

www.qconferences.com  
www.qconbeijing.com



伦敦 | 北京 | 东京 | 纽约 | 圣保罗 | 上海 | 旧金山  
London · Beijing · Tokyo · New York · Sao Paulo · Shanghai · San Francisco

# QCon全球软件开发大会

International Software Development Conference



# HULU推荐系统构建经验谈

郑华 Principal Dev Lead@hulu, 2013/04

# HULU介绍

- 美国最大的在线专业视频网站(Online Premium Video Site)
- 2007年由NBCU, FOX成立, 2009年Disney-ABC加入
- 450+内容提供商, 1000+广告商
- 每月30M独立用户, >3M付费用户
- 2012年收入690M USD



# 自我介绍

- 1982年，出生在江苏盐城
- 2000年，清华大学电子工程系
- 2006年，加入视频标注创业公司MOJITI.COM
- 2007年，硕士毕业后加入HULU.COM，先后负责播放器，推荐系统，广告精准投放和用户数据分析平台等



# 内容概要

- 目标
- 产品
- 数据
- 算法
- 架构
- 效果
- 总结

# 目标

- 我们的使命 - 帮助用户找到并欣赏世界上最专业的内容, 无论何时, 何地以及以何种方式 (Help people find and enjoy the world's premium content; when, where, and how they want it.)
- 推荐的目标:
  - 提供个性化的用户体验, 帮助用户更快找到自己感兴趣的内容
  - 帮助内容提供商更好的投放自己的内容, 实现内容价值最大化
  - 帮助Hulu提高用户的满意度和停留时间

# 产品

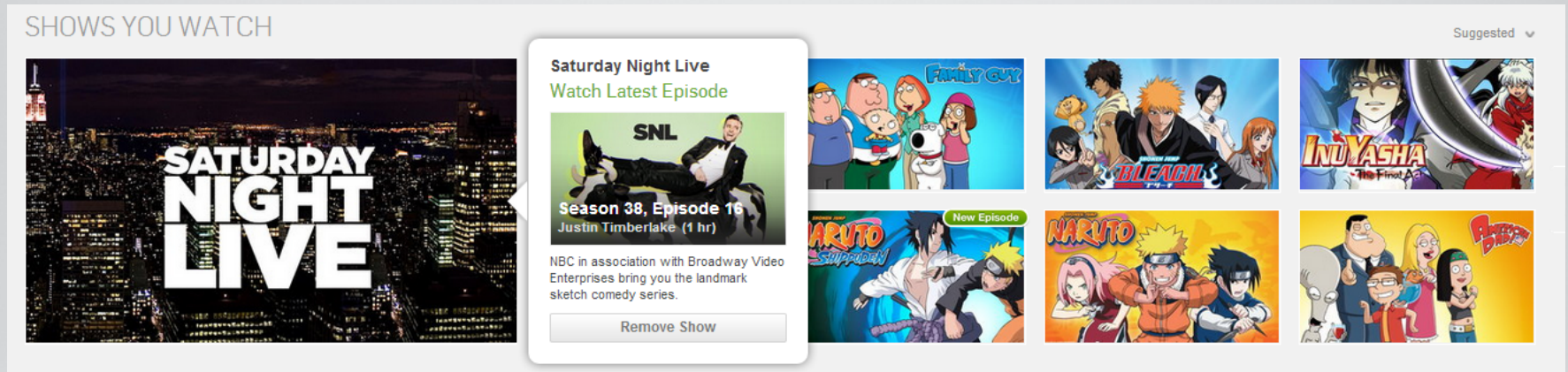
- 便捷浏览 (Navigation Convenience)
  - 帮助用户更快Catchup内容, 缩短用户的浏览路径
- 发现内容 (Content Discovery)
  - 帮助用户发现自己感兴趣的新的东西

# 便捷浏览

- Shows You Watch
- Show Smart Start
- UpNext



# SHOWS YOU WATCH



- 按照最后一次观看时间排序，或者基于机器学习模型排序，后者比前者转化率提高10%以上
- 特征: SHOW的本身的特性, 用户观看SHOW的习惯, 当前用户的平台和时间等

# SHOW SMART START

The screenshot shows the Hulu website interface. At the top, there's a navigation bar with the Hulu logo, a 'BROWSE' dropdown menu, a 'Try Hulu Plus' button, a search bar, and user information including 'Queue (694)', 'Social Off', and a user profile named 'Zheng'. The main banner features the text 'NIGHT LIVE' in large white letters over a night cityscape. Below the banner, there's a section for 'Watch Latest Episode' which highlights 'Season 38, Episode 16 Justin Timberlake (1 hr)' from 'SNL'. To the right of this section are social media sharing options: a star rating, a 'favorite' button, and buttons for 'facebook', 'tweet', and 'comments (37)'. Below these is a paragraph of text about the series and a list of bullet points describing the content and availability. On the far right, there's a circular profile picture of a person and a 'Like' button with the number '427'. At the bottom right, the text 'Lonn Lee, Renee Cao and Robert Wong like this show' is visible, along with the word 'Social' in blue.

hulu BROWSE Try Hulu Plus NIGHT LIVE Queue (694) Social Off Zheng

Watch Latest Episode

Season 38, Episode 16 Justin Timberlake (1 hr)

★★★★★ + favorite facebook tweet comments (37)

NBC in association with Broadway Video Enterprises bring you the landmark sketch comedy series. [more](#)

- Full episodes of SNL consist of material that is cleared for online streaming. Some sketches and musical performances may be omitted due to various restrictions. Episodes are posted Sunday afternoons and are available five at a time. All of the clips cleared for online streaming are posted on Hulu. Select musical performances are available for a limited time.
- huluPLUS** Subscribe to Hulu Plus to watch the entire series in HD on your TV, mobile and computer. [Try it FREE](#)

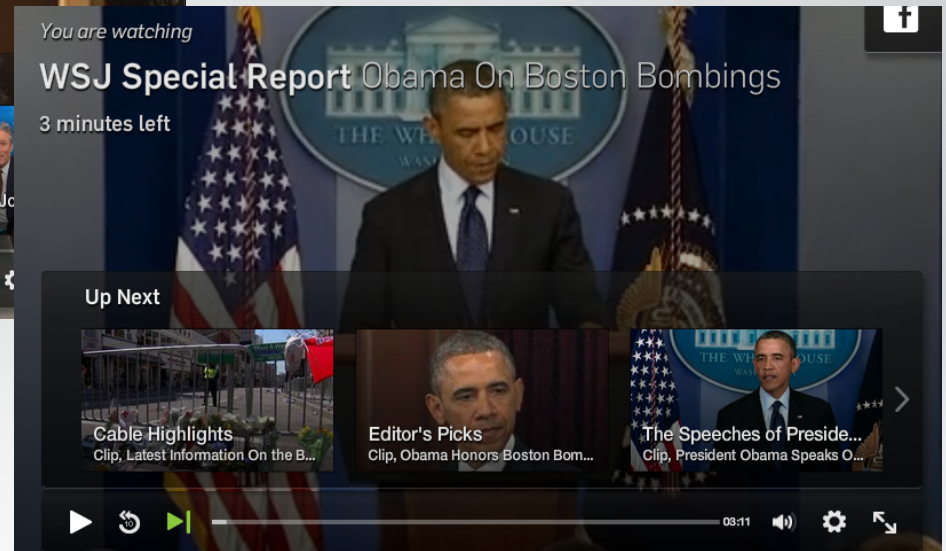
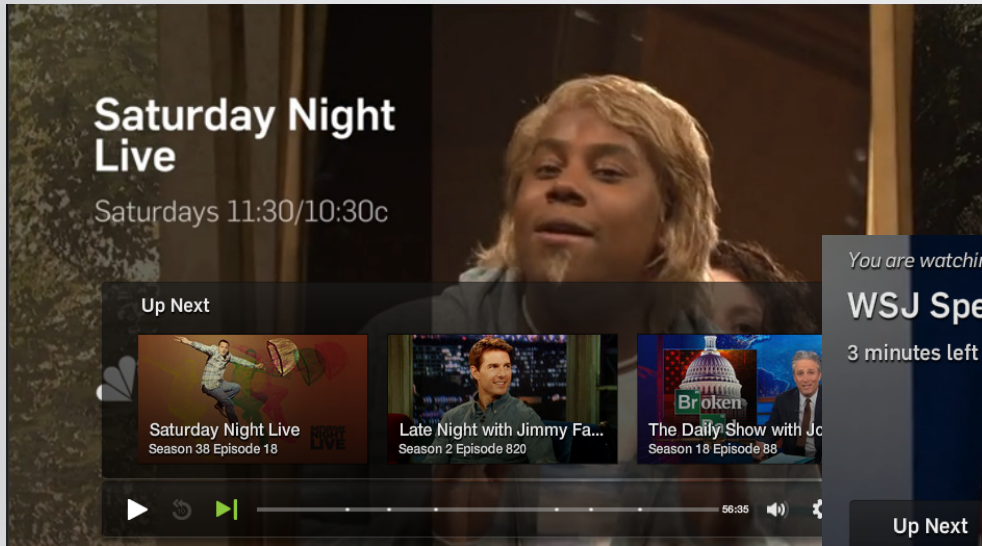
Lonn Lee, Renee Cao and Robert Wong like this show

