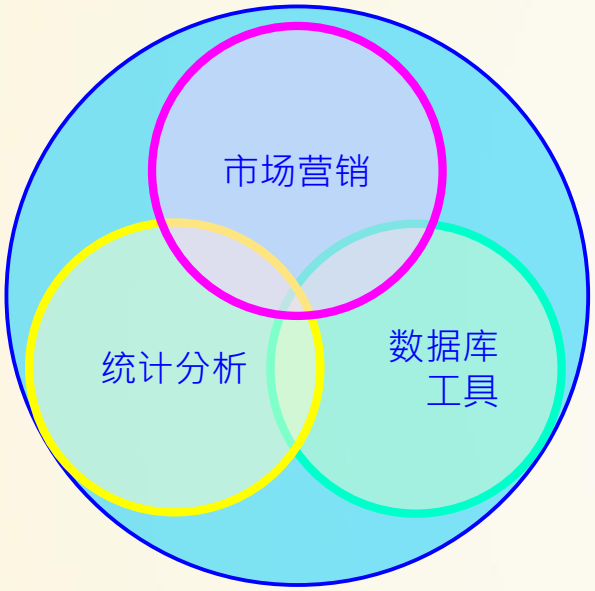
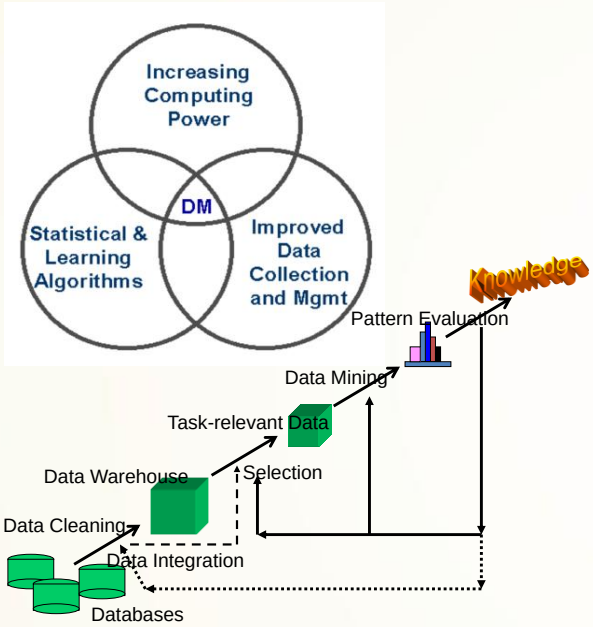
DTCC 2014中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2014
大数据技术探索和价值发现

数据分析漫谈
- 实践与个案分享 -

杜长嵘/ Lionel Tu
2014/01

那些年我们一起追过的数据





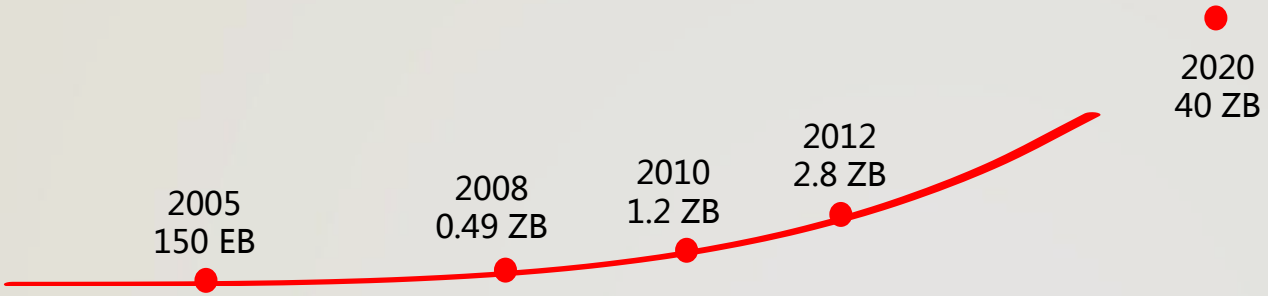
1995 **数据分析/数据库营销**

2000 **数据挖掘**

2012 **大数据**

那些年我们一起追过的数据

数据宇宙



数据源

调查研究
政府机关普查
水电工共单位

互联网
金融保险
医疗交通
电信/石化

移动设备
智能电视/电表
穿戴式设备
传感器/RFID
生物基因

统计分析方法

次数分配表
交叉分析

回归分析
时间序列
多变量分析
决策树

文本挖掘
机器学习
神经网络

分析工具

Excel
SPSS
SAS

Splus
Minitab

SQL
R

什么是数据分析？

数据分析：给个定义

数据分析是指用适当的统计方法对收集来的大量第一手资料和第二手资料进行分析，以求最大化地开发数据资料的功能，发挥数据的作用。是为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Understanding individual customers lifestyles, needs and preferences, their loyalty to brands and responsiveness to communications in order to make marketing recommendations, at an overall strategic level or versus particular activities. Including 2 areas, measurement & analytics

