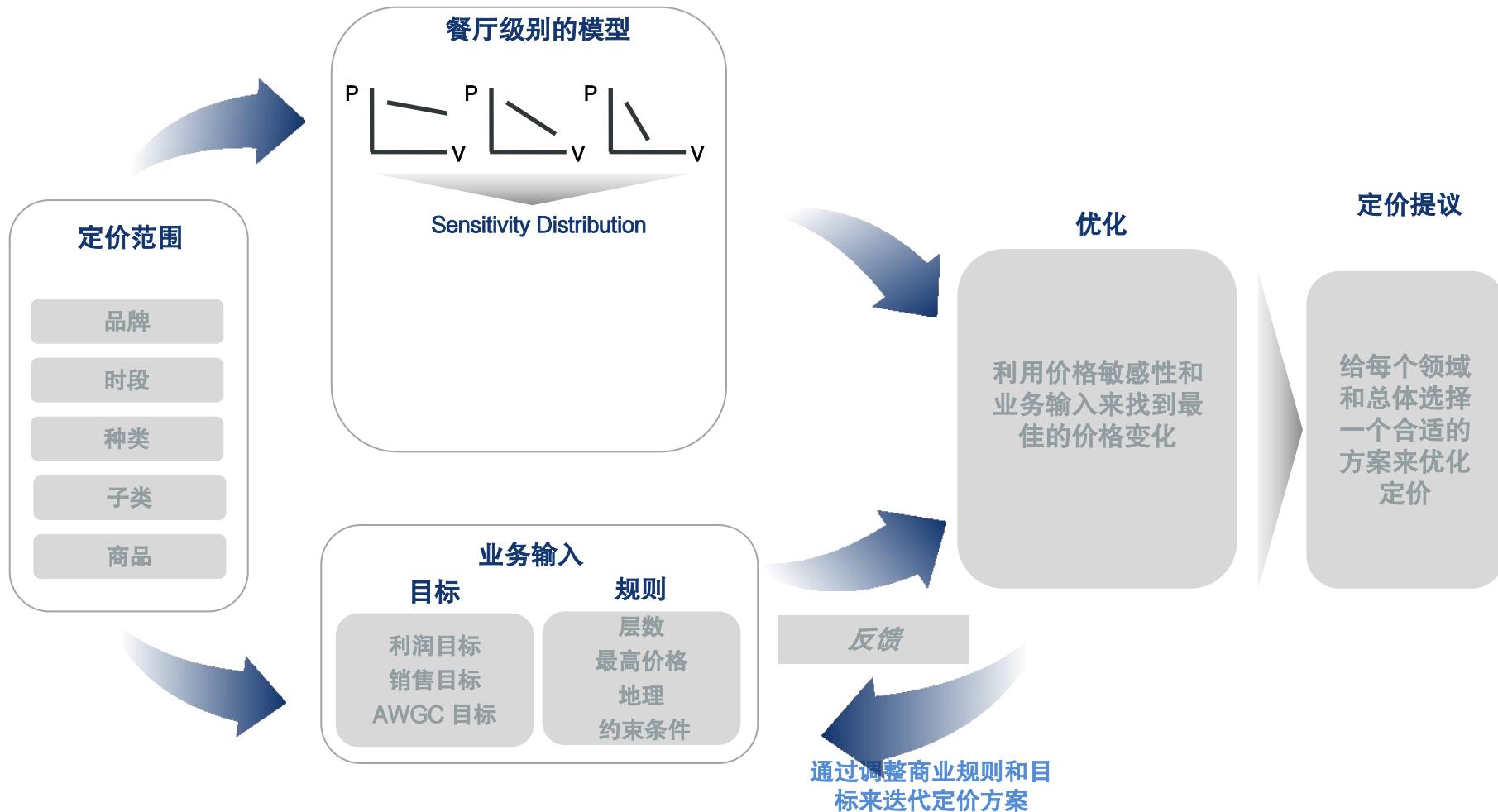
- 1. 产品名称，图片与产品编号**: Points to the product listing for 'Marvel Captain America With Spinning Shield'.
- 2. 价格与价格差异**: Points to the price section showing 'Buy new: \$15.99 \$16.50' and '19 new from \$6.50'.
- 3. 促销方法**: Points to the promotional offer 'Up to 50% Off Select Toys'.
- 4. 用户反馈**: Points to the product rating and reviews section showing 4.5 stars and '(9)' reviews.
- 免费送货**: Points to the shipping information indicating 'Eligible for FREE Super Saver Shipping'.

## 定价与推广模块 分析

- 1. 分类 - 产品名称，图片与产品编号 - 与竞争对手产品的横向比对**
- 2. 价格与价格差异 - 采集竞争对手价格用以:**
  - 分析竞争对手价格走势
  - 在关键产品与促销额度方面，建立经销商销量与产品差价之间的模型关联
- 3. 折扣/产品促销 - 知道竞争对手何时大规模打折促销，并且采取相应措施**
- 4. 用户反馈 - 分析用户反馈趋势，了解用户情绪走向**
- 5. 产品与用户建立关联 - 产品与用户ID一一对应，准确快速做出对价格、目标群体、推广时间、推广地点等等方面的调控**

# 优化定价的过程

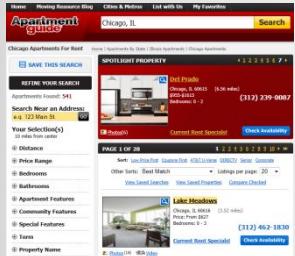
在粒度级别建立价格敏感性模型可以帮助优化定价决策和实现商业目标。



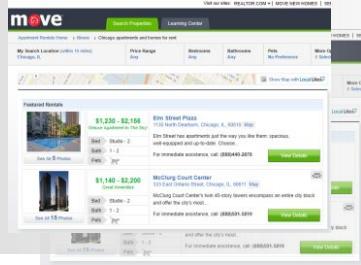
# 动态定价

为了了解创造更加动态的定价的潜能，我们应用了网络收集器来抓取每天的都市范围内的信息，并且用人力研究来补完它们

## 分析客户网站



## 分析竞争者的网站



## 其他的数据资源



### 例子：

- § “Mystery Caller”
- § 人口统计
- § 客户数据

## 提取数据

- § 每个网站每天更新的商品列表
- § 公寓细节

- 大小床/浴室
- 租金
- 地址

- § “社区” 细节
- § 价格

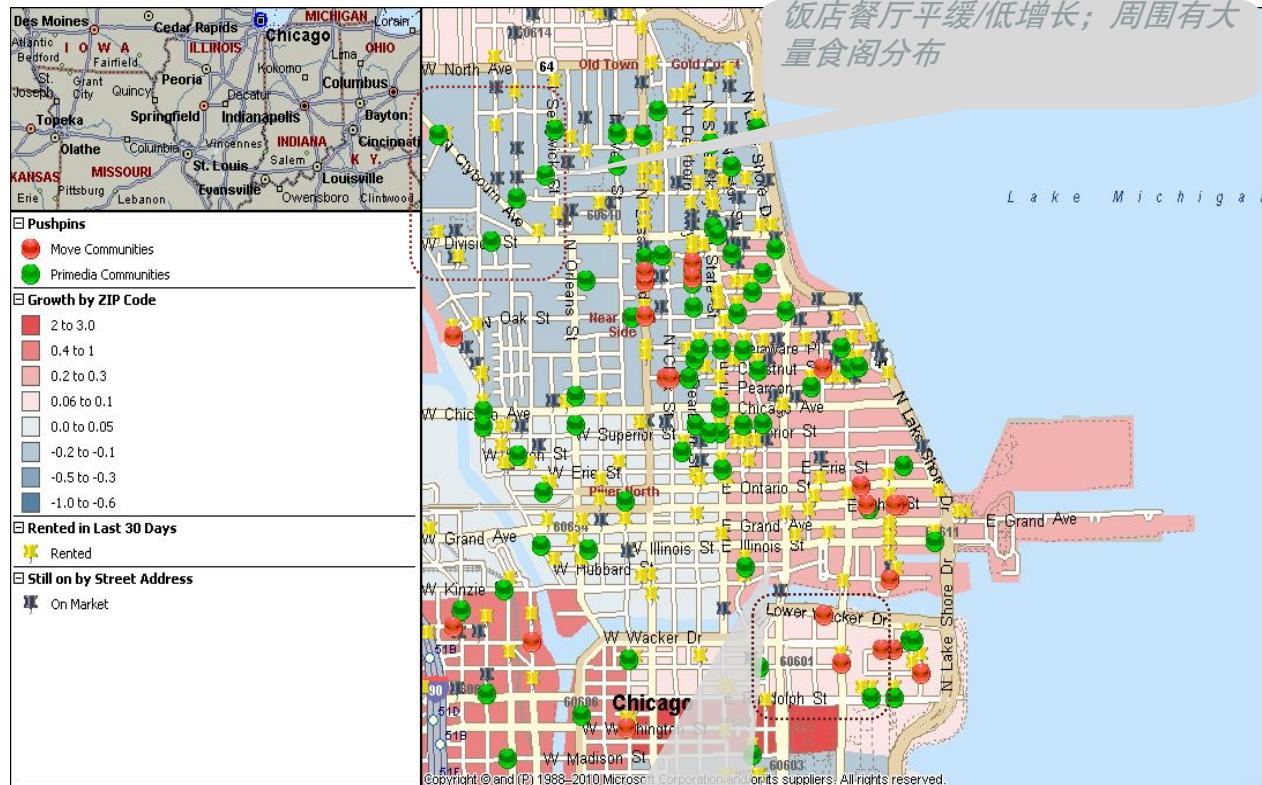
## 分析

- § New “communities” to target
- § 市场售屋时限(DOM) 趋势
- § 容量变化
- § 价格变化对DOM的影响
- § 当地“市场”定价参考

# 升级市场分析 - 微观的地区剖析

分析活动列表可以实现地区市场的剖析以及识别潜在的定价调整目标

## 各餐厅及其广告增长率的地区分析



- § 仔细分析当地的客流量变化速度来理解当地的饮食需求
- § 在微观角度测量相对市场份额和客户定价能力
- § 侧重于增长缓慢和定价错误的饭店，并相应进行价格调整

