TERADATA UNIVERSE
BEIJING MAY 6 2016

数据仓库 开源融合 极致演绎

2016 | Teradata 大数据峰会

以手机网络数据获取匿名预付费用户群的基本信息及使用偏好以进行细分营销

David Bloch – 分析及数据战略经理

提纲

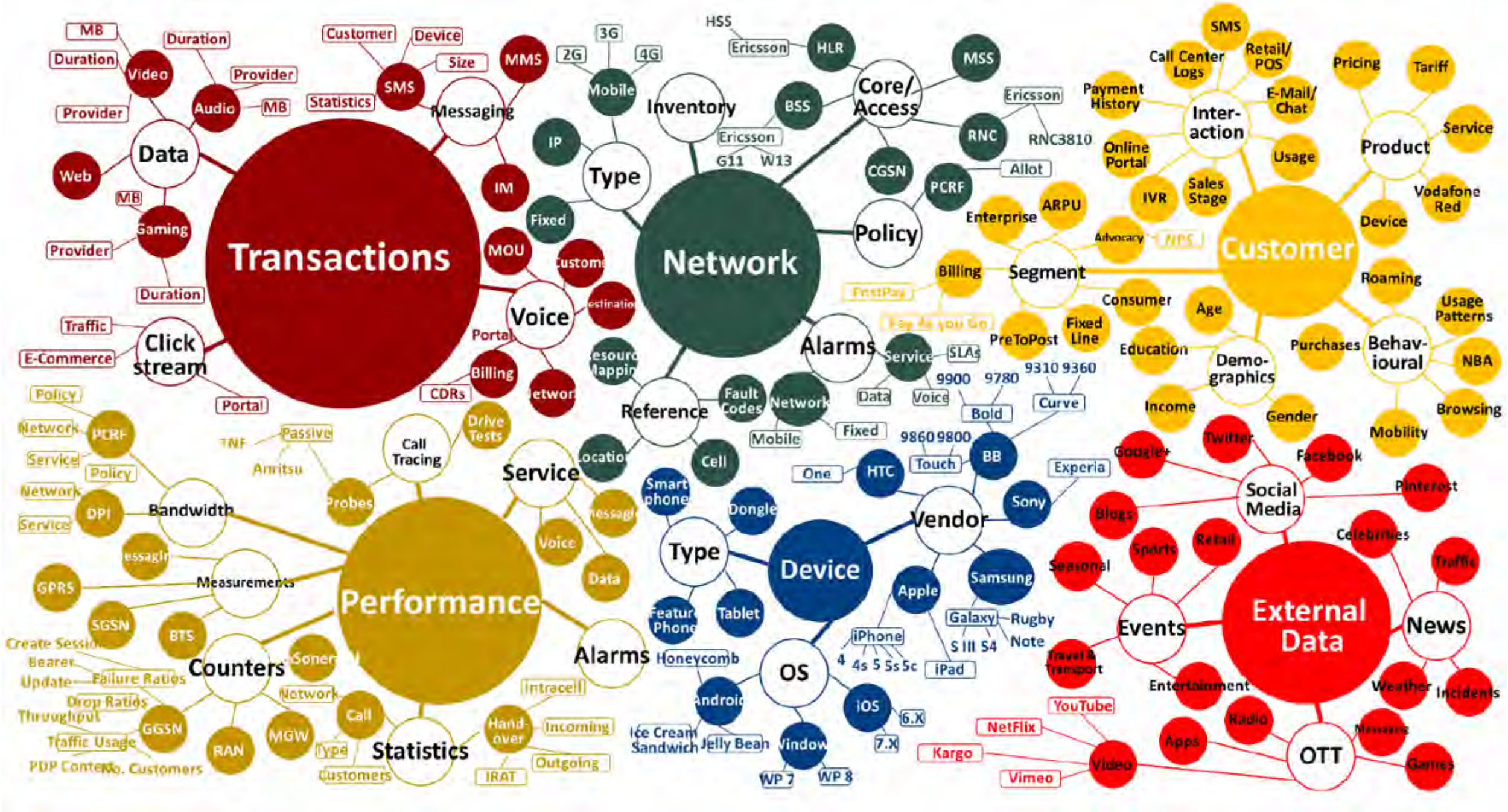
TERADATA UNIVERSE
BEIJING MAY 6 2016

- **简介**
  - 电信数据状况
  - 我们对大数据和高级分析的观点
  - 我们的大数据和高级分析平台
- **利用大数据识别匿名客户**
  - 问题陈述
  - 假设
  - 寻找行为数据
  - 数据量
  - 识别模式——年轻客户
  - 创建属性列表
  - 选择建模类型
  - 结果
- **结论**
  - 从流程中学到的经验教训
  - 问题和答案

简介

David Bloch
分析及数据战略经理
沃达丰新西兰公司

- **16年分析、商业智能和软件工程经验**
- **4年大数据生态系统构建经验**
- **曾在医疗保健、快速消费品、公共事业、媒体、广告和电信行业就职和提供咨询服务**
- **在沃达丰消费者部门领导由六名职业分析师组成的团队。**