Create value to consumers and marketers by delivering the consumer-centric data platform and insight services that maximize user engagement and enable innovative marketing solutions

数据分析不是做学术研究，而是要从数据中看到、发现Insights，提炼为information，内化为knowledge，并将其转化为具体可行的行动方案，最终对公司产生正面价值的影响

数据分析两核心理念

大多数人都会抱怨：我们有 terabytes 的数据，gigabytes 的报表，megabytes 的 Excel 和 PowerPoint 文档，但只有 bite 可以采取行动的 insights

1. Deliver the right Data for the right People in the right format at the right time through the right Way.
2. Data → Information → Knowledge → Actionable Plan → Positive Impact

数据分析要做哪些事？

数据分析团队的自我定位

愿景

- *让数据成为公司的竞争优势*

使命

- *让数据成为公司的共同语言*

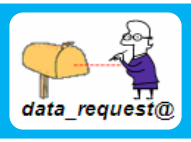
定位

- *基于数据的商业智慧信息提供者与公司运营策略咨询顾问*

数据分析的工作内容

数据快递（Data Express）

报表系统（Reporting System）



临时数据需求（Custom Request Solution）

数据挖掘、分析（Data Mining/ Analysis）

数据传道（Data Evangelization）

一定要的

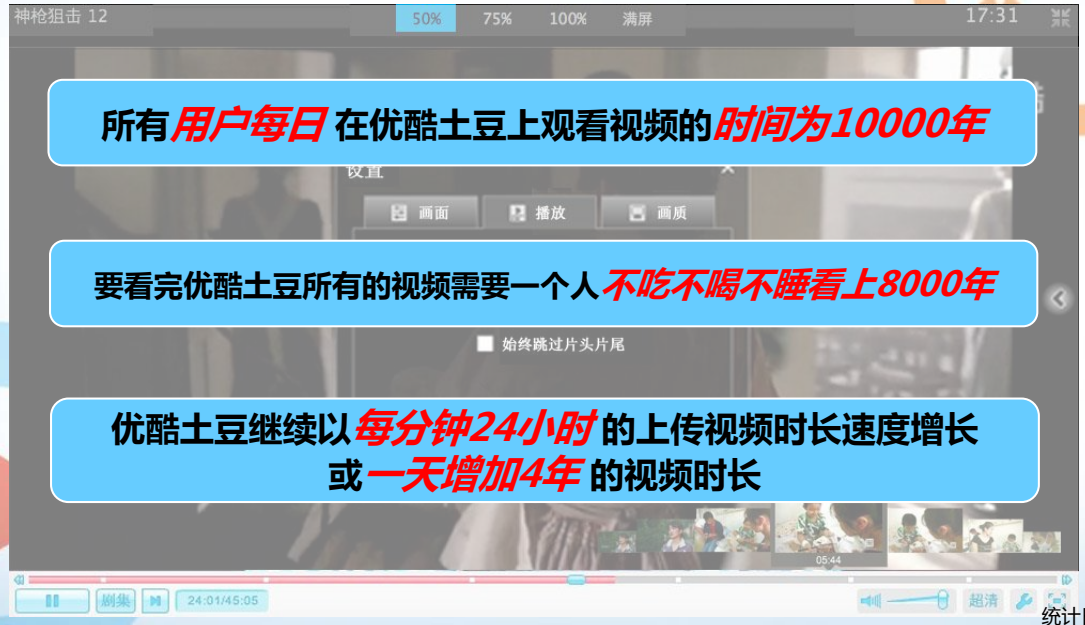
功力的考验

数据传道

数据培训	● 新员工数据培训 ● 各团队数据培训 ● 每月数据分享会
知识管理	● 数据藏经阁 ● 蒐藏所有数据相关研究报告的图书馆 ● 数据相关基础知识百科
数读	● 每两周发布精华浓缩版数据分析报告
对外数据运营	● 中国网络视频指数 ● 数据中心报告，网页及微博运营

数据培训 – 有趣的数据体验

生动有趣的数据



所有*用户每日*在优酷土豆上观看视频的*时间为10000年*

要看完优酷土豆所有的视频需要一个人*不吃不喝不睡看上8000年*

优酷土豆继续以*每分钟24小时*的上传视频时长速度增长
或*一天增加4年*的视频时长

统计日期：2013/11

优酷土豆数读 – 网络第一神剧『万万没想到』

优酷土豆数读 – 遇见影响者

Youku 优酷

上传

遇见影响者

每月4亿活跃用户，每日10万个
使得优酷成为中国第一的视频网
兴的影响者。这一部分人不仅获
评论，更激起了网民情感上无法
视频作品更具有深度与创意，或
或传递知识，不断地从各方方面
们