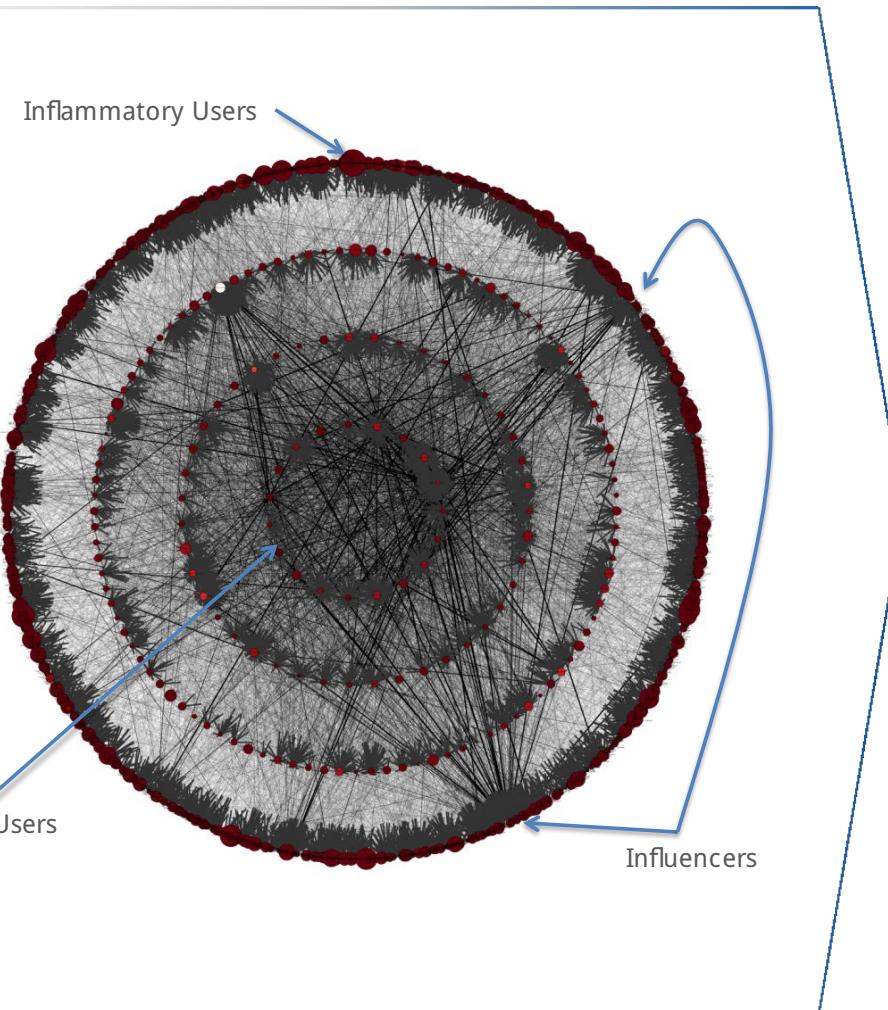


# 社交网络

# 大海捞针



我们使用社交网络分析技术来研究关系网络、定位目标人群

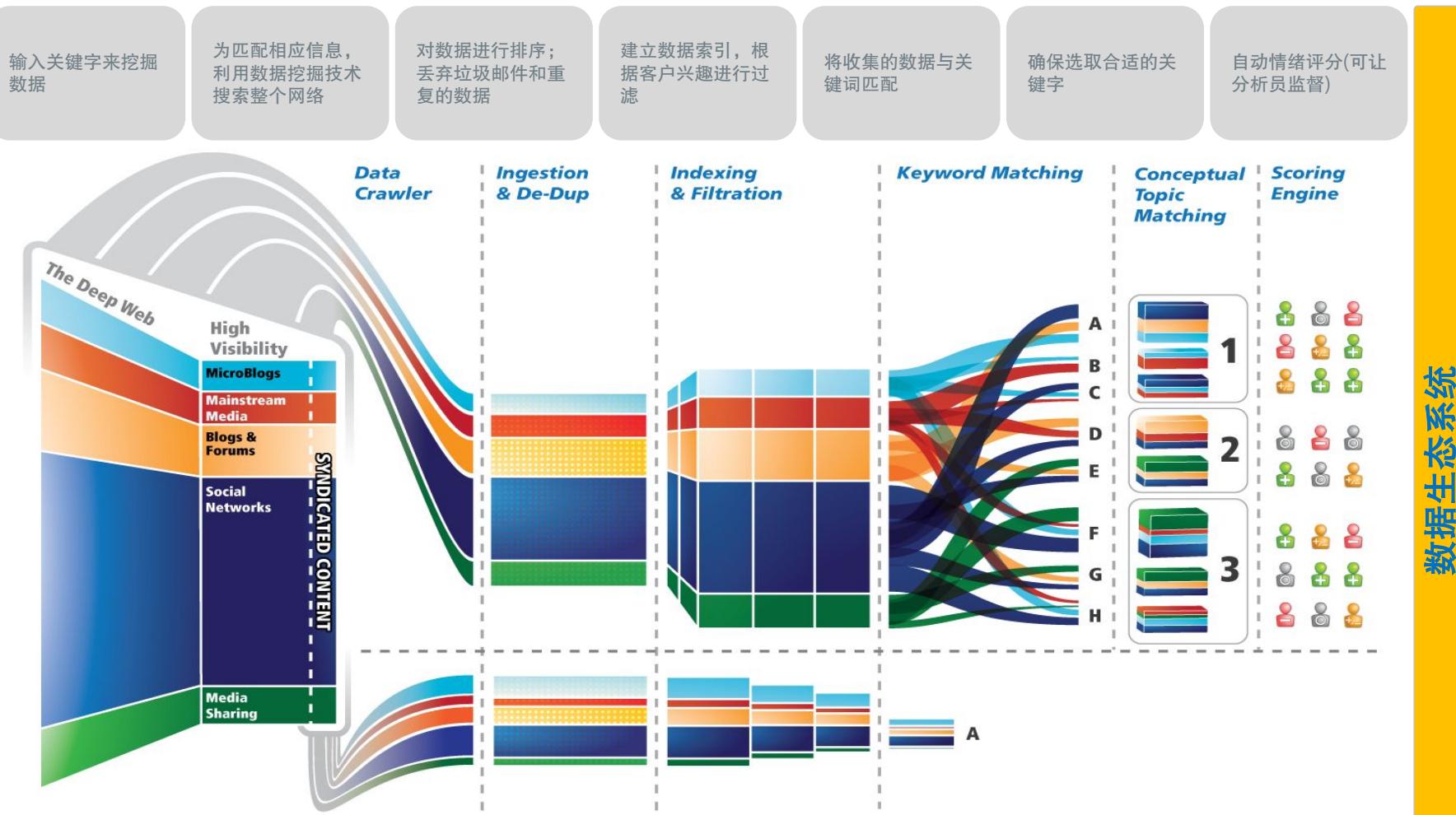


聚焦目标人群



# “倾听”客户的数据生态系统

我们的专有设备能捕捉所有类型的数据（文本，博客，日志，多媒体等等）、整合数据、关联数据、测评数据，再把它们整合到数据生态系统中



# 数据整合: 外部数据



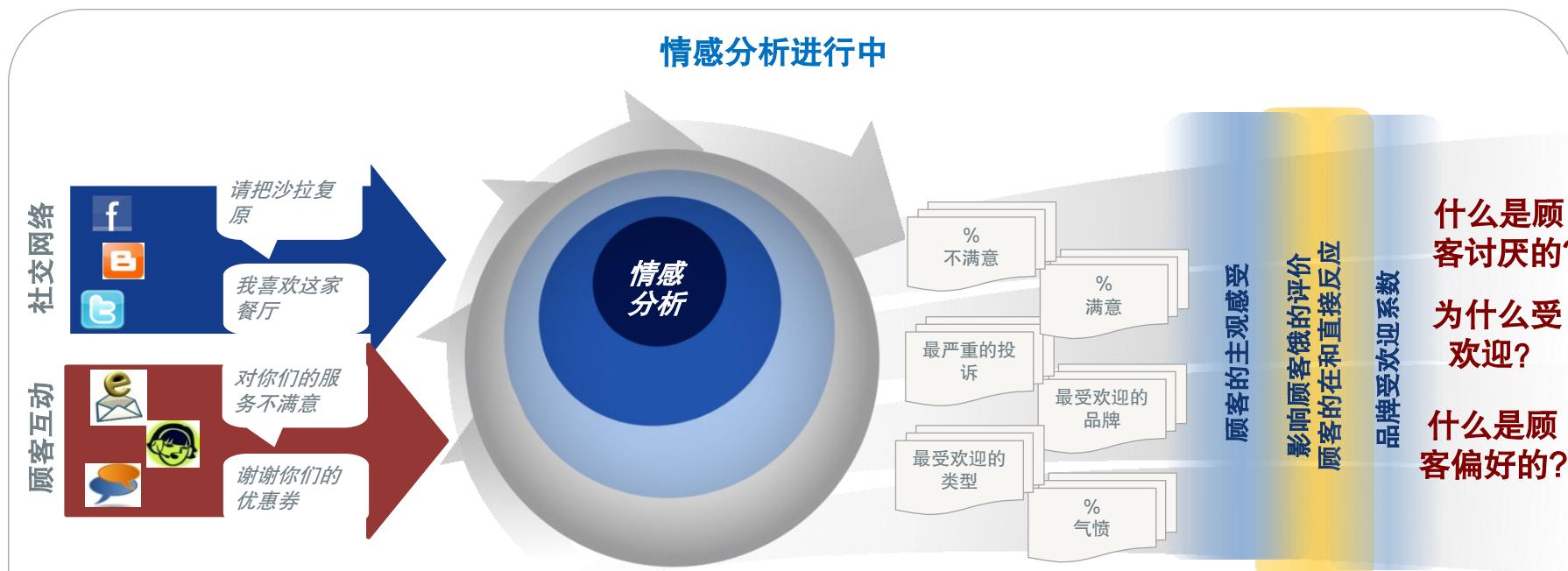
现如今, 因特网已经有超过**400亿**的网页, 成为数据爆炸的主要驱动者之一。然而, 从中提取出切实可行的商业方案并不容易