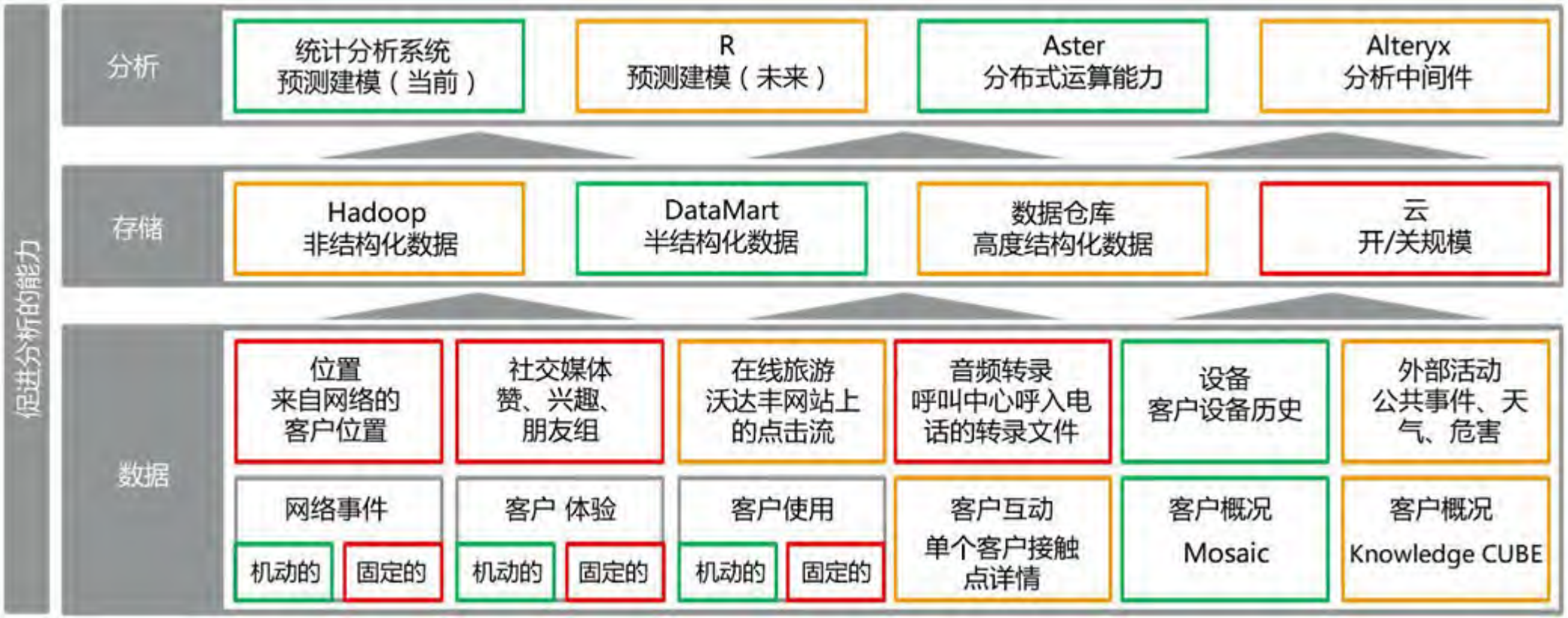

Transactions
Data
Video
Audio
Gaming
Web
MB
Duration
Provider
Click stream
Traffic
E-Commerce
Portal
Messaging
SMS
MMS
IM
Customer
Device
Size
Statistics
Voice
MOU
Billing
CDRs
Portal
Network
Inventory
Mobile
2G
3G
4G
Type
IP
Fixed
BSS
Ericsson
G11
W13
HSS
HLR
Core/ Access
MSS
RNC
Ericsson
RNC3810
CGSN
PCRF
Allot
Policy
Alarms
Service
SLAs
Data
Voice
9900
9780
9310 9360
Bold
Curve
Reference
Fault Codes
Network
Mobile
Fixed
Cell
Performance
Bandwidth
PCRF
DPI
Policy
Network
Service
Call Tracing
Drive Tests
Passive
Anritsu
TNF
Probes
Measurements
GPRS
SGSN
BTS
Counters
Create Session
Bearer
Update
Failure Ratios
Drop Ratios
Throughput
Traffic Usage
PDP Contexts
No. Customers
GGSN
RAN
MGW
Soner
Network
Call
Type
Customers
Statistics
Service
Voice
Data
Alarms
Hand-over
Intracell
Incoming
Outgoing
IRAT
Device
Type
Smart phone
Dongle
Feature Phone
Tablet
OS
Honeycomb
Android
Ice Cream Sandwich
Jelly Bean