我们计算的是各位UGC用户上传的视频在整个11月的播放量，播放量越高影响力越大
头像上带有 的为优酷认证用户

整体数据工作框架

哪些人/团队在做数据分析？

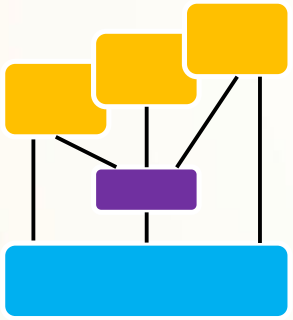
数据团队5种组织架构

- 数据参与者：数据分析团队、业务团队、技术数据团队

A – 在业务团队中

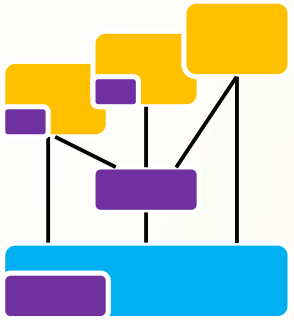
B – 在技术团队中

C – 独立团队



D – 分散在业务及技术团队

E – 为独立团队且各团队都有



5种组织架构优缺点比较

组织架构类型	优点	缺点	建议
A – 在业务团队中	对业务有更深的了解	知识分享局限 数据底层了解受限 跨团队项目较难推动	业务团队有非常专门特定的数据需求
B – 在技术团队中	数据底层有充分了解 数据产品开发 数据科学实验室	与业务距离远 分析是附加延伸工作 技术人员与分析师需要不同的技能组合	不同的业务团队有普遍相同的数据需求
C – 独立团队	知识分享、再利用 数据传道 贴近公司战略方向 更专业统计分析技能	项目优先级排定 业务、分析与技术团队互相牵制	有创新需求 数据密集型公司
D – 分散在业务及技术团队	技术支持长期战略性项目 业务分析团队支持短期战术性项目	技术团队数据分析师受冷落 知识分享局限 技术人员与分析师需要不同的技能组合	
E – 为独立团队且各团队都有	知识分享、再利用 数据传道	管理层大力支持 独立团队领导者能力要求	需要明确清楚的团队定位与认同 公司对数据的掌握与应用有一定成熟度