Social

- 定位用户当前需要看哪一集以及从这集的什么位置看



# UPNEXT



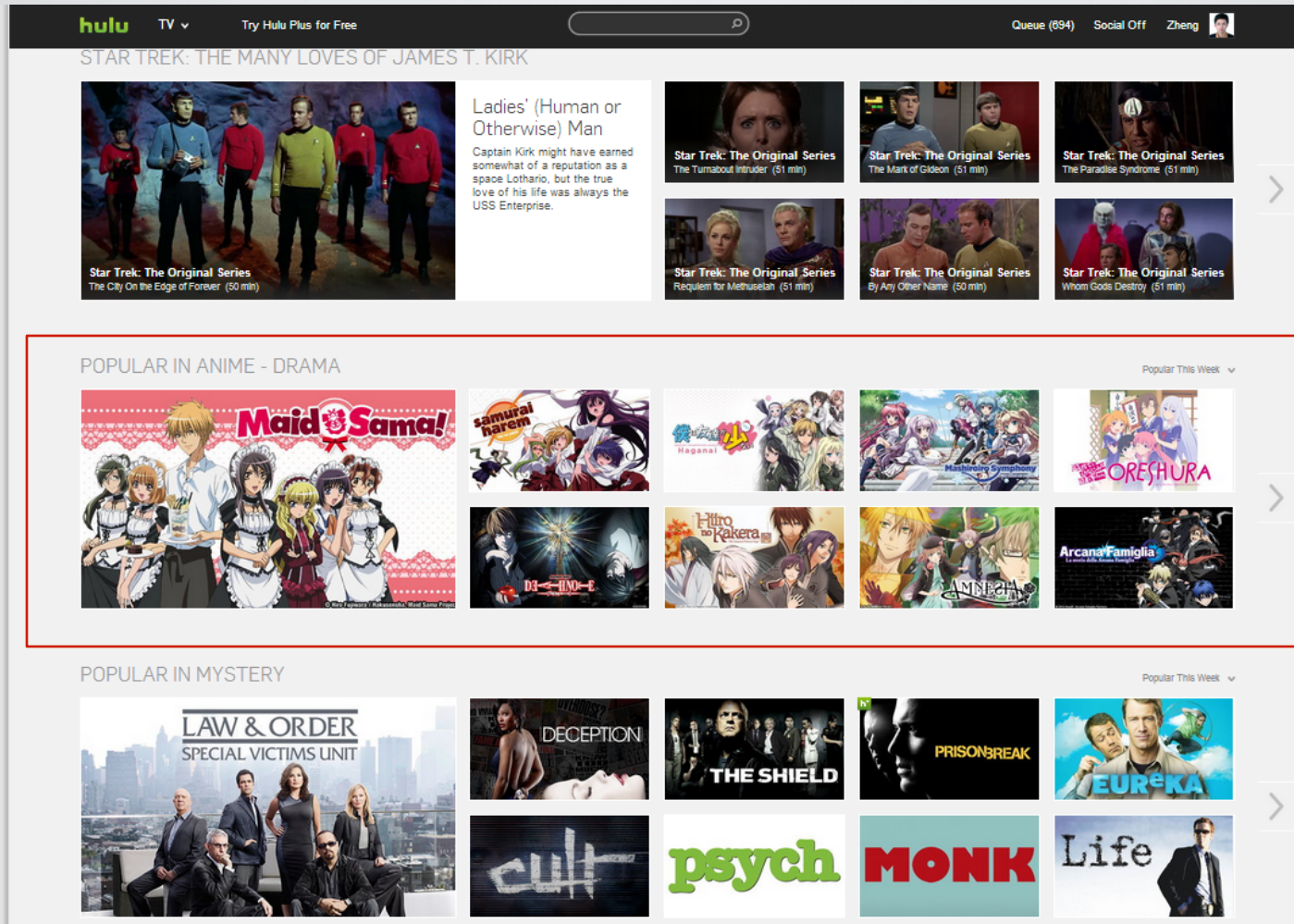
- 长视频: 自动播放同一个剧集的下一集或者相似剧集的Pilot
- 短视频: 相同主题的短视频
- 兼具Navigation Con和Discovery的功能

# 发现内容

- Personalized Homepage
- Personalized Masthead
- Top Recommended For You
- You Might Also Like (aka. viewers of this show also viewed ... )
- Steamboat



# PERSONALIZED HOMEPAGE



Tray

- 选择, 排序Tray和Tray里面的内容
- 保证相关性, 多样性, 新鲜性等提高点击率

# PERSONALIZED MASTHEAD

The screenshot displays the Hulu homepage interface. At the top, the Hulu logo is on the left, followed by navigation links: "BROWSE", "Try Hulu Plus", a search bar, "Queue (707)", "Social Off", and a user profile "Zheng". The main section is titled "TRENDING NOW" with the date "MARCH 29". Below this, there are two featured video thumbnails. The first thumbnail shows four people in chicken costumes performing on stage, with the text "Blake Shelton, Nick Offerman, Other Chickens Sing 'Ho Hey'" and "It's bacawfully impressive." The second thumbnail shows Nick Offerman on a talk show set, with the text "Nick Offerman Briefly Descends Pyramid of Greatness to Breakdance" and "Breakdance has been a very important part of my life." Both thumbnails include the text "Late Night with Jimmy Fallon" and their respective durations. Below the trending section is a promotional banner for "THE AMAZING RACE" with the text "Watch on Hulu Plus" and a "Try it Free" button. At the bottom, there is a section titled "SHOWS YOU WATCH" with a "Suggested" dropdown menu. Three show thumbnails are displayed: a dark scene, a cartoon show, and another cartoon show, each with a "New Episode" badge.

hulu BROWSE Try Hulu Plus Queue (707) Social Off Zheng

**TRENDING NOW**  
MARCH 29

#TrendingOnHulu

**Blake Shelton, Nick Offerman, Other Chickens Sing "Ho Hey"**  
It's bacawfully impressive.  
Late Night with Jimmy Fallon (2 min)

**Nick Offerman Briefly Descends Pyramid of Greatness to Breakdance**  
"Breakdance has been a very important part of my life."  
Late Night with Jimmy Fallon (4 min)