# 什么是以及为什么要进行情感分析



情感分析是指从大量文字中提取并集合主观信息，发现顾客喜欢什么和不喜欢什么的过程



我们开发出了能够快速而准确地从社区论坛数据中提取情感的过程

## 全过程

理解客户的业务用词

锁定数据中的信号

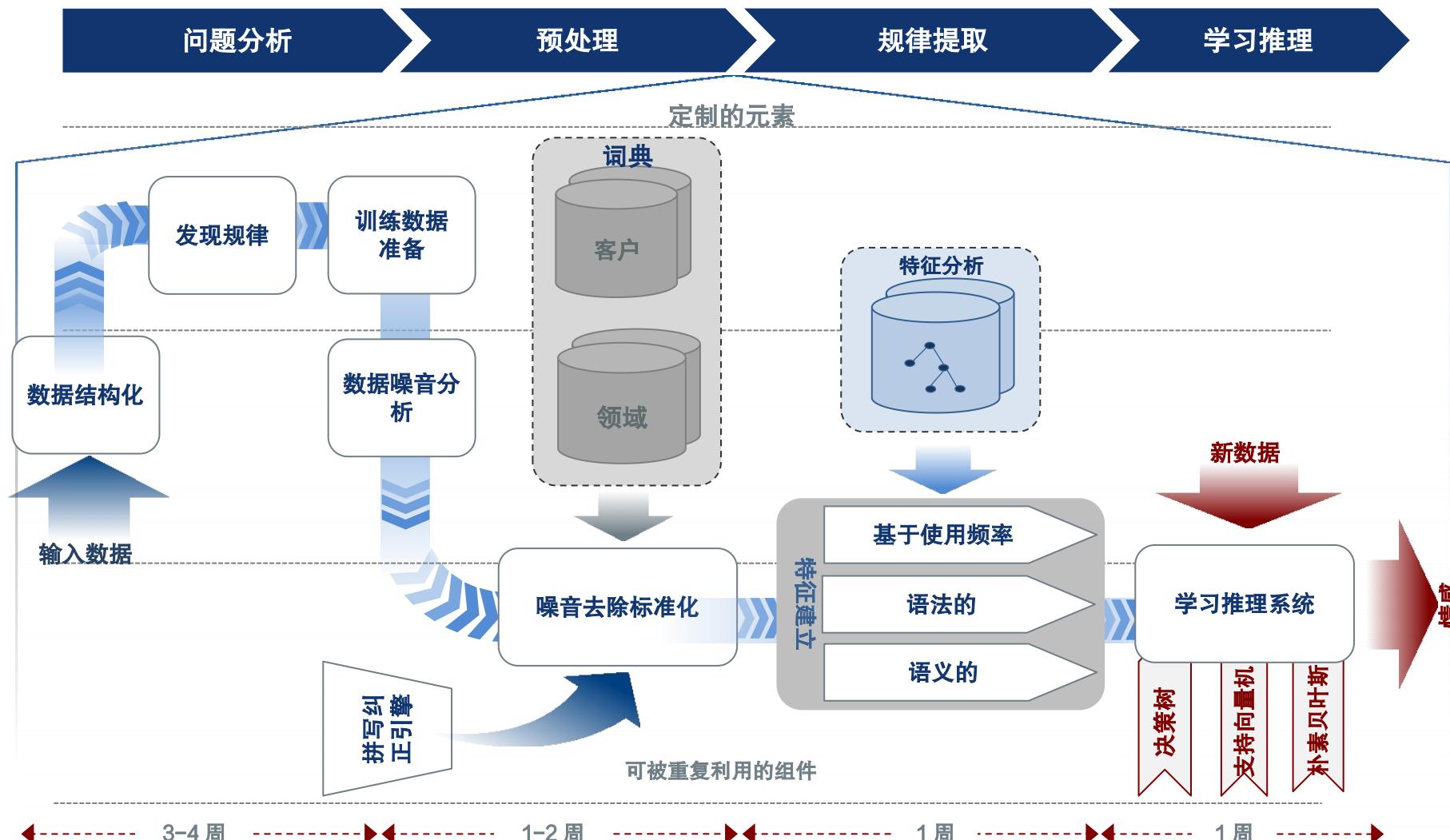
准备训练数据

训练并运行情感提取程序

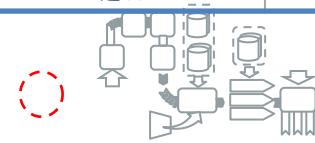
# 情感分析过程



我们开发了一套可重复利用的、能够将多源数据综合起来的强大的算法，可用来在无结构的文字数据中形成顾客情感整体观



# 输入数据类型



客户的情感存在于微博信息、聊天记录、客户反馈等数据中。此案例分析着重于网上社区数据，即餐厅微博页面的文字评论

## 使用的数据组

### 网上社区数据

从某连锁餐厅的微博页面摘取。文字从HTML格式被转换成一列列包含评论、评论作者、时间、点赞数等信息的表格。在采集的数据组里，人们有的对餐厅的各方面作出了评论，有的问问题，有的提供反馈或者提出要求等

例子



我喜欢你们的面包！

你们的沙拉是我吃过最好吃的！

请重新推出沙拉台！

上个星期在你们餐厅坐了10分钟也没有一个人来给我们点单。最后吃了麦当劳的沙拉，也不错。

### CHARACTERSTICS

中性评论一般只有几个句子长 | 类似于微信

## 其他数据组

### 聊天数据

- 顾客和客服的交流
- 每次交流可能包含多个主题
- 每次交流一般包含15-20条信息

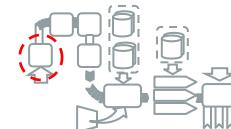
### 顾客反馈

- 包含对20个左右问题的回答
- 包含酒店房间、发布时间等其他辅助信息
- 有些反馈含有定量信息

### 博客数据

- 数据来自13个博客和8个受欢迎的社区
- 很多关于其他竞争者和受众对品牌的看法
- 有一条评论有多个段落

# 输入数据结构化



此步骤将非结构文字数据转化成结构化的一列列信息，利于接下来的分析

## 分析社区数据步骤

### 数据格式分析

#### 描述

不同社交工具的数据会有差别，例如微博和微信上的信息就不同

#### 微博页面数据：

- HTML格式
- 信息包括：收藏、转发、评论量、用户名、发布时间、评论内容等

### 数据转化

#### 描述

我们拥有能把多种格式的数据转化成结构化信息的工具

#### 微博 HTML 转化：

1. HTML 格式化文字
2. 格式化文字 CSV  
用户名, 评论内容, 发布时间, 收藏数等信息

### 反常现象去除

#### 描述

需要对一些无法转化的内容进行后期处理

#### 反常现象：

- 人为添加的标点

#### 例子：

我稀饭这个饭。。。店。。。  
我 / 饭店

## 处理其他数据格式的困难

### 调查问卷

这种格式存在于顾客反馈数据中。主要的困难是：  
辨认问题并将其与回答分开

### 聊天记录

在顾客聊天记录和顾客与客服交流中出现。主要困难是：  
将两个人的话语和姓名等分开

### 电邮

存在于通过电邮方式提交的防窥中。主要的任务是：  
理解邮件的标题以及邮件的结构

# 在数据中发现规律



当文字数据被结构化了之后，我们使用自有的分析工具定义出顾客情感的关键因素。在此案例中，49%是中性评论，31%负面评论，17%正面评论

## 微博数据分析

### 中性

包括

- 餐厅自己给出的评论
- 餐厅职工给出的评论
- 没有主观情绪的顾客评论
- 询问、空评论

星期四来用餐有特别优惠

我怎么找到你们餐厅

100% = 2,879条评论

### 正面

包括对餐厅多方面作出的肯定，例如

- 装修氛围
- 食物质量
- 食物味道
- 服务质量

餐厅整体很干净！

喜欢你们的面包！

### 其他

包括：

- 两面性评论：顾客没有给出明显的反馈意见，而是表达正反面都有的评论
- 要求餐厅提供一些服务

食物不错，但服务差

请提供一个投诉电话号码。  
食物还不错！

讨厌你们！

你们服务真差劲…

### 负面

包括对餐厅多方面作出的批评，例如

- 装修氛围
- 食物质量
- 食物味道
- 服务质量

许多主观信息会被发现：语气、满意程度、心情、顾客对政策的反应。找到准确的信号是最关键也是最费时的步骤



## 训练数据准备

为建模做数据准备时，我们需要将一部分数据标上合适的标签（正面，负面，中性）。这个步骤加上发现规律是最费时的（约需1-2周）

### 现有做法

- 训练数据是为提取关于顾客满意指数的情感做准备的。两面性评论则会被忽略
- 因为仅从饭店1提取的负面评论数量不够，所以一些从饭店2提取的例子也被用于分析

标签	数量	百分比
正面	1,423	51%
负面	483	17%
中性	877	31%
双面	80	数量少

训练数据最终数量

### 训练数据范例

姓名	评论	情感	日期时间	点赞数
李晓亮	好吃！	正面	09/05 7:15pm	
韦珊珊	饺子	中性	03/27 1:23am	
王坦坦	这是我儿子拍的照片	中性	07/07 4:44am	4
陈伟	我在这儿呢	中性	08/28 10:55pm	
陈国栋	我喜欢!!!!!!	正面	05/05 6:12pm	
刘华强	我是这家店的领班，我很热爱我的工作！希望可以升职！	中性	09/ 29 8:29am	2

# 数据噪音分析



实际的数据组里总是会有错误和一些跟数据来源或者行业有关的特殊的语言。这些特殊情况和错误只有在被更正并被标准化后才易理解

## 需要被清理/标准化

错误的语言

适用于某类数据源、行业或客户的语言

### 错误类型

1. 语法错误
2. 不完整的句子
3. 拼写错误

### 语言习惯特例: 细节

1. 标点: ..... 或者!!!!
2. 特殊符号: ❤
3. 口语词: 稀饭!
4. 加重语气: 超超超超级喜欢!

### 特殊词汇: 细节

1. 跟行业相关的: 罗非鱼
2. 客户自己的词汇: 锦绣沙拉
3. 数据源词汇: 餐厅名字

## 现有数据组中的噪音范例

我讨厌咖啡物语

我 ❤ 咖啡物语

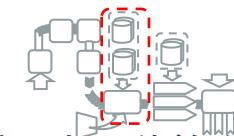
需要大众点评优惠券

我稀饭它

超超超超级烂

你们的罗非鱼沙拉不错

# 噪音去除与标准化: 步骤



我们的噪音去除步骤综合了语言更正、特殊行业只是和客户自己的语言，从而能够降低数据中的噪音、提高准确率

## 流程

### 标点符号和语义着重

包括

- 标点更改
- 语义着重更改
- 空格修改
- 使用字典来理解



给输入信息排序

可嵌入的  
拼写更正引擎

### 拼写更正步骤

- 运用第三方拼写检查软件
- 综合客户排好优先顺序的自定义字典

标准干净  
的数据

报告

## 范例

牛逼!!! 超超超稀饭  
罗非鱼

牛逼! 很 稀饭 罗非  
鱼

牛逼! 很喜欢罗非  
鱼

很好! 很喜欢罗非鱼

很好! 很喜欢罗非  
鱼

语义着重去除

拼写更正

常用词词典

行业知识

# 噪音去除与标准化: 运用



拼写检查还有其他的运用，例如找出对以后分析有用的元素

## 预贴标签

此步骤给一些词贴上含有辅助信息的标签，以帮助引导理解这些词语。以下例子表明了一些词语是怎样因被标为普通名词来帮助理解词性的

罗非鱼    罗非鱼/普  
通名词

咖啡物语    餐厅名字/普  
通名词

## 标准化

将一些口语形式转化成标准形式。例如，口语词、特殊符号和缩写

稀饭    喜欢

牛逼    很好

餐厅缩写    餐厅名字

¥    人民币

## 概括

为了在分类过程中不产生偏差，我们往往会将一类词概括成一个常见词。可以通过使用词典或者将其视为拼写错误来实现

沙拉    菜品

罗非鱼    菜品

咖啡之翼    竞争  
对手

上岛咖啡    竞争对手

## 拼写更正

此步骤能够在更正拼写后，给拼写检查引擎因缺乏客户自定义的、行业相关的信息而将一个有意义的词标为错词时的自带的词典更新

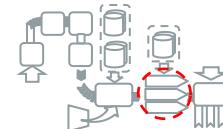
稀饭    喜欢

苏州    苏州

罗非鱼    罗非鱼

麦麦    麦当劳

# 特征建立: 步骤



特征建立会将文字信息转化成一种分析模型能够读取信息的格式。DataESP的特征建立工具通过三步将文字数据转为机器可读的特征

## 文字特征建立

### 多种特征建立

- 目标:** 使用多种方法能将文字转化成数字的方法，再将处理过的文字导入到算法中
- 使用:** 语法要求、使用频率、语义要求等信息来综合理解文字
- 结果:** 多种把文字转化成向量的表达方式

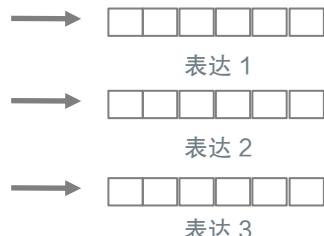
### 综合特征

- 目标:** 将上个步骤得到的多种特征混合或者创建一个复杂特征
- 使用:** 将要被混合的一个或多个表达方式；能够进一步处理上个步骤得到的多种表达方式的模块
- 结果:** 把文字转化成向量的表达方式

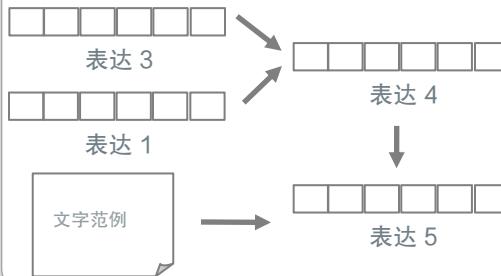
### 降低维度

- 目标:**
  - 移除结尾词:** 将没有太多意义的词语移除
  - 特征去除:** 将之前一些不必要的特征去除
- 使用:** 常用结尾词列表, 倒排文档频率, 文字使用频率