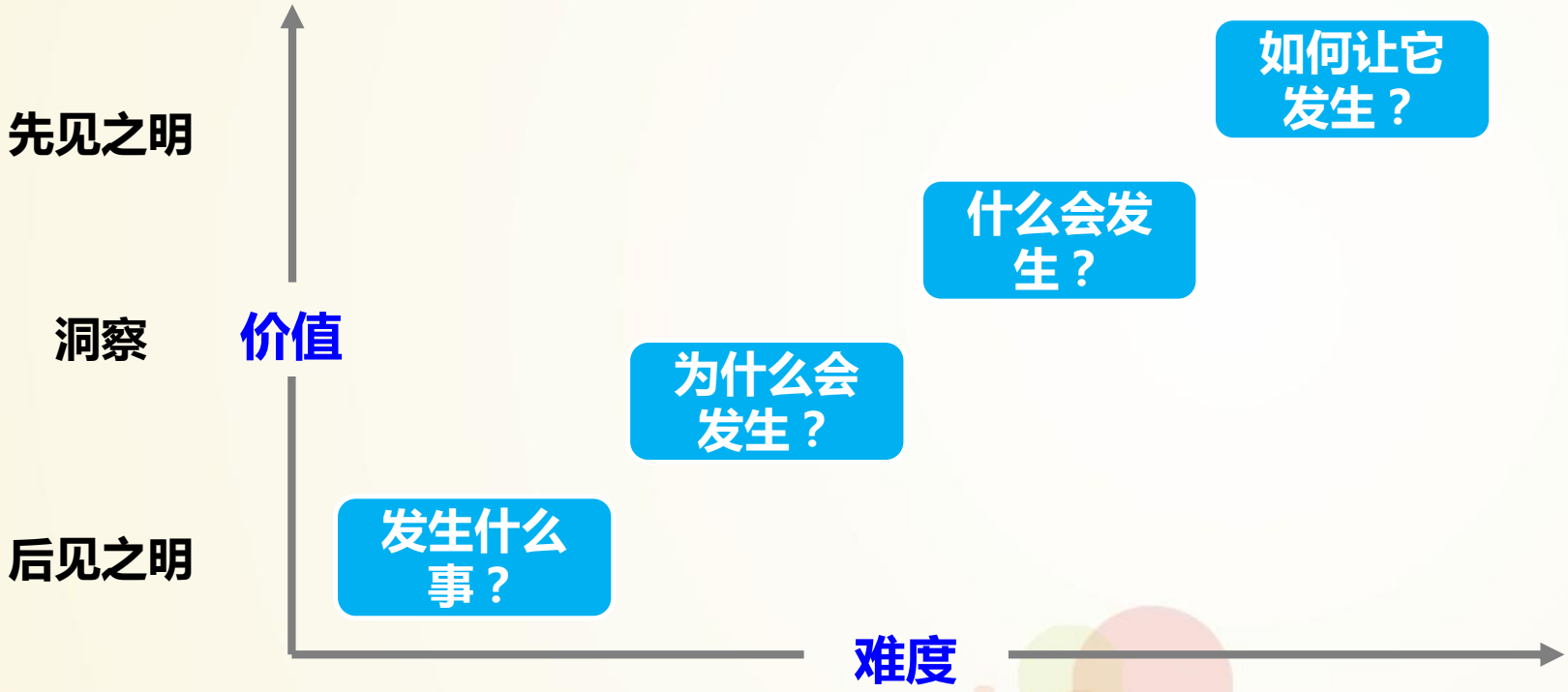
企业内数据分析工作会遇到哪些挑战？

数据工作的五大挑战

1. 需要充分了解公司各项业务及需求
2. 除了拥有核心的分析技能外，还必须对数据采集、处理、存储等技术有充分的了解
3. 知识分享机制（接收与给予）
4. 执行跨团队项目
5. 获得最高管理团队支持

如何提升数据分析能力？

数据分析四部曲



数据分析师思维

同样问题，两种思考逻辑

1. 成吉思汗的继承人窝阔台，公元哪一年死？最远打到哪里？
2. 成吉思汗的继承人窝阔台，当初如果没有死，欧洲会发生什么变化？试从经济、政治、社 会三方面分析

1. 甲午战争是哪一年爆发的？签订的叫什么条约？割让多少土地？赔偿多少银两？
2. 日本跟中国100年打一仗，19世纪打了甲午战争，20世纪打了一场抗日战争，21世纪如果日本跟中国开火，你认为大概是什么时候？可能的远因和近因在哪里？分析之

数据力 (Data Capability)培养

- 数据分析能力与数据解读能力
- 运用数据的能力与使用数据來判断事情及做决策的能力

数据分析师思维

美国有个学生是这样回答的：这位**蒙古领导人如果当初没有死**，那么可怕的**黑死病就不会被带到欧洲去**，后来才知道那个东西是老鼠身上的跳蚤引起的鼠疫。但是六百多年前，黑死病在欧洲猖獗的时候，谁晓得这个叫做鼠疫。如果没有黑死病，**神父跟修女就不会死亡**。神父跟修女如果没有死亡，**就不会怀疑上帝的存在**。如果没有怀疑上帝的存在，就**不会有意大利弗罗伦斯的文艺复兴**。如果没有文艺复兴，西班牙、南欧就不会强大，**西班牙无敌舰队就不可能建立**。如果西班牙不够强大，意大利不够强大，盎格鲁一撒克逊，会提早! 200年强大，**日耳曼会控制中欧**，**奥匈帝国就不可能存在**。

有个日本高中生是这样分析的：我们跟中国很可能在**台湾回到中国以后，有一场激战**。台湾如果回到中国，中国会把基隆与高雄封锁，**台湾海峡就会变成中国的内海**，我们的油轮就统统走右边，走基隆和高雄的右边。这样，会**增加日本的运油成本**。我们的石油从波斯湾出来跨过印度洋，穿过马六甲海峡，上中国南海，跨台湾海峡进东海! ，到日本海，这是石油生命线，中国政府如果把台湾海峡封锁起来，我们的货轮一定要从那里经过，我们的**主力舰和驱逐舰就会出动**，中国海军一看到日本出兵，马上就会上场，那就打！按照判断，公元2015年至2020年之间，这场战争可能爆发。所以，我们现在就要做对华抗战的准备。

人家培养的是能力，而我们灌输的是知识。

如何成为一位出色的数据分析师

分析师与大师的差异

1. 你是否具有敏锐的商业感觉？
2. 你是否具备缜密的逻辑分析能力？
3. 你是否能够从现象中抽出核心的问题所在？
4. 你能否会分清重点避免在一些无谓的问题上钻牛角尖耗费精力？
5. 你是否能具有宏观的思维，又能在微观层面上进行有条理的分析挖掘？
6. 你是否能把你所获得的见解和结论以最好的故事和讲法呈现出来？

出色数据分析师的10种特徵

1. 会使用一个以上的统计分析工具
2. 经常浏览数据分析相关的网站及blogs
3. 在做任何分析前会先调研相关研究与熟悉研究对象
4. 分析是从用户角度出发而不是公司角度
5. 了解各式数据采集方式的差异与数据内容型态
6. 熟悉定性与定量的调查研究方法
7. 饥渴的探索者
8. 有效的沟通者
9. Street smart
10. 防御中带进攻

视频网站数据分析个案分享

Powering XXX's Data for Business Advantage !