**THE AMAZING RACE** Watch on Hulu Plus  
Try it Free

SHOWS YOU WATCH Suggested

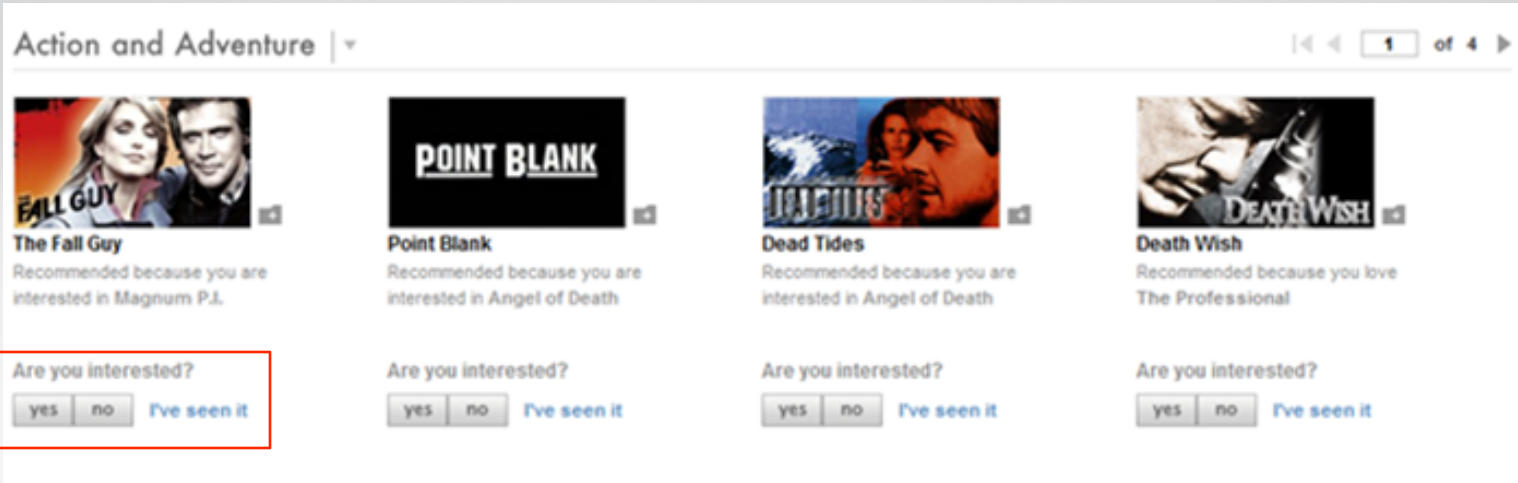
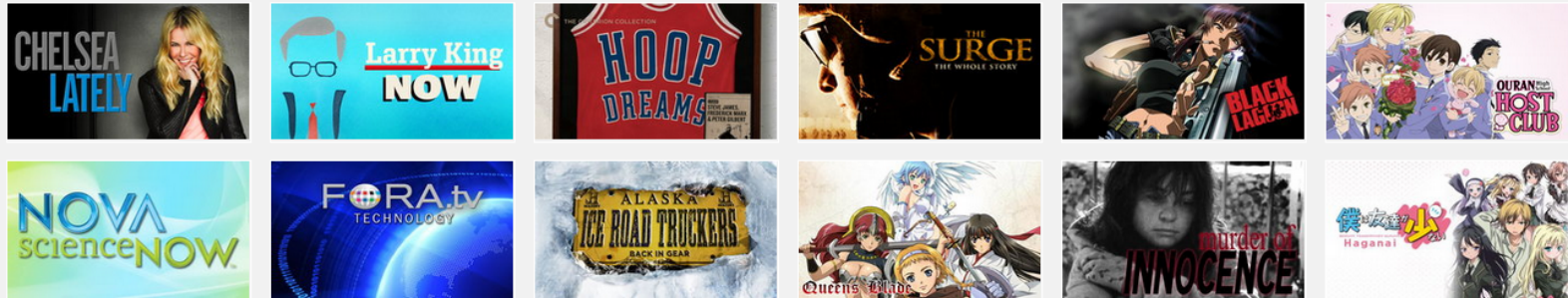
New Episode New Episode

- 选择并排序Masthead里面Trending的内容
- 理解用户的兴趣 (e.g. genre, entity, topics, etc)



# TOP RECOMMENDED FOR YOU

## RECOMMENDED FOR YOU

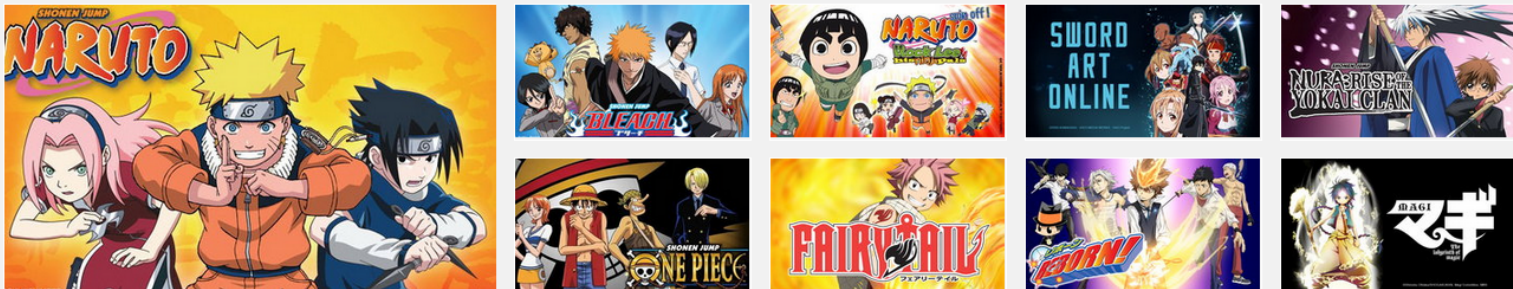


- 基于用户的观看，搜索，打分等行为，推荐用户最可能观看的剧集
- 用户可以对推荐结果反馈

# YOU MIGHT ALSO LIKE



YOU MAY ALSO LIKE



- 相关推荐, 看过当前剧集的用户也看了这些剧集



# STEAMBOAT

Ad : Your video will resume in 7 seconds.

Does this recommendation interest you?

## Chicago Hope

Season 1

Available on

# hulu

- 在用户看完当前剧集最新一集的时候使用音视频的形式在广告位推荐其他他可能感兴趣的剧集

# 小结

- 设计不同的产品满足用户不同的需求
- 算法和策略也需要随着产品进行改变
- 好的产品形式可能会带来意想不到的效果
- 鼓励用户反馈

数据

# 数据类型

- 用户:
  - 匿名用户, 注册用户, 订阅用户
  - Demographics (性别, 年龄, 教育, 种族, etc)
  - Social, Facebook Likes
- 物品:
  - 视频(长视频, 短视频)
  - 剧集(首播, 重播)
  - 播放列表(Playlist, 相同主题, 比如习主席出访, 欧债危机, 财政悬崖)
  - 类型(剧情片, 喜剧片, 动作片, 科幻片)
- 用户和物品关系:
  - 显式: 打分, 喜欢, 感兴趣/不感兴趣...
  - 隐式: 观看, 搜索, 浏览, 点击...
- 上下文信息:
  - 设备, 时间, Session