Window
WP 7
WP 8
iOS
6.X
7.X
Vendor
HTC
One
BB
Touch
9860 9800
Sony
Experia
Apple
iPhone
4
4s 5 5s 5c
iPad
Samsung
Galaxy
S III S4
Rugby
Note
Customer
Inter-action
SMS
Call Center Logs
Retail/ POS
Payment History
E-Mail/ Chat
Online Portal
IVR
Sales Stage
Usage
Product
Pricing
Tariff
Service
Vodafone Red
Device
Segment
Enterprise
ARPU
Advocacy
NPS
Billing
FirstPay
Pay As You Go
PreToPost
Fixed Line
Consumer
Age
Education
Demo-graphics
Income
Gender
Behav-ioural
Roaming
Usage Patterns
Purchases
NBA
Browsing
Mobility
External Data
Social Media
Google+
Twitter
Facebook
Pinterest
Blogs
Sports
Retail
Seasonal
Events
Travel & Transport
Entertainment
Celebrities
Traffic
News
Weather
Incidents
OTT
YouTube
NetFlix
Kargo
Vimeo
Video
Apps
Radio
Messaging
Games

我们对大数据和高级分析的观点

大数据带来为客户和业务创建新价值的能力，用户对呈指数式增长的数据集进行捕捉、储存和分析，同时视隐私、许可和安全为第一要务

将数据转变为行动	利用数据为客户服务	利用数据驱动市场
从核心业务提炼价值 （营收提升、成本效率和网络优化）	**推出创新产品和服务** （基于位置的服务、基于产品和个性化的实时数据）	**数据货币化** （面向第三方的数据分析解决方案）