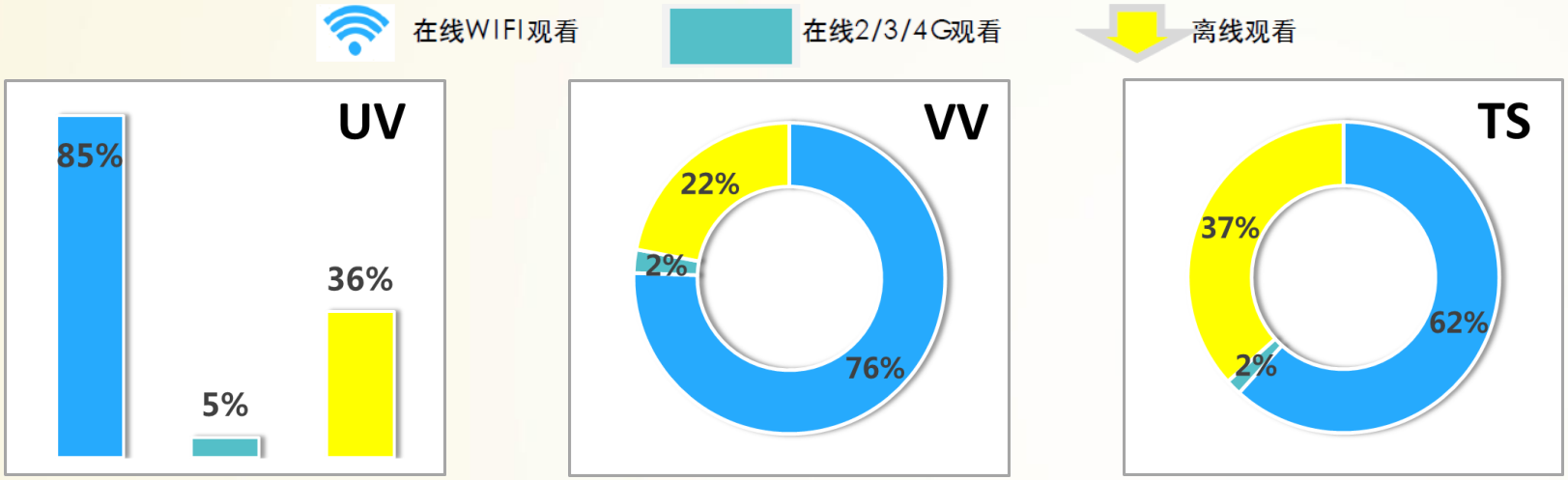
视频网站数据分析框架


产业趋势
流量监测
广告客户
人群属性及
内容偏好
广告效果分析
产品分析
(播放器、网页
改版、点播互动）
用户体验
用户行为分析
（看找完传拍）
多屏行为分析
用户聚类
数据平台
用户
剧集生命
周期分析
UGC播放影响
因子分析
用户内容偏好
分析
视频购物篮
分析
内容挖掘
市场竞争