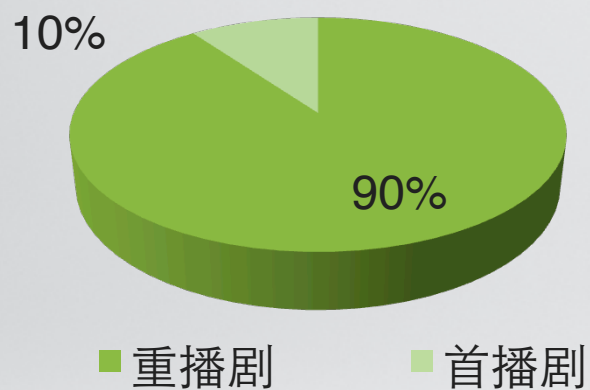


# 数据规模

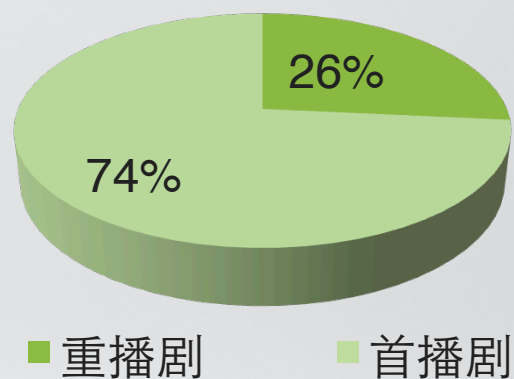
- 每月独立用户数30M, 观看了7亿视频和14亿视频广告\*
- 每天有超过10亿个用户行为比如观看, 搜索, 浏览, 展示, 点击等被记录, 处理的数据超过10T

\* ComScore 2013/02统计数据

# 剧集类型



剧集个数



剧集流量

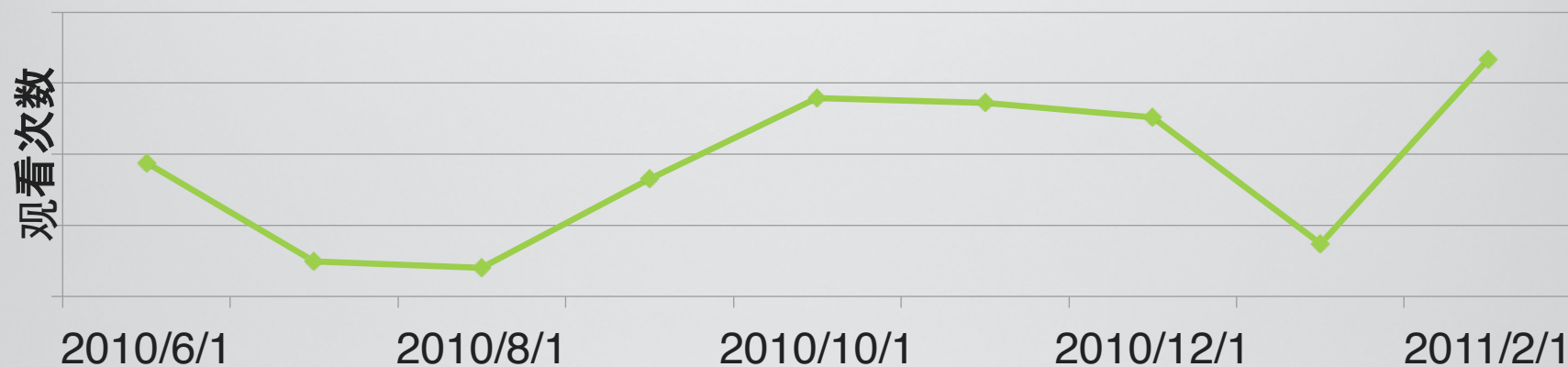
# 追剧(CATCHUP)