我们的大数据和高级分析平台



问题陈述

新西兰的预付客户无需登记任何详细信息或提供身份信息即可购买预付费Sim卡。

虽然我们可以通过客户在本公司网络中的流量消耗来瞄准客户，但我们不能识别他们的人口统计学分组，而这也对我们在向预付费客户提供相关产品和服务方面的能力造成限制

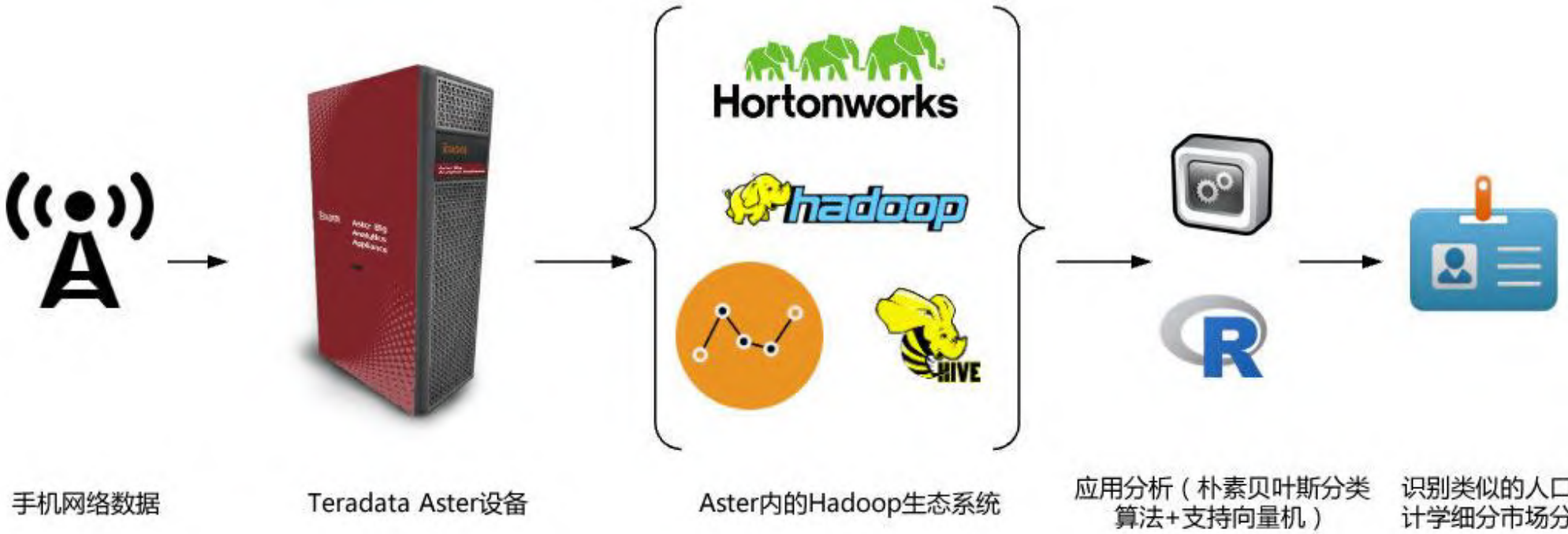
假设

如果我们利用合约客户建立人口统计学分组，分析各个组别的行为模式，我们就可以使用这些模式识别匿名客户的类似人口统计学分组。

如果我们探查16至24岁之间的所有已知客户，了解他们使用的应用类型、使用时间及他们消耗的流量，我们就可以预测有多少匿名客户属于这个年轻组别

假设

行为数据捕获


Hortonworks
hadoop
HIVE
R
手机网络数据
Teradata Aster设备
Aster内的Hadoop生态系统
应用分析（朴素贝叶斯分类算法+支持向量机）
识别类似的人口统计学细分市场分组