暴走的移动视频

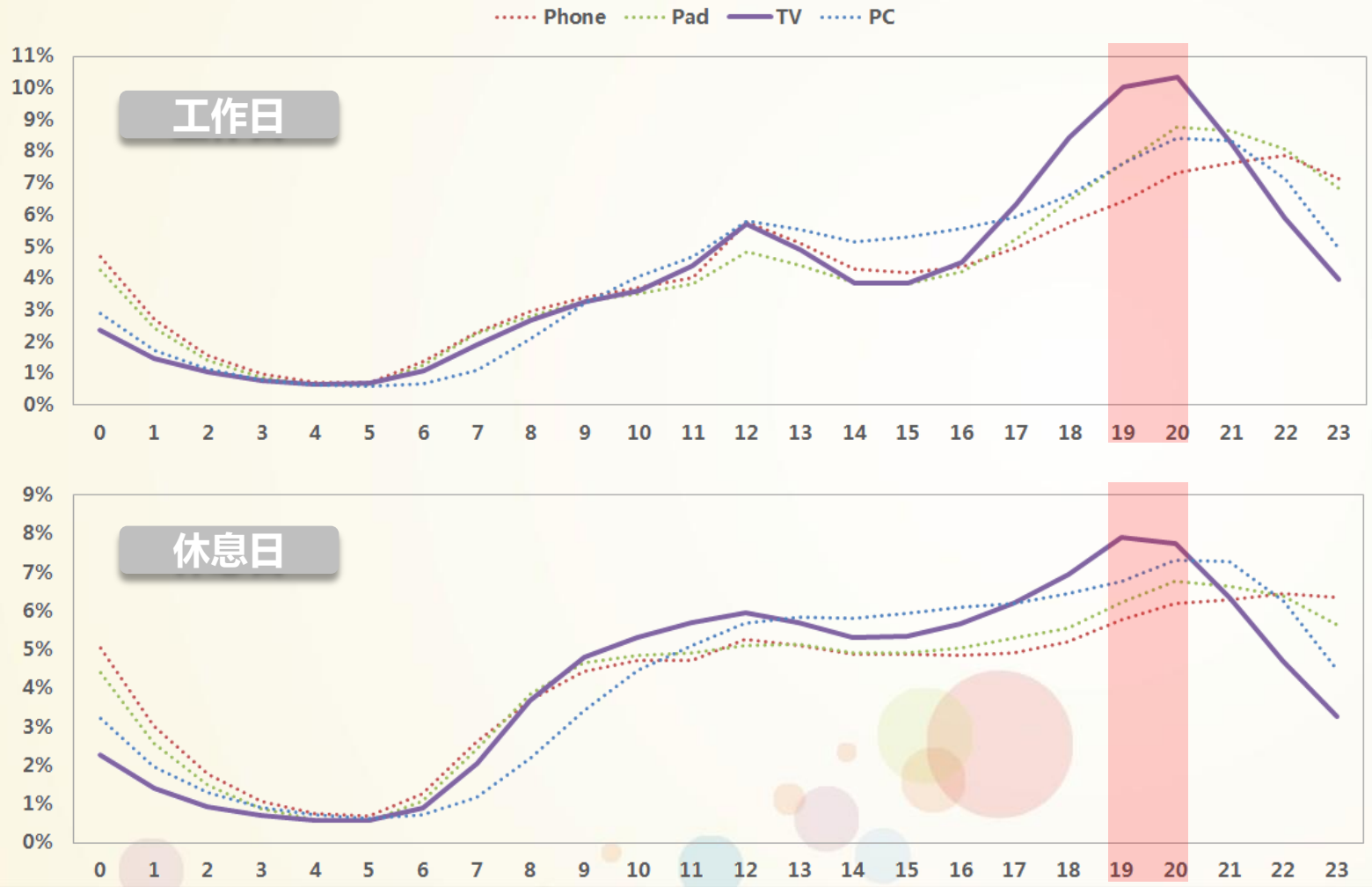

2012
2014
80%
70%
60%
50%
40%
30%
20%
10%
0%
6/30 (10%)
1/1 (3%)
1/1 (1%)
3/22 (67%)
12/22 (45%)
12/20(40%)
12'1
2
3
4
5
6
7
12
14'1
2
3
马斯洛需求层次理论现实升级版
自我实现需求
自我实现，发挥潜能等
尊重需求
成就、名声、地位和晋升机会等
社交需求
对友谊、爱情以及隶属关系的需求
安全需求
人身安全、生活稳定以及免遭痛苦、威胁或疾病等
生理需求
食物、水、空气、性欲、健康
WiFi