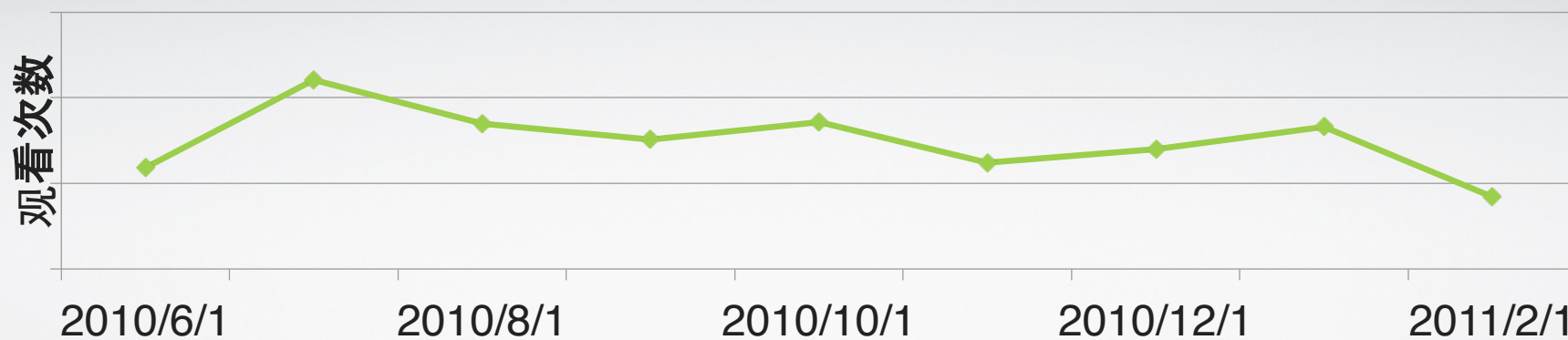


# 季节效应

## 首播剧

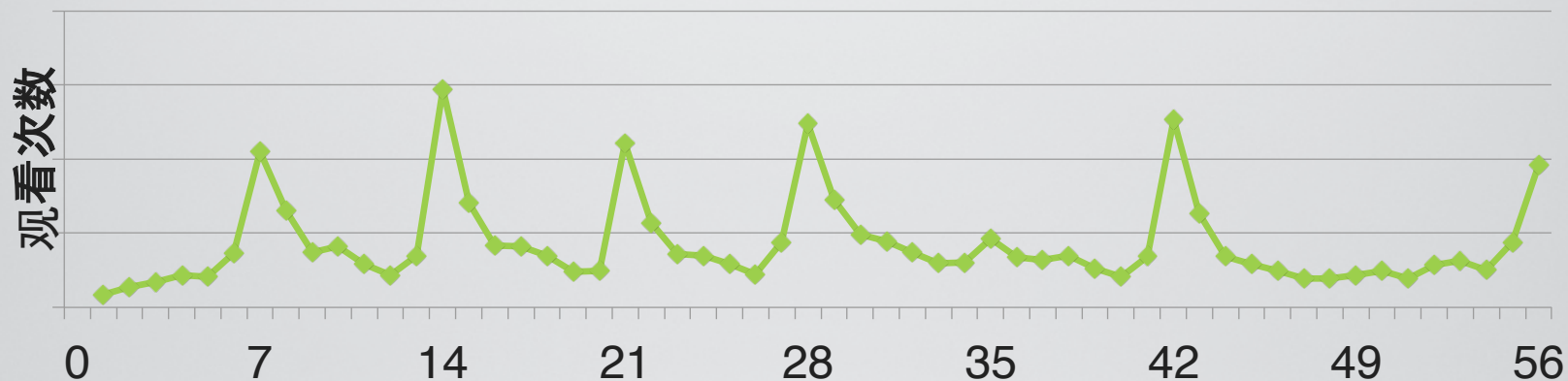


## 重播剧

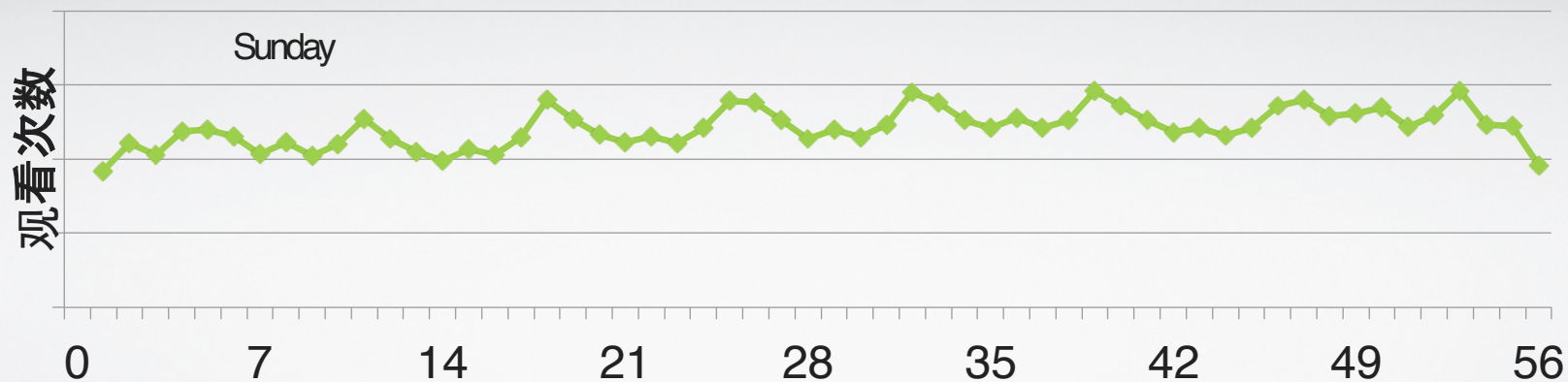


# 周末效应

首播剧

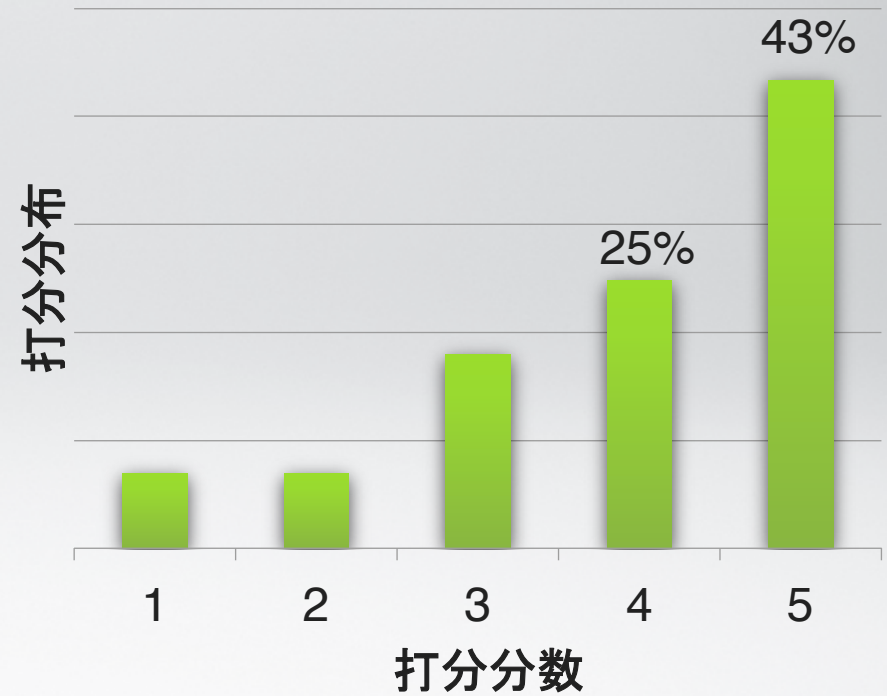
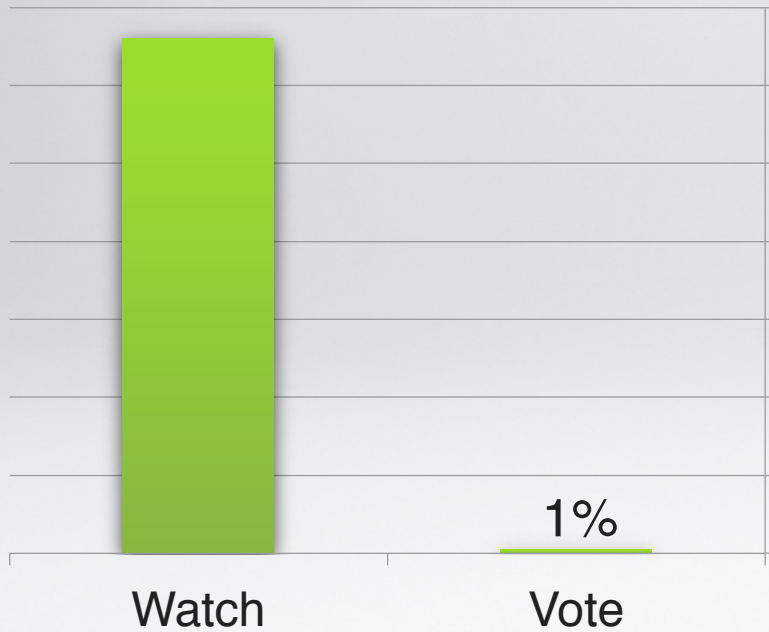


重播剧



# 反馈类型

## 隐式 vs. 显式



# 小结

- 美剧有很强的季节性和周期性，首播剧和重播剧是很好的补充
- 首播剧有充分的关注度，用户对主动追首播剧，需要帮助用户更快的定位到哪部剧的哪一集
- 重播剧质量也很高，相比首播剧关注度不够，更需要主动帮用户发现这些内容
- 用户的隐式反馈无论从数量和种类上要远远多于显示反馈



# 算法

# 推荐流程



- 特征提取:从用户的行为或其他数据中提取出用户兴趣
- 相关物品检索:通过兴趣特征快速查找出候选的相关物品
- 物品排序:按照一定的算法及指标对候选物品排序

# 特征提取

- 用户的性别, 年龄
  - Female Age25-34, Male Age18-24, etc
- 用户感兴趣的剧集
  - Desperate Housewives, Star Trek, Man vs. Wild, etc
- 用户感兴趣的类型
  - 喜剧片, 动作片, 悬疑片, etc
- 用户感兴趣的主题
  - 电视基因 (Show Genome): 事实类(导演, 演员, 类型, 地点, 时间)和语义类(剧情, 主题)
  - 流行的话题 (Trending Topics): 地沟油, 好声音, 十八大
  - 隐含主题模型(Latent Topic Model): 根据物品内容或者用户对物品的行为提取的隐含主题