捕获数据量

4 **分钟**，这是从事件发生到事件进入本公司BD环境中的时间距离

260 **万**个产生本公司网络数据使用流量的连接

85.4 **万**个不同ID地址得到用户访问

384 **亿**次用户应用会话

2016年3月至4月期间，四个星期的数据捕捉

识别具有预测价值的潜在数据来源

通过本公司账单平台收集的记录信息系统

示例：

- 帐单历史
- 设备类型
- 在网时间
- 购买套餐
- 平均营收
- 充值频率
- 年龄（如已知）
- 性别（如已知）
- 总用量

通过我们的大数据平台收集的机器生成数据

示例：

- 应用使用（如微信、新闻）
- 通话圈和呼叫模式
- 每次Web会话的上传/下载量
- Web会话的当日时间
- 分类使用情况（社交媒体vs新闻网站）

识别样本模式——年轻客户

年轻客户的夜间数据使用量高于非年轻客户

年轻客户在微信等消息应用上的会话数量（即所使用的总时间）更高

年轻客户在微博等社交网站的上传量更高

年轻用户充值频率更高、单次充值金额更低

年轻客户的通话圈更大，比非年轻客户联系的人更多

年轻客户更可能参与通过短信渠道开展的优惠和促销

创建属性列表，使数据价值标准化

195 **个潜在属性**，这是我们在记录系统和机器生成的数据中识别的、具有行为价值的客户人均潜在属性数量

按属性类型识别**价值变量**。

- **帐单信息**——总价值/中间值，得出利用率
- **使用分类信息**——分类应用点击总量/分类应用点击中间值
- **特定应用信息**——应用点击总量、下载总量、上传总量/信息中间值，按当日时间划分

标准化数据样本视图

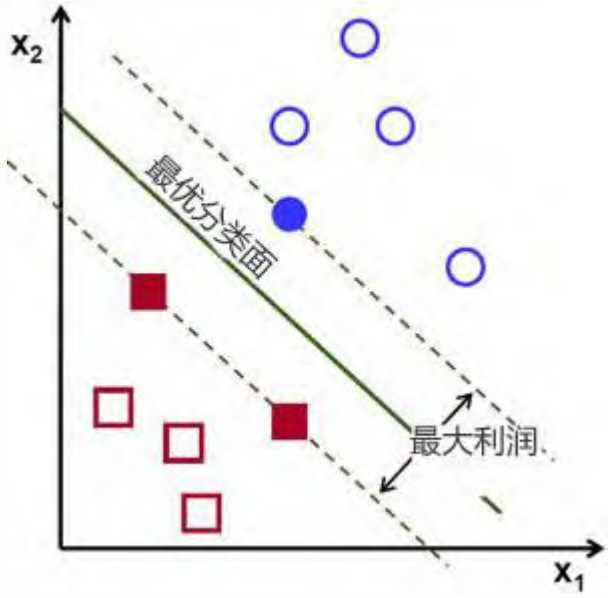
通过本公司账单平台收集的记录信息系统

识别因素	属性	标准化价值
节选	智能手机标记	1.000
节选	3G标记	0.000
节选	4G标记	1.000
节选	手机在网时间	0.258
节选	通话分钟数	1.236
节选	短信总计	0.375
节选	流量总计	1.281
节选	所购买流量套餐	0.000
节选	所购买的通话套餐	0.000
节选	每用户平均收入	0.433
节选	亚马逊在线商店	0.010

通过我们的大数据平台收集的机器生成数据

识别因素	属性	标准化价值
节选	信息（10）	8.626
节选	在线商店（7）	0.050
节选	电子邮件（3）	6.263
节选	流媒体（9）	3.446
节选	社交网络（8）	1.739

建模类型——支持向量机



支持向量机是一种有人监管的学习模型，允许使用各种类型、各种维度的多个变量

它结合横跨多类变量的关联、分类和回归分析，对利用识别因素、数据和价值字段的强大模型进行训练，然后将之用于预测

结果

78%至88% 利用多个已知客户样本集合和经过训练的模型对整个客户群体进行分析的预测高度准确性

98% 模型具有较高的预测置信度（大于80%）时的预测准确性

81% 在客户年龄组方面有合理置信度时对整个预付费客户群里进行预测的准确性

170k 在匿名预付费客户群中呈现活跃状态的年轻客户大约总数

经验教训

TERADATA UNIVERSE BEIJING MAY 6 2016

1. 行为数据的预测潜力，特别是当日时间和上传量为用户提供有关预付费客户群中类似年轻客户的深入洞见

2. 通过将支持向量机与通话圈分析相结合，我们得以进一步提升预测准确性

3. 收集更长时间的数据，利用更多的已知年轻客户案例训练模型，有助于提高预测准确性

TERADATA UNIVERSE
BEIJING MAY 6 2016

谢谢！