移动观看视频联网状态

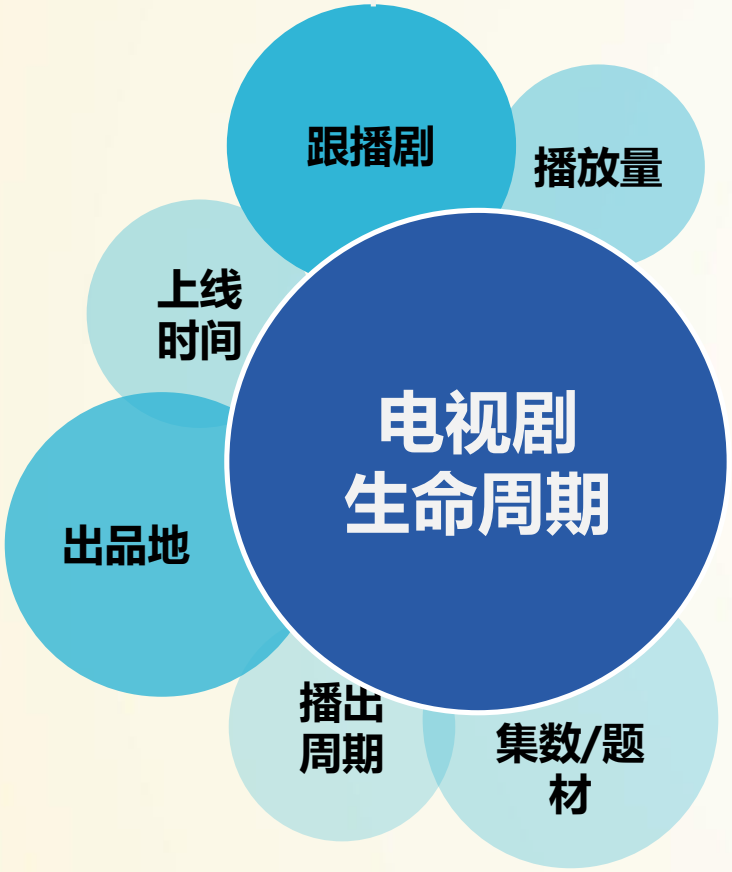


设备	不同网络状况下VV			不同网络状况下TS		
	WIFI	2/3/4G	离线	WIFI	2/3/4G	离线
Android Phone	75%	3%	22%	61%	2%	38%
iPhone	75%	2%	23%	56%	2%	42%
Android Pad	79%	3%	18%	69%	3%	28%
iPad	77%	0%	22%	71%	1%	28%