# 相关物品检索

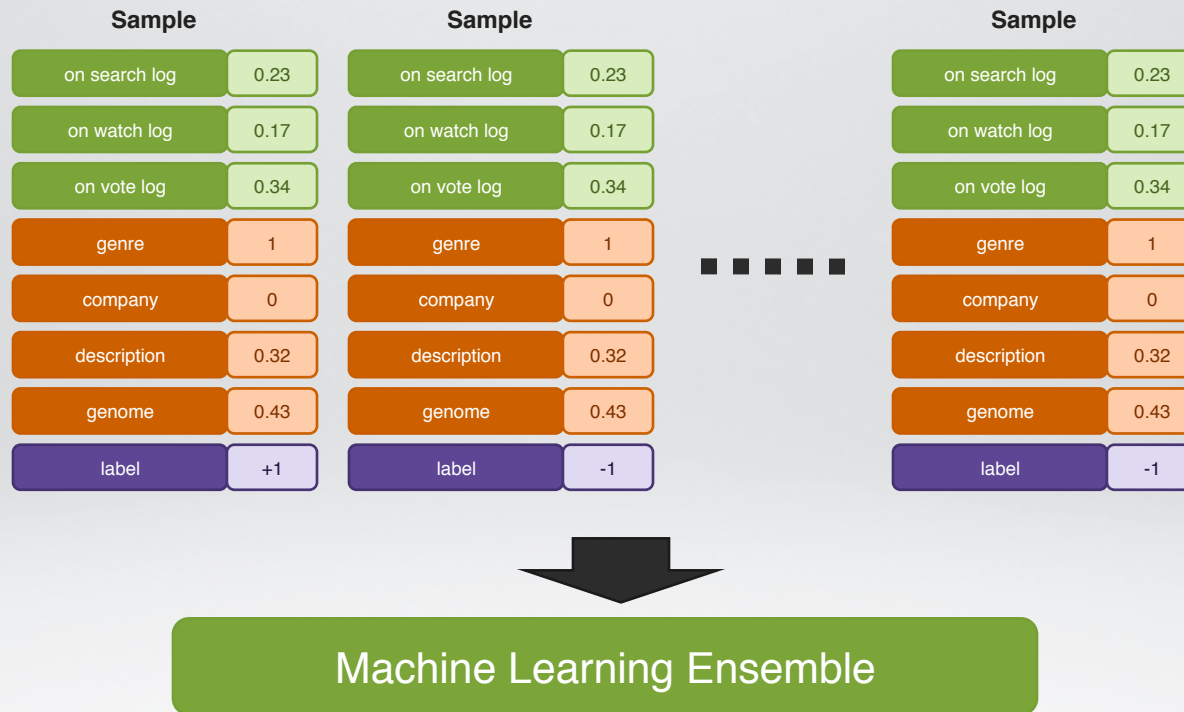
- Demographics:
  - 统计每种Demographics最相关的物品
- 剧集类型:
  - 该剧集类型中最流行或者打分最高的物品, 需要该剧集类型比较有区分力, 比如Drama不如Medical Drama好
- 剧集:
  - 与该剧集相关的物品
- 主题:
  - 属于该主题(电视基因, 流行话题或者隐含主题)的物品



# 物品的相关度

- 基于内容的相关度：
  - 剧集：结构化数据-电视基因
  - 视频：自然语言处理提取关键词计算相似度，对短文本还可以用主题模型；并且对流行主题进行了捕获
- 基于用户行为的相关度：
  - Item协同过滤 (Item-Item CF)
  - 主题模型 (Topic Model)
  - Tips：数据规模很重要；各种用户的反馈(观看，搜索，打分, etc)都可计算相关表，帮助推荐结果
- 利用领域知识(Domain Knowledge)融合相关表

# 物品相关表融合



- 以专家的人工标定和用户的反馈作为学习目标, 融合相关表
- 维数不多的时候Decision Tree效果好于LR

# 物品排序

- 在满足一定约束比如多样性, 新鲜度等的条件下, 并考虑用户所处的上下文(设备, 时间等), 最大化某项指标;
- 使用Logistic Regression预估点击率, 在Hadoop集群上并行训练, 并且可增量更新;
- 也可使用Learning to Rank算法(Netflix)

# 算法模拟和评估

rtl.prod.hulu.com/PersonalizedTray/index.php

Hulu Internal To Do List Setup MySQL R... Machine Learni... MachineLearning PersonalizedTray

251 popular episodes 0.2

1	2	Feedback
No data available in table		

Check More Relevant Features

498 saturday night live ones to remember 0.4759018

1	2	Feedback
WATCH GENRE	comedy~late night comedy	Relevant Irrelevant
WATCH SHOW	Late Night with Jimmy Fallon	Relevant Irrelevant
WATCH GENRE	comedy~sketch comedy	Relevant Irrelevant
TRAY CLICK	saturday night live ones to remember	Relevant Irrelevant
WATCH SHOW	Saturday Night Live	Relevant Irrelevant

Check More Relevant Features

Feature Name	1	2	Feature Weight
this	TRAY CLICK RECENTWEEK		0.40331757
this	TRAY CLICK		0.22876295
this	TRAY CLICL IN PAGE		0.22876295
saturday night live ones to remember	TRAY LOAD WITHOUT CLICK	video game clips	0.05482818
saturday night live ones to remember	TRAY LOAD WITHOUT CLICK	fight the fear	0.043460708
saturday night live ones to remember	TRAY LOAD WITHOUT CLICK	the oscars 2013 red carpet highlights	0.030057654
saturday night live ones to remember	TRAY LOAD WITHOUT CLICK	popular shows in tv subgenre music musicals	0.022716727
saturday night live ones to remember	FAV SHOW	Saturday Night Live	0.022510866
saturday night live ones to remember	TRAY LOAD WITHOUT CLICK	like totally 80s	0.018684613
saturday night live ones to remember	WATCH SHOW	Bleach	0.0061908765

5970 top celebrity headlines 0.43530023

1	2	Feedback
WATCH GENRE	news and information~celebrity and gossip	Relevant Irrelevant
FAV VIDEOTYPE	clip	Relevant Irrelevant

Check More Relevant Features

Feature Name	1	2	Feature Weight
top celebrity headlines	WATCH GENRE VIDEOTYPE	action and adventure	5.1910763e-7
-3047459964635341108	undefined	undefined	0.043676157
top celebrity headlines	TRAY LOAD WITHOUT CLICK	recent popular episodes	0.033893127

- 对用户要透明, 对Product manager和Domain experts更需要透明



# 小结

XAPI

Audience  
Platform

Site DB

Show  
Genome  
DB

Raw Data

User Behavior Data  
(Watch, Vote, Favorite, ...)

Item Meta Data  
(Genre, Company, Genome, Actor, Director, ...)