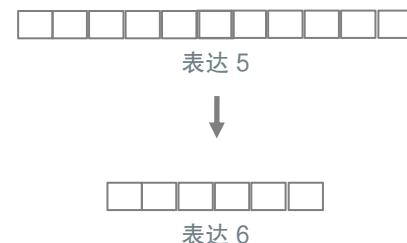
### 范例:



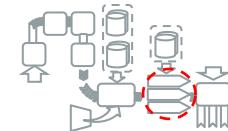
### 范例:



### 范例:



# 多种特征



使用多种特征来尽可能多地保留文字原意

## 基于使用频率的特征

- 在一条文字信息或一组数据中，各词语被使用的频率可被用来描述这条文字信息的意义
1. 一个词被使用的频率越高，这条文字信息和该词表达的意义的重合度就越高
  2. 一个词在多种文件中出现的频率越高，该词在此情景中所包含的重要意义就越少

## 基于语法的特征

- 不区别对待作为不同词性出现的同个词会导致错误
1. 当“抢”单独作为动词出现的时候一般是贬义词，从而使得评分在此情况下较低  
然而
  2. 当“抢”与“手”作为“抢手”形容词同时出现时，就可能是褒义词，从而评分也会相应不同

## 词与词的相关性

- 很多时候需要了解某个词是描述前后哪个词来理解文字信息的意义

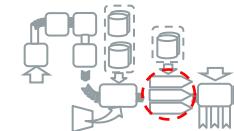
1. 好不错 里有 不 好
2. 坏不好 里有 不 坏

在上面的例子中，哪些词是相关的、哪个词是被其他词影响的等信息很关键

## 规律以及词语层面的评分

- 对于一些比较抽象的特征，我们需要找出能够确定
- Independent word level scores are another important way to find out if sentence has a negative pattern or a positive pattern. To get individual word level scores ontologies such as Sentiwordnet is being used.

# 通过词语或词组的使用频率得出的特征



词语或词组的使用频率是判断一条信息所属情感类别的关键

## 创建过程

词语、词组转化为表格的列项

每个词、词组的使用频率变成评分

### 文字

....这里的食物很差 服务也很差。没有什么是好的。整体都不爽 ....

### 所有可能的词语

....	这里	食物	差	服务	没	什么	好	整体	....

### 词语拼接

这里的食物	食物很	很差	服务也	也很差	没有什么	什么好	整体都	....

可以创建、使用更多拼接组合的词语

食物很好， 服务也很好。装修氛围差

...	2	1	1	1	...
...	好	差	食物	服务	...

食物很差 服务也很差。装修氛围好

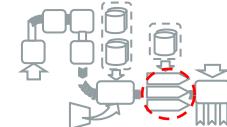
...	1	2	1	1	...
...	好	差	食物	服务	...

食物非常好， 服务也好。装修氛围也不差

...	1	0	0	1	...
...	很好	很差	不好	不差	...

食物很差， 服务业差。装修氛围也不好

...	0	1	1	0	...
...	很好	很差	不好	不差	...



# 基于语法的特征

词性等基于语法的特征包含关于理解文字意义的重要信息。同一个词可能在以不同词性出现的情况下有不同的情感意义；它还有可能在某种词性下不带任何特殊意义

## 特征

### 形容词

#### 带有定性信息

一个形容词常常用来描述、确认或者给词语定量。一个形容词往往有超过它所描述的名字或代词的意义

最好的, 短暂而  
无意义

## 范例

RESTAURANT had like the **best**  
ribs n da world  
**Short and pointless** night at  
RESTAURANT

### 副词

#### Carries qualitative information

A word or phrase that modifies or qualifies an adjective, verb, or other adverb or a phrase, expressing a relation of place, time, circumstance, manner, cause etc.

Never

Amazing. I **never** get tired of the  
broccoli cheese soup

### 动词

#### Carries experience oriented actions

A word used to describe an action, state, or occurrence, and forming the main part of the predicate of a sentence, such as hear, become, happen

Loved, loved

I **loved** the food and I **loved** the  
ambience. We had a gala time  
there.

### 介词

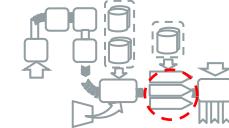
#### Modifies the meaning of other words

A word governing, and usually preceding, a noun or pronoun and expressing a relation to another word or element in the clause like on, by, to etc.

Without taste

The food was **without taste** and  
did not have the right ingredients

# Word Dependence Features



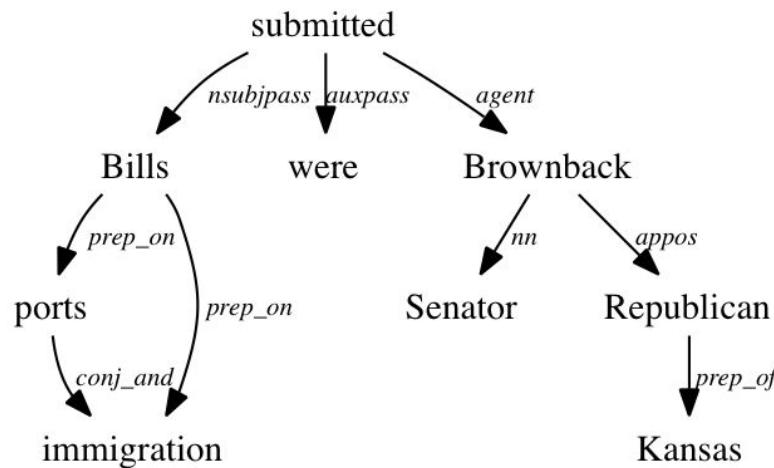
*How different words in a sentence are related to each other contains useful insights. We identify these relationships and convert them into features so that the overall machine-based classification improves*

## WORD DEPENDENCY

### SENTENCE

Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas.

### DEPENDENCY TREE



- **SOME DEPENDENCIES**
- nsubjpass(submitted, Bills)
- auxpass(submitted, were)
- agent(submitted, Brownback)
- nn(Brownback, Senator)
- appos(Brownback, Republican)
- prep\_of(Republican, Kansas)
- prep\_on(Bills, ports)
- conj\_and(ports, immigration)

# Compounding Features and Stopword Removal



We can create complex features by mixing, merging and cascading outputs of individual feature creation modules into each other. These are a few features which can be done using compounding

## COMPOUNDING FEATURES

### POS TAG PATTERNS

#### INTUTION

- These are more abstract features which try to capture difference in style, if present, in texts from different classes.

#### CREATION

1. Create grammatical features
2. Retain part of speech tags
3. Run n-grams creator

### NEGATION HANDLING

#### INTUTION

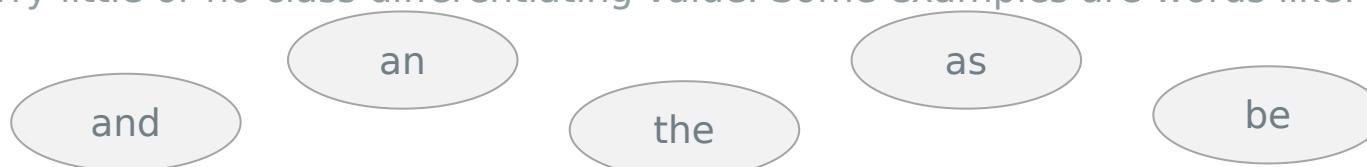
- This feature captures negation of qualitative words such as  
*NOT good, WITHOUT honour*

#### CREATION

1. Create grammatical features
2. Dependency analysis
3. Retain the NOT ones

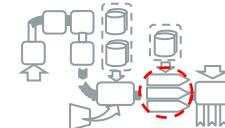
## STOPWORD REMOVAL

Is the process of removing a pre-specified list of words from the feature set as they carry little or no class differentiating value. Some examples are words like:



Stopword list used for sentiment analysis differs from the stopword list used for general text classification which ignores many prepositions useful for identifying the right sentiment polarity

# Feature Selection: Techniques Used



*Even a small amount of text has thousands of unique words, resulting in a large number of features. We use two techniques for reducing the size of the feature set to find relevant attributes*

## TFIDF BASED ATTRIBUTE SELECTION

TFIDF is an abbreviation for “Term Frequency, Inverse Document Frequency” and is a commonly used measure to identify words significant or representative of a particular class

### FORMULA

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

$$(tf-idf)_{i,j} = tf_{i,j} \times idf_i$$

### PROCESS

1. Find TFIDF of words for each class
2. Pick up the top TFIDF words of each class
1. Prepare sorted list mixing sorted list of words from each class
2. Select the top n words from this

In TFIDF based selection of attributes we select the words which reject words with low TFIDF score.

## FREQUENCY BASED ATTRIBUTE SELECTION

This approach selects a minimum needed frequency threshold on words and all words below the threshold are rejected. We reject words which are too specific to a text instance and don't generalize well

# Classification



*Classification is the process of training a computer to identify sentiment, using the features and the training data created in earlier steps*

## CLASSIFIERS USED

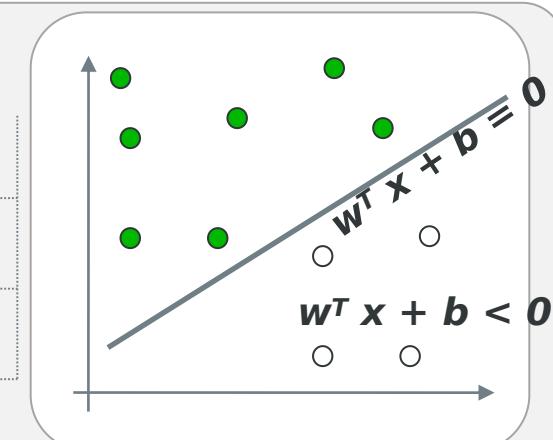
Support Vector Machines are discriminative class of classifiers and one of the most powerful tools available for classification.

	Opinion	Neutral		+ve	-ve
Opinion	1,741	171	+ve	1,355	68
Neutral	494	385	-ve	185	304

### RESULTS

#### ACCURACY

ACHIEVED = 78%



## Naïve Bayes

Naïve Bayes is a generative class of classifiers which constructs the probability of an instance coming from a class using Bayes' Rule

	Opinion	Neutral		+ve	-ve
Opinion	1426	388	+ve	1216	155
Neutral	305	563	-ve	42	424

### RESULTS

#### ACCURACY

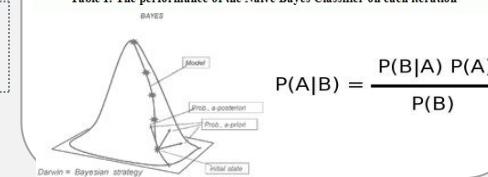
ACHIEVED = TBD

All results are from a two-step classification procedure using 5 fold cross validation.