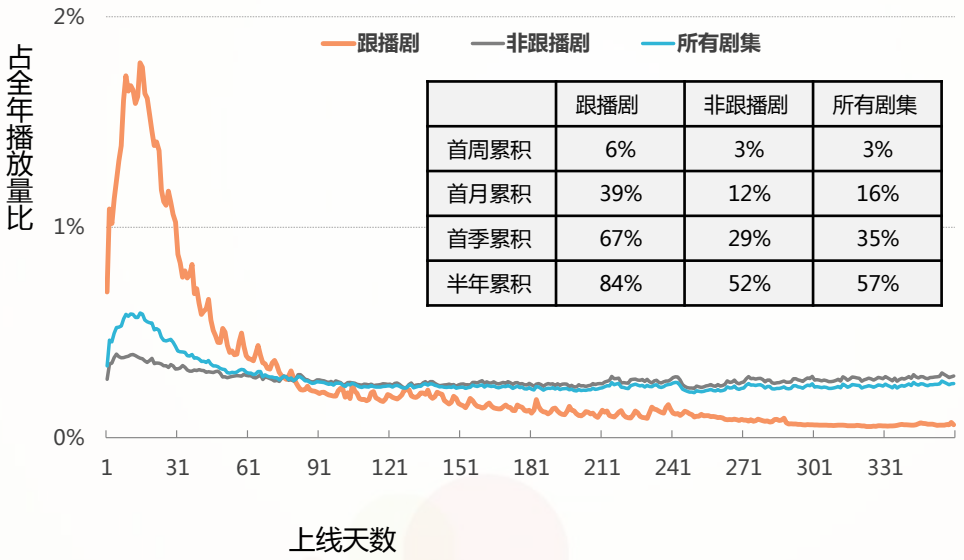
不同屏幕的24小时VV趋势对比



电视剧剧集播放量的生命周期

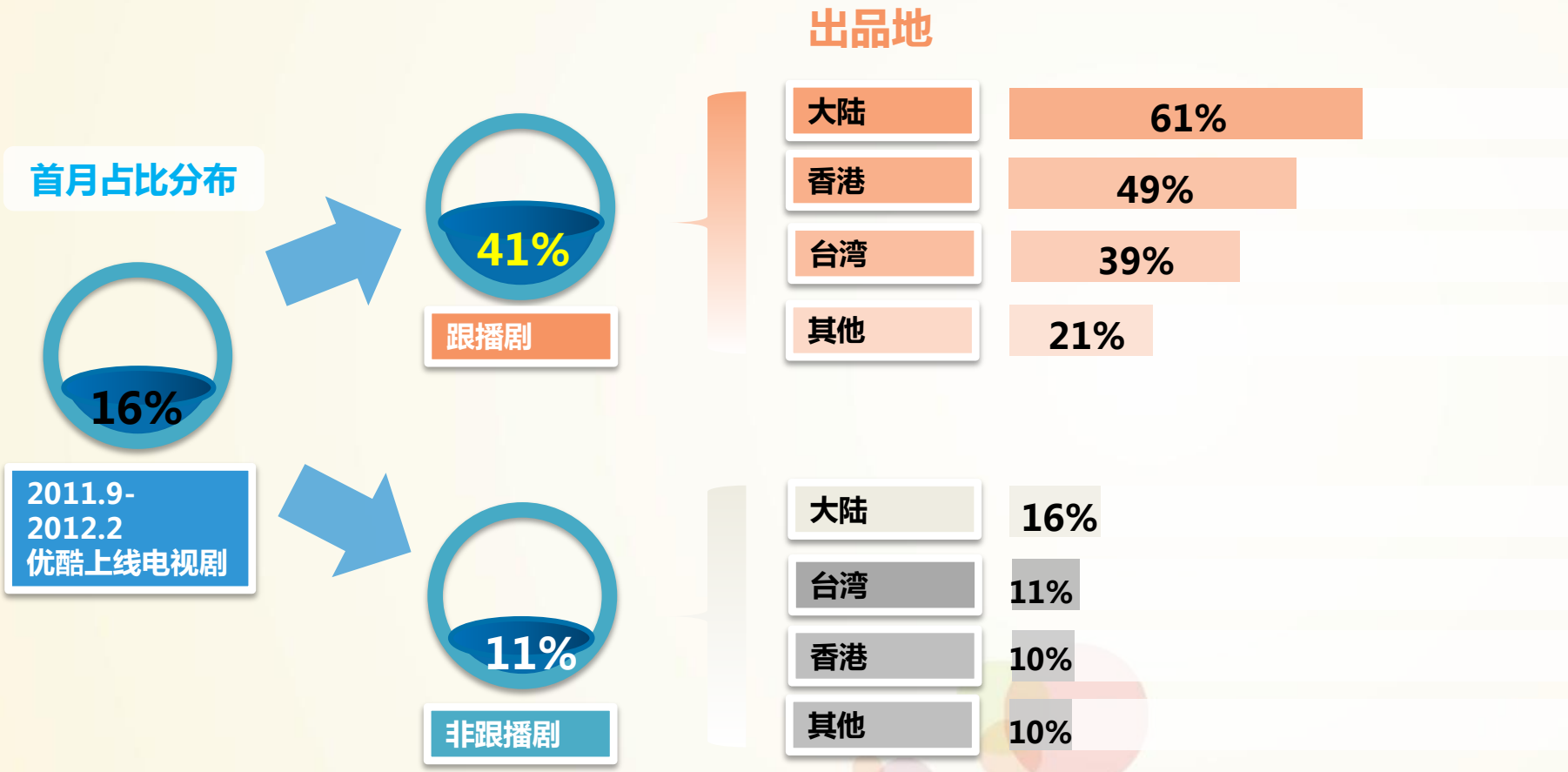


- **跟播与否**是影响电视剧生命周期的众多因素中最重要的一个



	跟播剧	非跟播剧	所有剧集
首周累积	6%	3%	3%
首月累积	39%	12%	16%
首季累积	67%	29%	35%
半年累积	84%	52%	57%

电视剧剧集首月播放量的决策树



我爱电视剧之国-港-台-韩-英-美剧观众画像

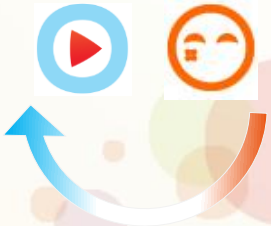
	国产剧	台剧	港剧	韩剧	美剧	英剧
	☆70后熟男同志的菜 ☆初中学历观众的偏好突出 ☆地域性差异较不明显	☆95后小女生的菜 ☆华南粉丝贡献突出	☆轻熟女的菜 ☆华南粉丝贡献突出	☆跨越年龄界线，深获广大女性同胞的青睐 ☆上海、东北五线城市的粉丝贡献突出	☆85后到95前的男粉丝欲罢不能 ☆更偏向高学历、高收入的观众 ☆华中二线和西北三线城市的粉丝偏好明显	☆特受90到95的女粉丝青睐 ☆特别偏向高学历、高收入的观众 ☆上海和华中二线城市的粉丝偏好度最突出
性别						
♀	86	109	109	141	87	116
♂	116	89	90	52	115	81
年龄						
95后	101	163	83	108	89	99
90后	87	104	110	106	126	140
85后	95	85	122	85	114	84
80后	99	80	116	93	101	82
70后	117	82	73	106	69	83


学历
硕士+
本科
专科
高中
初中
100 85 87 89 134 149
88 69 98 117 134 165
96 89 113 107 97 77
108 120 101 92 74 54
117 154 91 76 59 49
月收入
1万以上
5千-1万
3千-5千
3千以下
无收入
150 50 100
地域
城市级别
一 二 三 四 五
华北 华东 华南 华中 东北 西南 西北
说明
151+ 131-150 111-130 91-110 71-90 51-70 50- 无城市

有人比你更了解你

- 圣经.诗篇139篇

- 耶和华啊，你已经鉴察我，认识我。
- 我坐下，我起来，你都知晓；你从远处知道我的意念。
- 我行路，我躺卧，你都细察；你也深知我一切所行的。



数据运营中心

洞见发现，知识传递

Q&A

THANKS