Structured Data

Machine Learning

Text Mining

Statistics

Collaborative  
Filtering

Method

Item Related Table

User Interest Model

CTR Prediction Model

Model

Collaborative Filtering  
based Table

Latent Topic Model

Logistic Regression  
Model

Content based Table

Feature-Item Table

# 架构

# 技术



APACHE  
**HBASE**



OpenTSDB

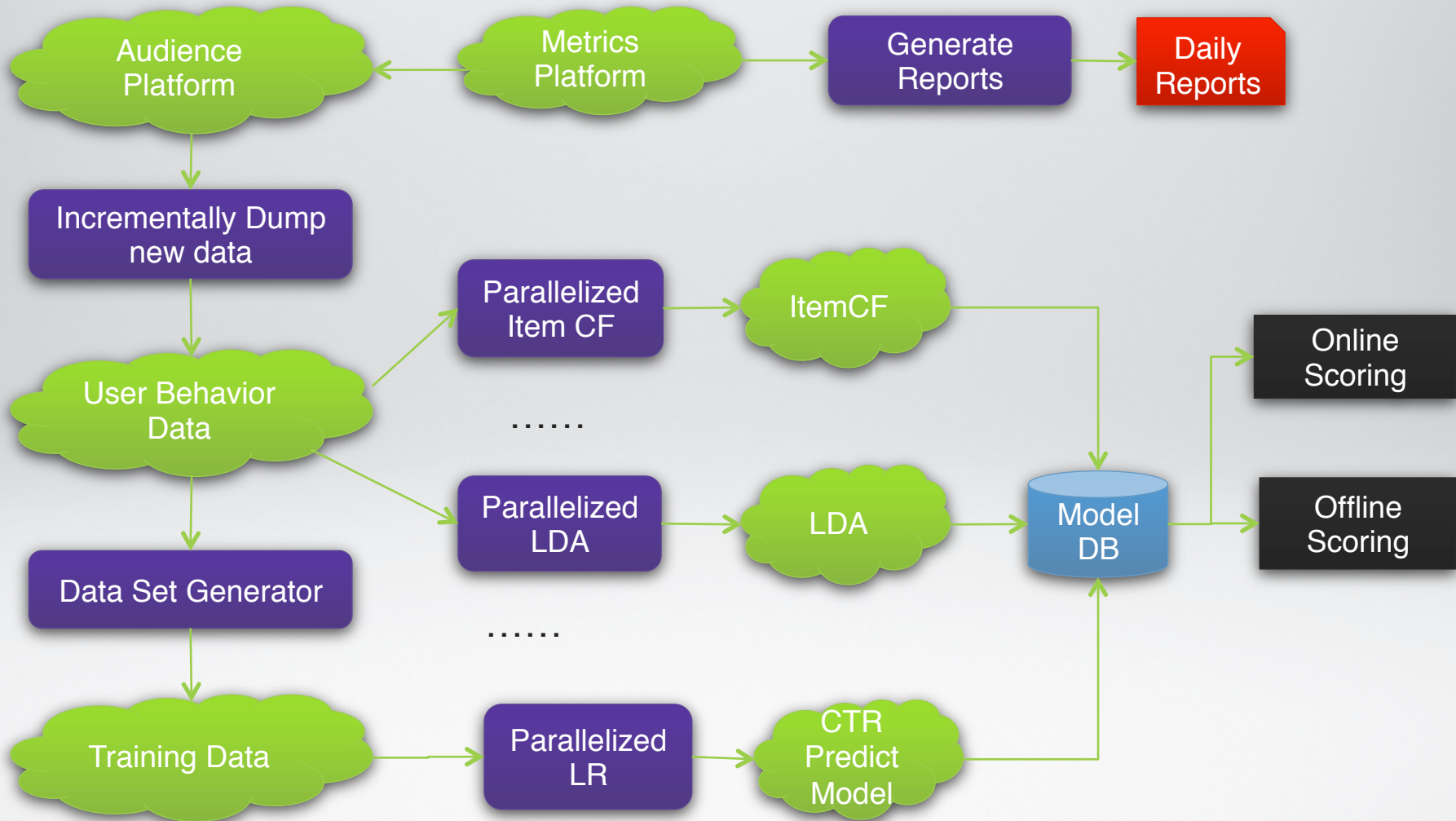