Data set Size: 2780

Features Used	Number of Features	Classifier Accuracy	Sensitivity	Specificity	AUC	Brier Score
All features	8	0.7590	0.8000	0.6828	0.8334	0.3388
Removal of '2-hour Serum Insulin'	7	0.7474	0.7690	0.6567	0.8335	0.3346
Removal of 'Number of times Pregnant'	6	0.7512	0.8100	0.6418	0.8357	0.3278
Removal of 'Diastolic blood pressure'	5	0.7656	0.8300	0.6455	0.8421	0.3152
Removal of 'Diabetes Pedigree Function'	4	0.7669	0.8240	0.6604	0.8411	0.3187
Removal of 'Triceps Skin Fold Thickness'	3	0.7790	0.8640	0.6306	0.8471	0.3072
Removal of 'Age'	2	0.7630	0.8460	0.5746	0.8233	0.3266

Table I: The performance of the Naïve Bayes Classifier on each iteration



# 持续倾听、数据抓取



我们能实时监测媒体/博客的信息；我们持续地抓取数据并对非结构性数据进行整合，然后再将其送入能挖掘规律的引擎，以得到相应的解决方案

确定需要搜集的话题和关键词



## 元数据

### 领域

- 频率
- 评论数
- 作者类型
- Google网页

### 作者

- 最近是否访问
- 感情
- 朋友和粉丝
- 朋友圈扩大

### 社区

- 点赞
- 转发/发帖
- 感情
- 话题

## 客户感受

### 影响力分析 顾客反应

### 品牌受欢迎程度

% 不满意

最严重的投诉

最受欢迎的品牌

最受欢迎的功能

% 开心

# 持续的倾听，收集，整合数据

我们可以把外部网络数据与内部“未挖掘的”数据结合起来，来判断顾客的行为，偏好以及心理；企业从而可以采取服务，产品，库存，布置等等不同方面的行动

## 社交网络（外部）



## 企业内部(内部)



客服中心



在线聊天



邮件



开放式调查  
问卷和反馈

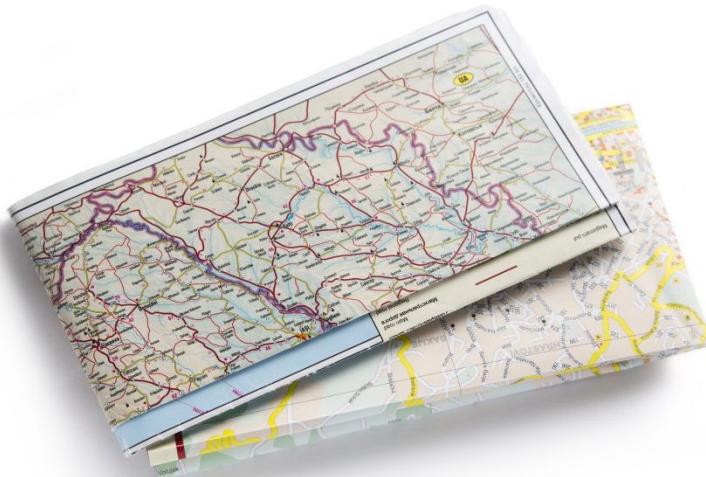
一个对顾客的行为，观点，  
和偏好开放的窗口

我去了XYZ spa-真是太棒了，最棒的是它离我家很近并且从早八点开放至晚八点

SideDoor提供了‘Tebla’餐厅真实太好了，但是我一天前从Groupon上拿到了同样的优惠…

希望SideDoor提供更多的刺激的体育活动…





# 地址选择

# 数据要求

除了客户提供的数据，我们还通过公共资源，网络数据抓取和文本分析来收集位置数据。

数据来源	示例变量/描述	细节层次
零售商店	<ul style="list-style-type: none"> <li>每个分类的销售量(e.g. 总数, 化妆品, 防晒, etc)</li> <li>每个种类的销售单位数量</li> <li>店铺大小, 位置(e.g. 市中心, 搬家, etc)</li> </ul>	商店 商店 商店
药店 (药剂师)	<ul style="list-style-type: none"> <li>每种药品的购买数量 (示例. 总量, 苯丙胺, 咳嗽及感冒用药等)</li> <li>支付方式 (示例. 现金, 医疗保险, PTP)</li> </ul>	商店 商店
诊所	<ul style="list-style-type: none"> <li>客流量</li> <li>历史记录</li> <li>顾客满意程度</li> <li>员工满意程度</li> <li>市场意识</li> <li>国家保险环境 (示例: 有医保的客户人数)</li> </ul>	商店 商店 市场 商店 市场 商店
公开数据	<b>示例变量:</b> <ul style="list-style-type: none"> <li>人口普查</li> <li>劳动统计局数据</li> <li>工商局数据</li> <li>旅行和交通数据</li> <li>凯撒家庭基金会</li> </ul>	邮政编号 省/邮政编号 邮政编号 邮政编号 省
网络数据抓取和文本分析	<ul style="list-style-type: none"> <li>竞争者密度(e.g. 小诊所, 急诊, 内科医生)</li> <li>竞争者的实力和表现(e.g. 竞争者的平均等待时间)</li> <li>附近的邻居(e.g. 咖啡店, 学校, 公园, 杂货商店 等)</li> </ul>	商店 商店 商店

# 关键变量

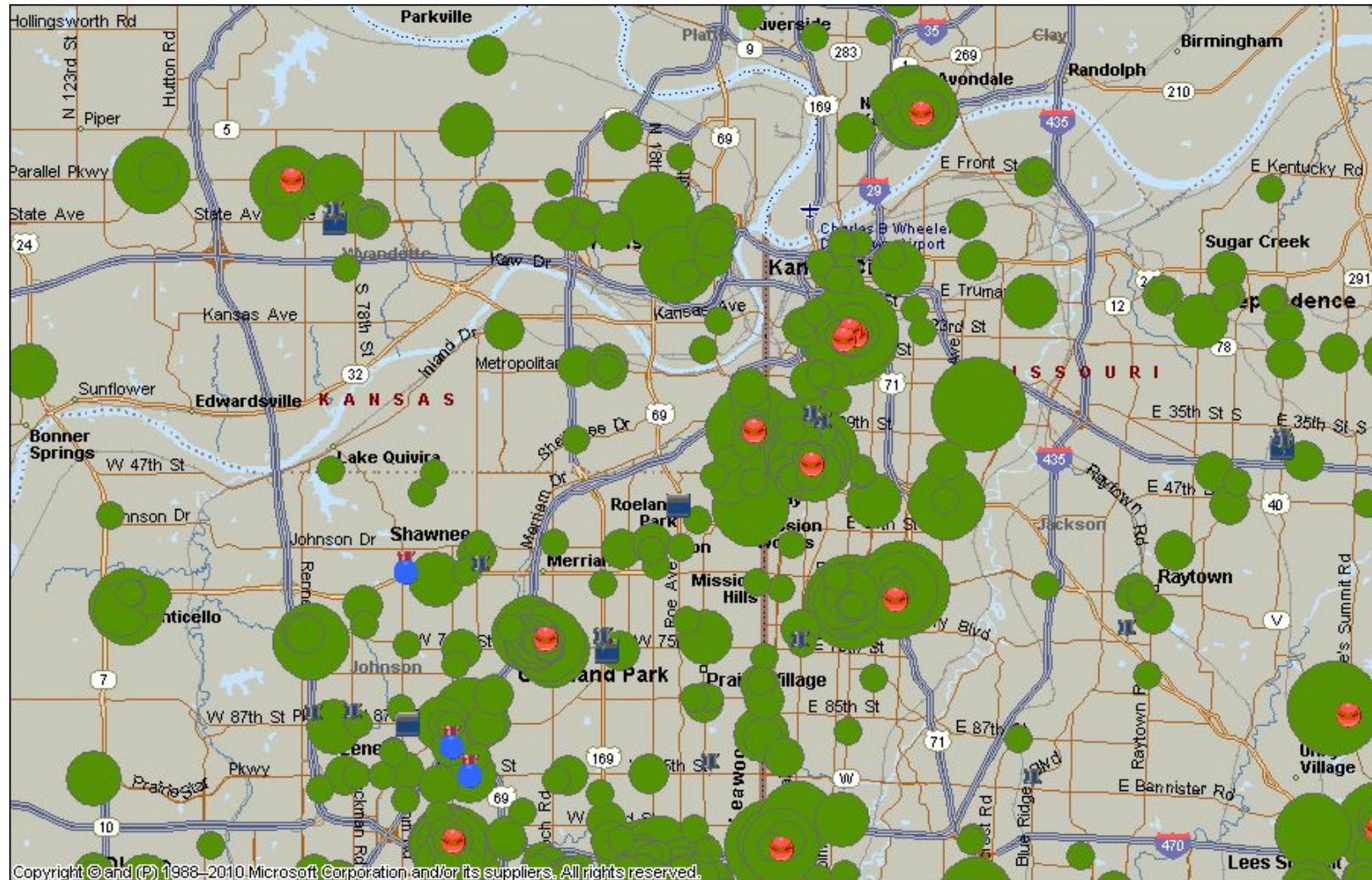
*130 variables contributed to the final model which collectively generated significant predictive power.*

Variable Category	Description / Sample Variables	Level of Importance	Relationship	Level of Detail	Attributes of Success
Insurance Environment	<ul style="list-style-type: none"> <li>Total Population Uninsured</li> <li>% of total PPO Network lives contracted by Client</li> <li>HMO Penetration (%)</li> <li>% of total Commercial lives contracted by Client</li> <li>% of total insured lives contracted by Client</li> </ul>	Ultra High Ultra High Ultra High Ultra High Ultra High	Negative Positive Negative Positive Positive	State State State State State	<b>High contract coverage and low HMO penetration</b>
Population Demographics	<ul style="list-style-type: none"> <li>Caucasian population concentration</li> <li>Loyalty Card household income &gt;125K</li> <li>BA and above education</li> <li>Population with Food Stamps</li> <li>Households Per Zip Code</li> <li>High School and above education (%)</li> <li>Total Population within 2 mile radius</li> <li>Population age under 5 (%)</li> </ul>	Ultra High Ultra High Ultra High Ultra High High Moderate - High Moderate Moderate	Positive Positive Positive Negative Positive Positive Positive Positive	邮政编码 商店 邮政编码 省 邮政编码 邮政编码 邮政编码 邮政编码	<b>High Concentration of Affluent, Educated and Young Families</b>
Front Market Sales Volume	<ul style="list-style-type: none"> <li>FS Sales (Suncare, Nicotine Replacement, Giftcards)</li> <li>FS Sales (Allergy/Cold, HairCare, Cosmetics, SkinCare)</li> <li>Total 52 Week FS Sales</li> </ul>	Ultra High Ultra High High	Positive Positive Positive	商店 商店 商店	<b>High FS Sales Volumes - Overall and Select Categories</b>
Neighborhood	<ul style="list-style-type: none"> <li>Maximum delta in monthly avg. temperature</li> <li>Prevalence of Coffee Shops</li> <li>Active residential delivery mailboxes</li> <li>Parks, Family Restaurants within 2 miles</li> <li>ER and Retail Clinic Concentration within</li> </ul>	High Moderate-High Moderate Moderate	Positive Positive Positive Positive	邮政编码 商店 邮政编码 商店 邮政编码	<b>Strong Seasonality and Neighborhood Affluence Indicators</b>

# 市场评估 - 竞争者



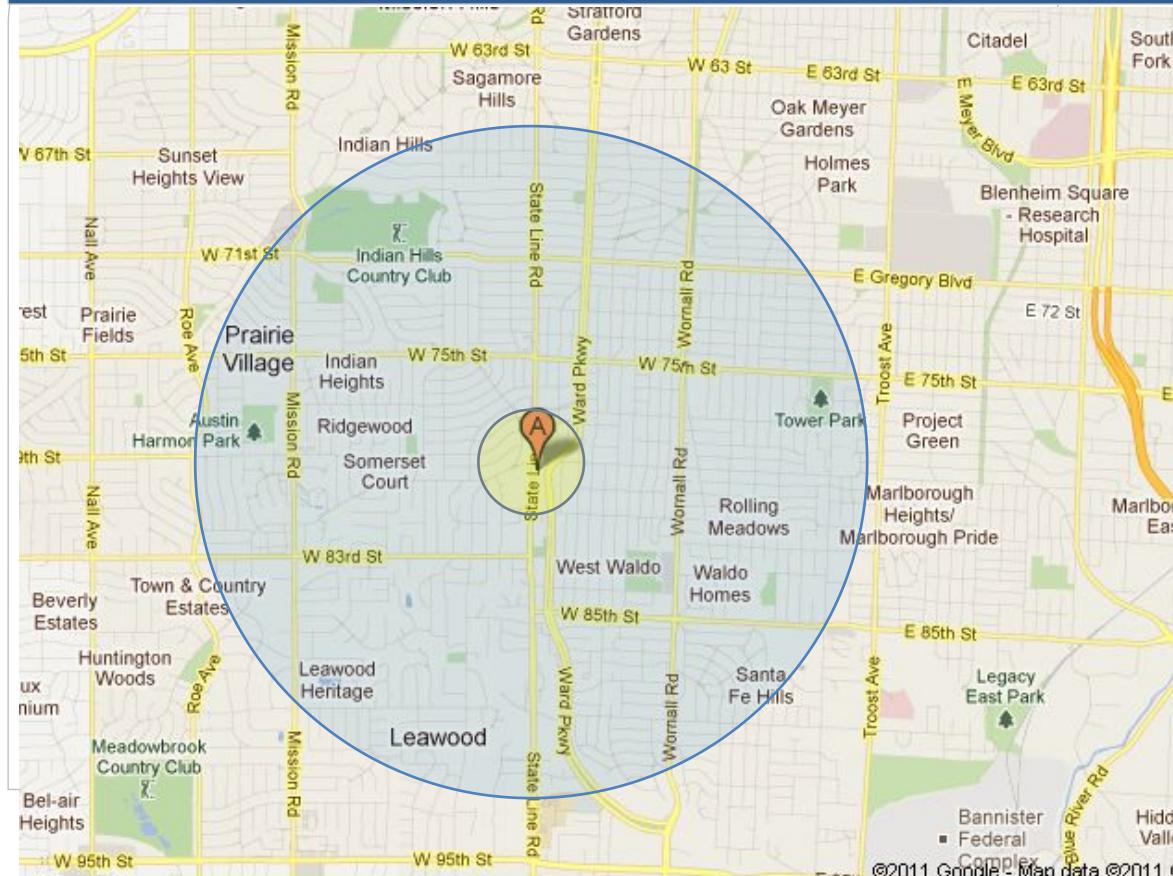
我们计算了客户与竞争店铺之间的驾驶时间来把竞争形式特征化



# 市场评估 - 地区的微观档案



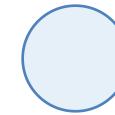
Client shop, Kansas City, MO



## 示例变量

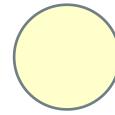
### 地区资料

- 收入
- 年龄
- 民族
- 房屋价值
- 供应商 (MD's, ER's, 急救, 零售)
- 保险
- 学校
- 老年中心



### 微观资料

- 高档咖啡店
- 折扣商店
- 交通站
- 家庭餐厅
- 快餐



# 理解地区微观经济

每个交易区的人口，经济，以及其他汇总的特征可能会表现出和过去的表现很重要的相关性。

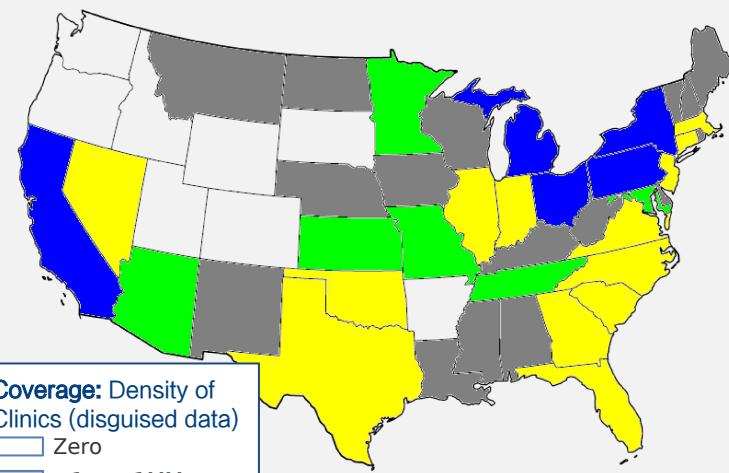
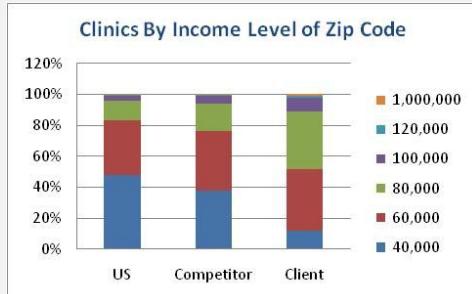
## 商业区 - 顾客

## 商业区 - 竞争者

## 市场

当地的微观经济：一些变量，比如收入级别，房屋价值，失业率，当地的借贷率数据库，拥有社区或私人健康保险的人数比例, **number of people per PCP, ER的人数,人口密度**等等，可以对现在的诊所分布的绩效和预测新的最佳地点提供宝贵的参考和见解。

*Current footprint\*: Clinic locations and comparative economics*



# 倾向模型和竞争影响地图

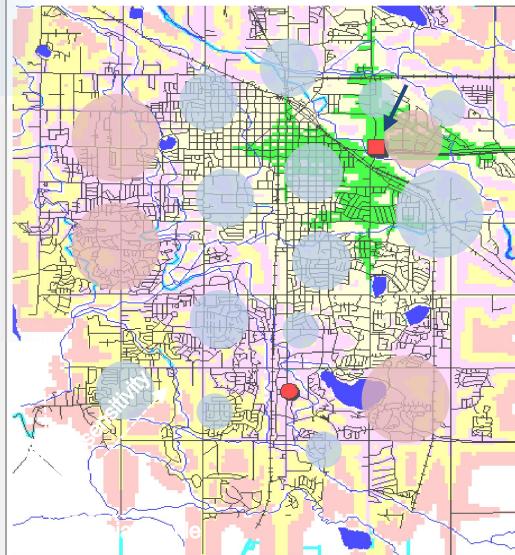
空间分析可以帮助理解那些可以影响诊所潜在的需求的因素，并且便于记录----在学习阶段，我们的目标是建立这些分析的预测值。

商业区 - 顾客

**服务需求:** 为每条服务线的周边地区，设定慢性的或者急性的倾向分数。

商业区 - 竞争者

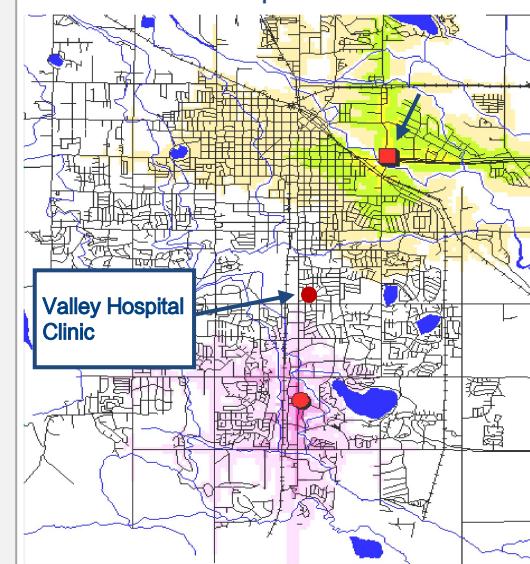
Etc  
Respiratory Care  
Rehabilitation  
Long-Term Acute Care



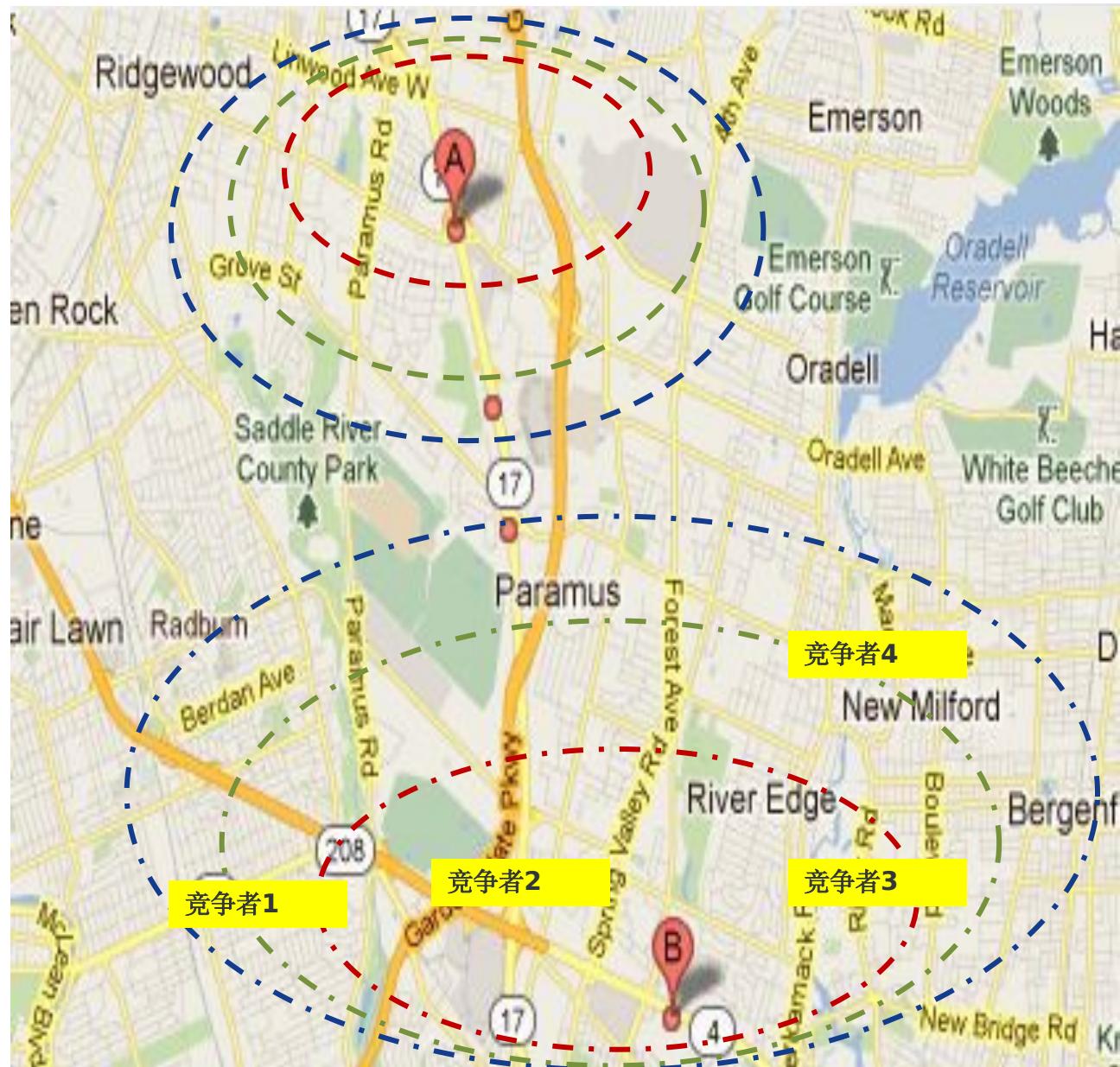
市场

**COMPETITOR GRAVITY:** 对于每条服务线以及相关的直接、非直接的竞争者，分数影响了目标客户群。

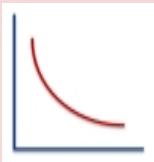
Etc  
Ambulatory Care Centers  
Stand-Alone Rehab Centers  
Acute Care Hospitals



# 虚拟价格区



销售点价格弹性



顾客行为分析



人口 &  
竞争者密度



# 市场影响



*Effectiveness of prior marketing campaigns need be considered to ensure appropriate contribution of area, shop and customer factors in predicting Clinic success .*

商业区-顾客

**市场影响:** Impact, channel and target of prior campaigns need be normalized to properly account for marketing efforts in the success of each clinic and to ensure that the predictive factors for success are not biased by asymmetrical marketing investments

商业区-竞争者

**Avg. Patient Census**

9/1 - 9/30

市场

**Site 223: One-Month Promotion**

7/1/08

7/1/09

7/1/10

10/1 -

**Site 1327: Open-Ended Direct Mail Campaign**

7/1/08

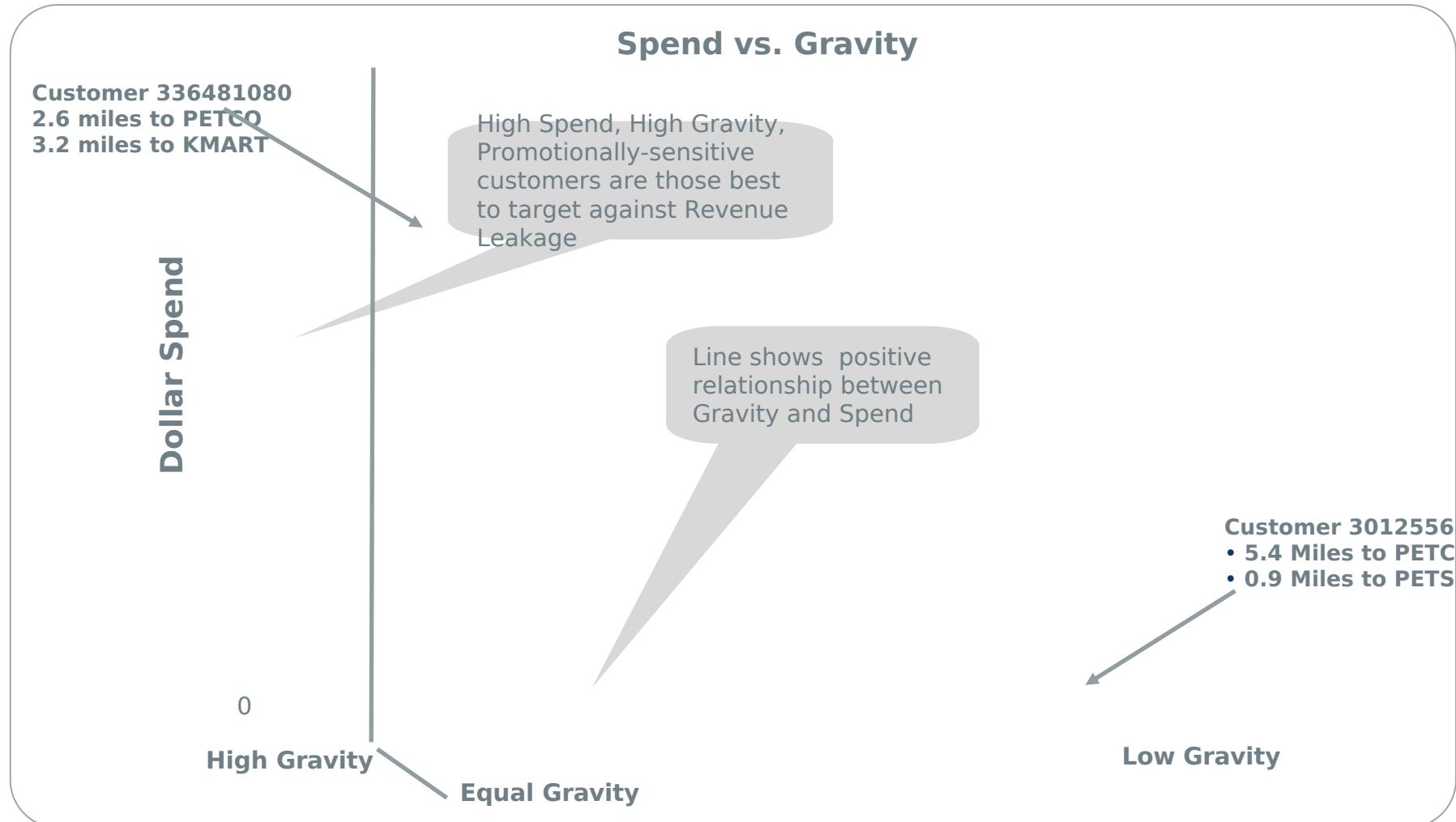
7/1/09

7/1/10

# Gravity effect is significant on Household Spend

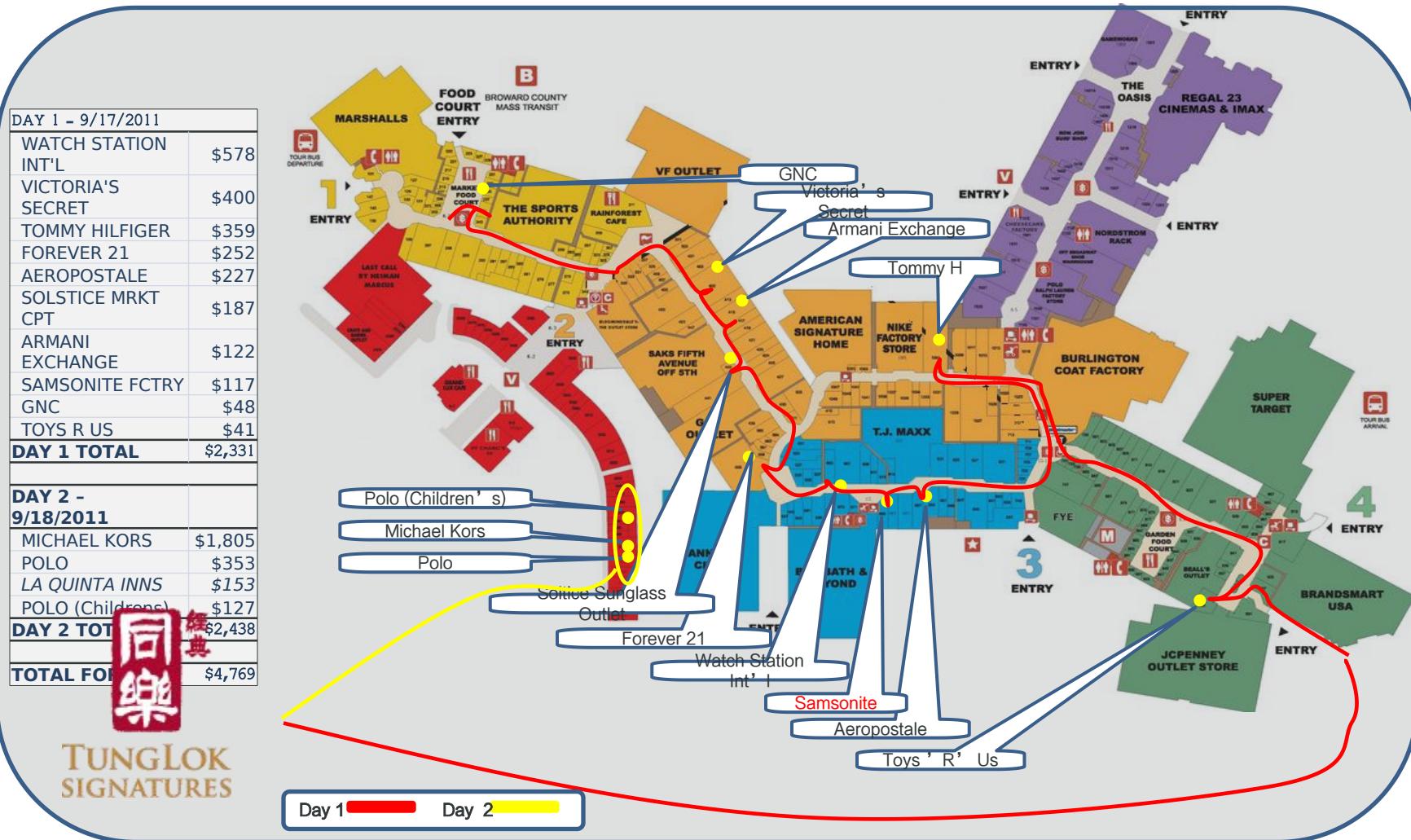


*Data shows that customers who live in High Gravity locations spend more than those with a lower gravity moreover have different promotional response profiles*



# 购物中心内部的店址选择

通过数据库里顾客的宾馆，行李花费，以及仅仅活动了两天作为特征，可以发现这个商场的顾客多数为“目标明确的购物者”。隔离了那些能暗示顾客行为的信号，我们可以更好的分割定位更多的人。



# 影响品质的因素

## 批发商店的店龄



**老店**

店龄评分 = 5

**新店**

店龄评分 = 9



## 标识是否醒目



**标识不醒目**

标识评分 = 3

**标识醒目**

标识评分 = 8



多种因素会影响顾客对门店的主观体验和客观需求

基准线 80%

3%

7%

5%

各因素对需求的影响权重

## 整洁度



**脏乱的门店**

整洁度评分 = 5



**干净的门店**

整洁度评分 = 8

## 驾车进入是否方便



**不方便**

方便评分 = 4



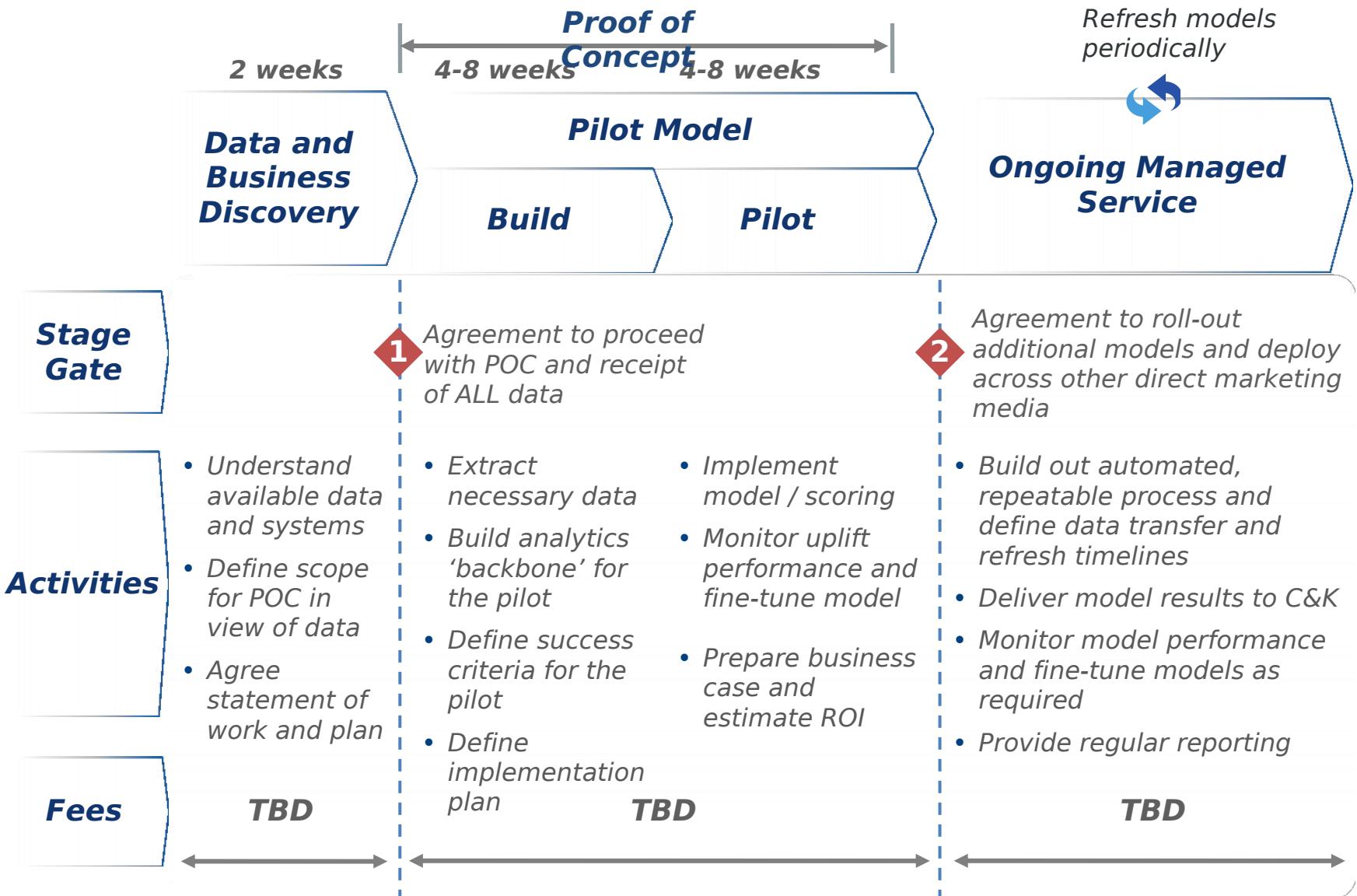
**方便**

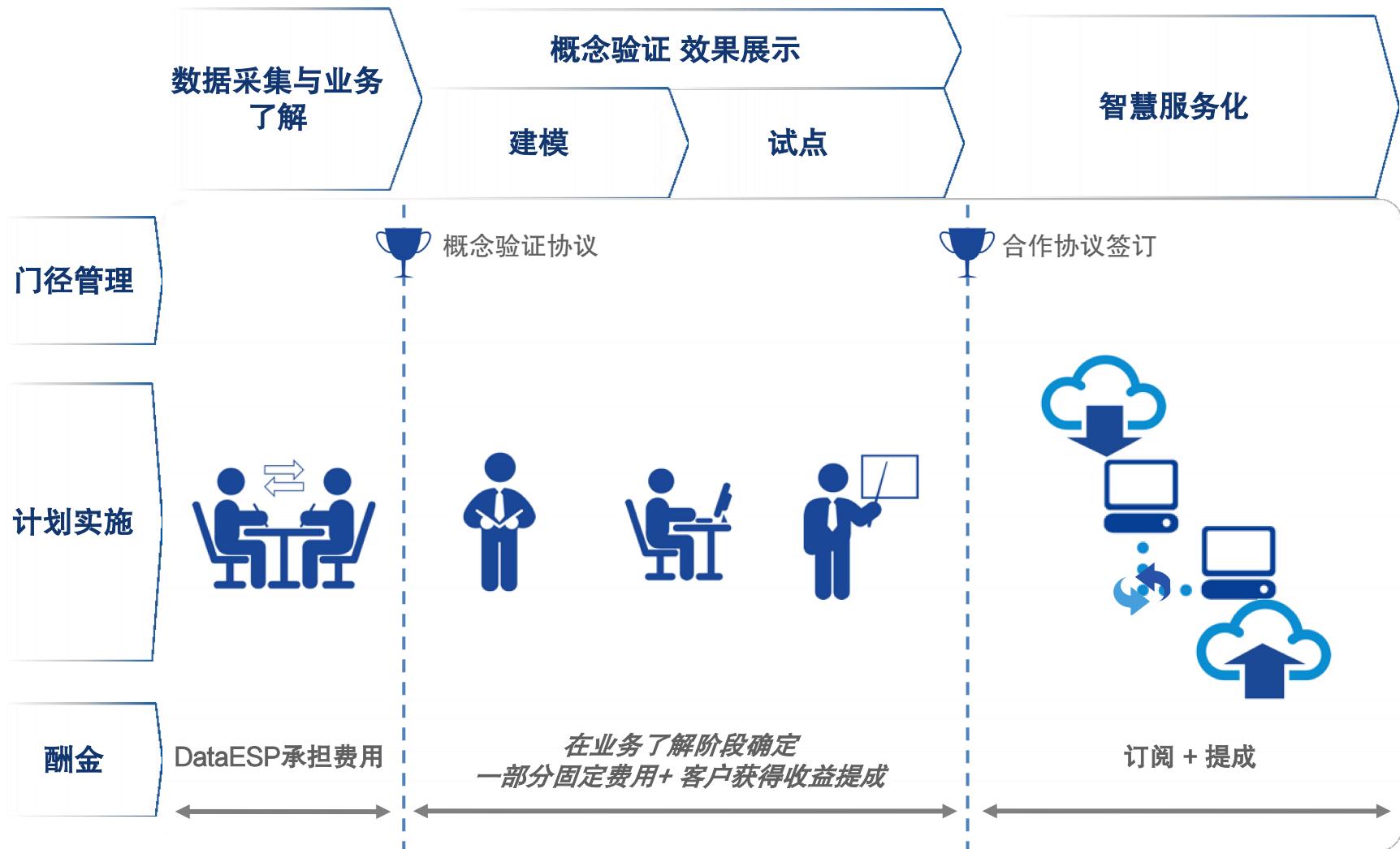
方便评分 = 7

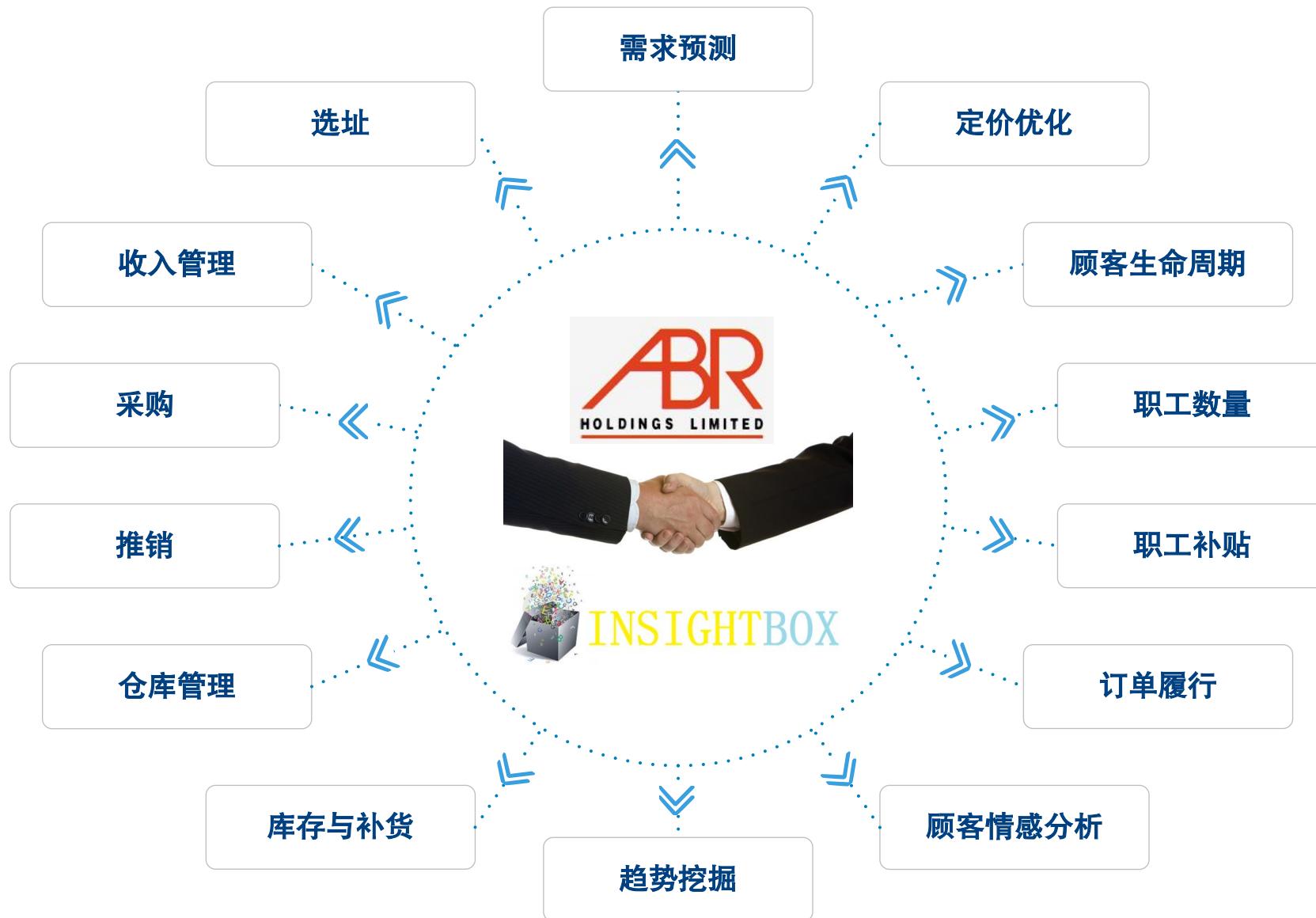


下一步

# Sample Engagement Plan









问题?

期待与您的合作!





# Appendix

# Hypothesis Generation - Competitive Advantage



*Our ability of exhaustively scanning the problem-solution space to identify unusual, statistically significant patterns in arbitrarily large database sets us apart from all competitors. The same task that will take our competitors years to accomplish with tools such as OLAP can be done in seconds using DataX-ray*

## Questions



As the availability of data (3 V's) approaches infinity:

- Correlation → Causality
- OLAP<sup>1</sup> usability → 0
- OLAP + DataX-Ray usability stays constant

1. OnLine Analytical Processing: [http://en.wikipedia.org/wiki/Online\\_analytical\\_processing](http://en.wikipedia.org/wiki/Online_analytical_processing)

# DataX-Ray Technical Highlights



*DataX-Ray is a few order of magnitude more effective than conventional techniques for uncovering hidden knowledge within large data stores*



## Automated

Automatic hypothesis generating and validating

- § Complex Pattern Discovery
- § Data mining with noisy and/or incomplete data
- § Efficient search space truncation techniques



## Exploratory

Exploratory pattern scanning

- § Fastest high order statistical significant association detection in the world
- § Flexible pattern representation - hypergraphs, rules, associations and spreadsheet
- § Pre and post-processing of patterns and rules



## Transparent

Easy to understand and for domain validation

- § Ideal for engineering optimization
- § Suitable to build domain specific expert systems and guidance system
- § Speed up deployment in large enterprises and complex environments



## Customizable

Flexible implementation and customization

- § Based on XML and other industry standard technologies
- § Architecture lends itself to expansion and flexibility
- § Object-oriented architecture and web application

# Model-T Solutions: Input Output Specifications



Demand  
Forecasting



Churn  
Reducti  
on



Price  
Optimizat  
ion