**Cassandra**



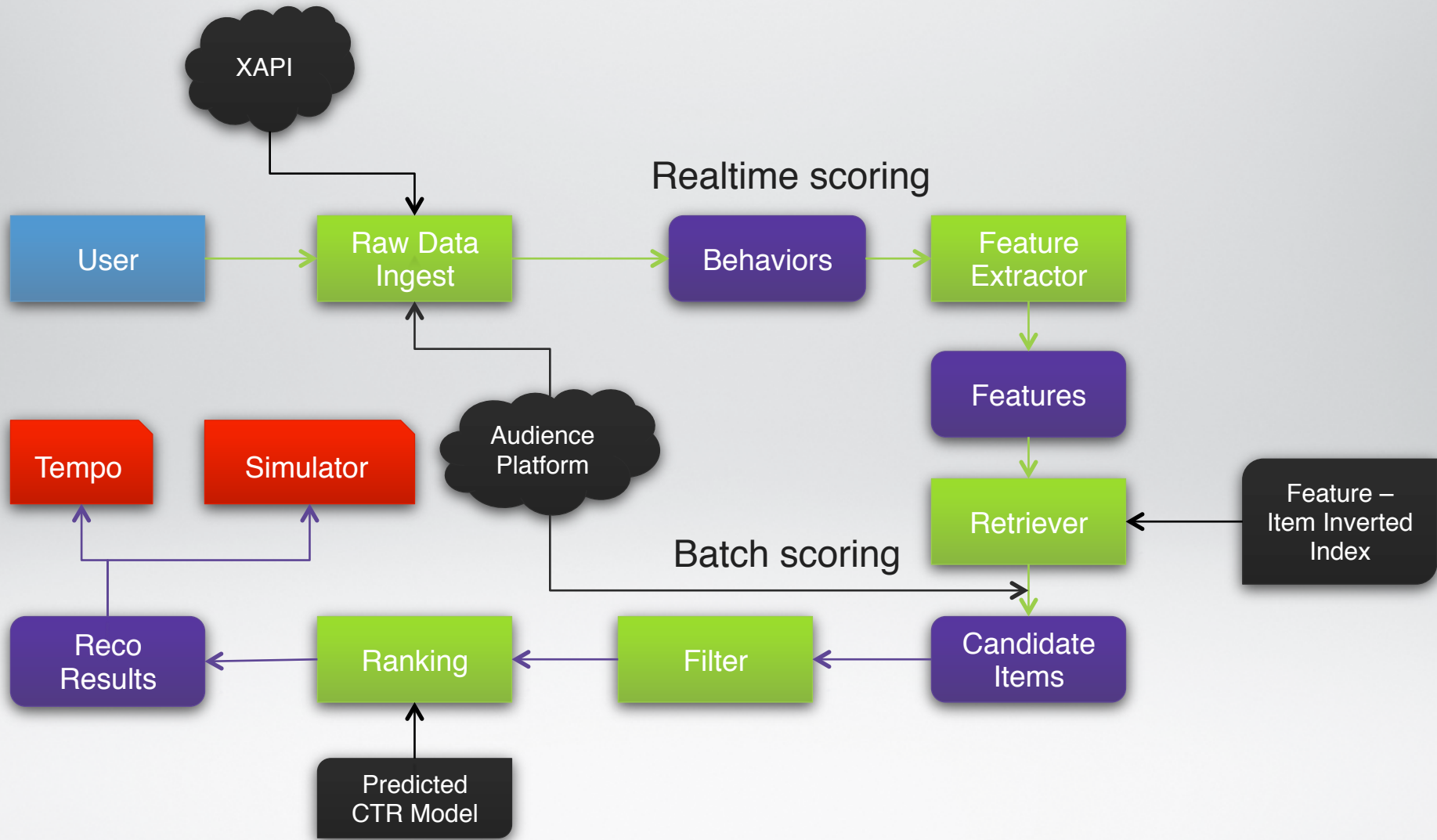
**Jenkins**

# 离线处理





# 在线处理

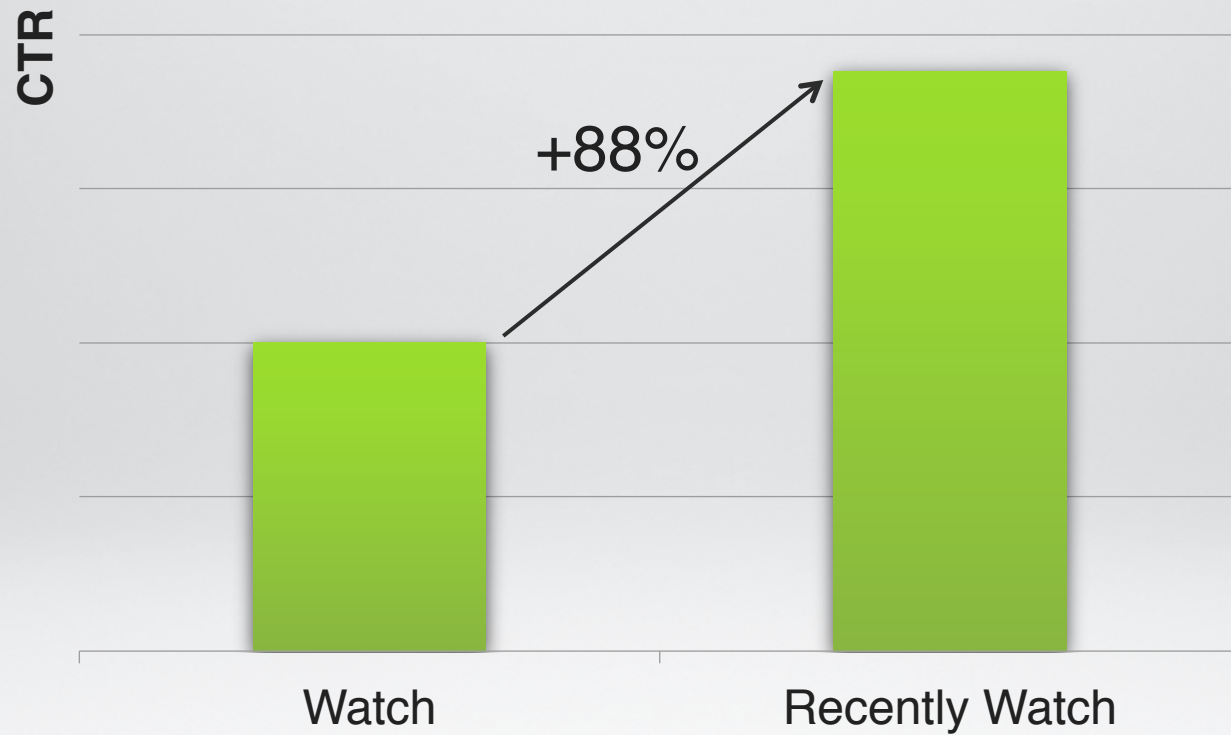


# 评估测试

- 离线算法评估：
  - 离线调整算法后使用离线指标决定能不能上线测试
  - 包括准确率, 覆盖度, 多样性, 新颖性, AUC, 等等
- 在线测试：
  - A/B测试, 指标是点击率, 用户停留时间等, 需要注意分析统计显著性
  - 数据驱动决定(data-driven decision making)

效果

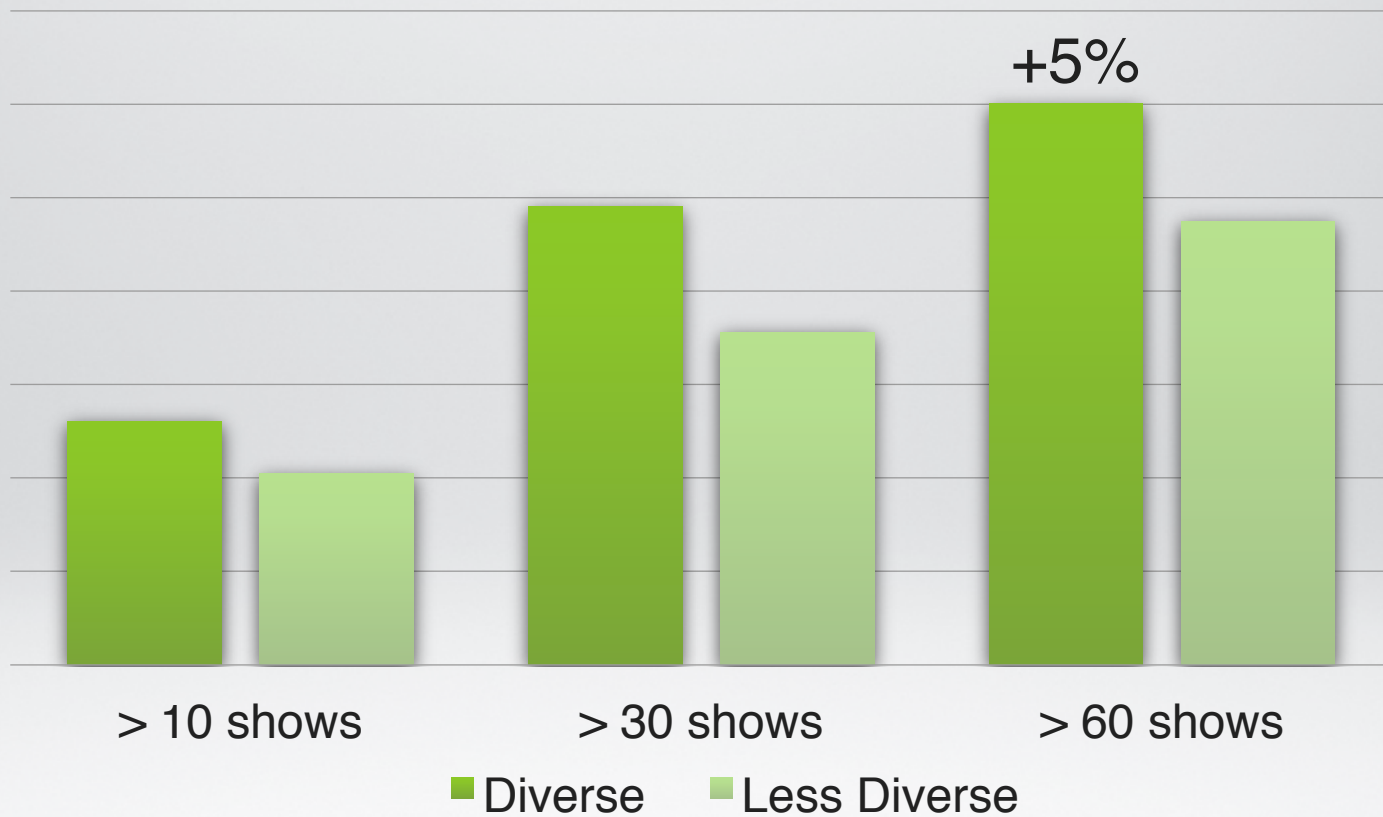
# 时间效应



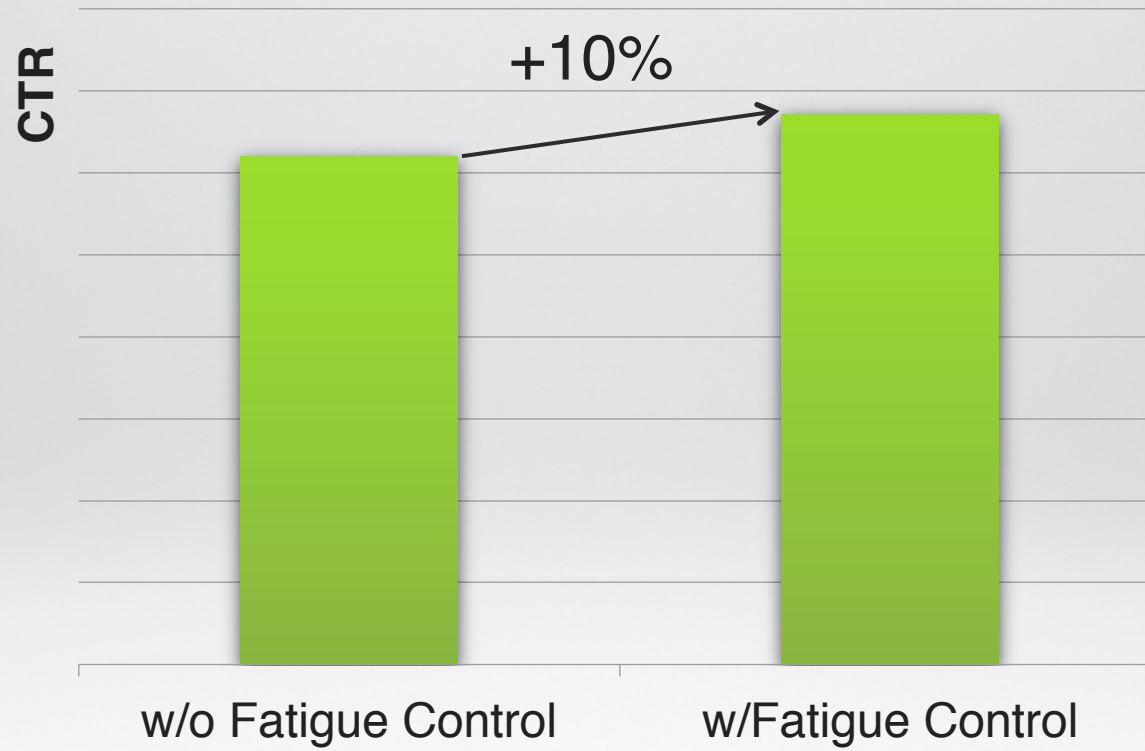


# 多样性

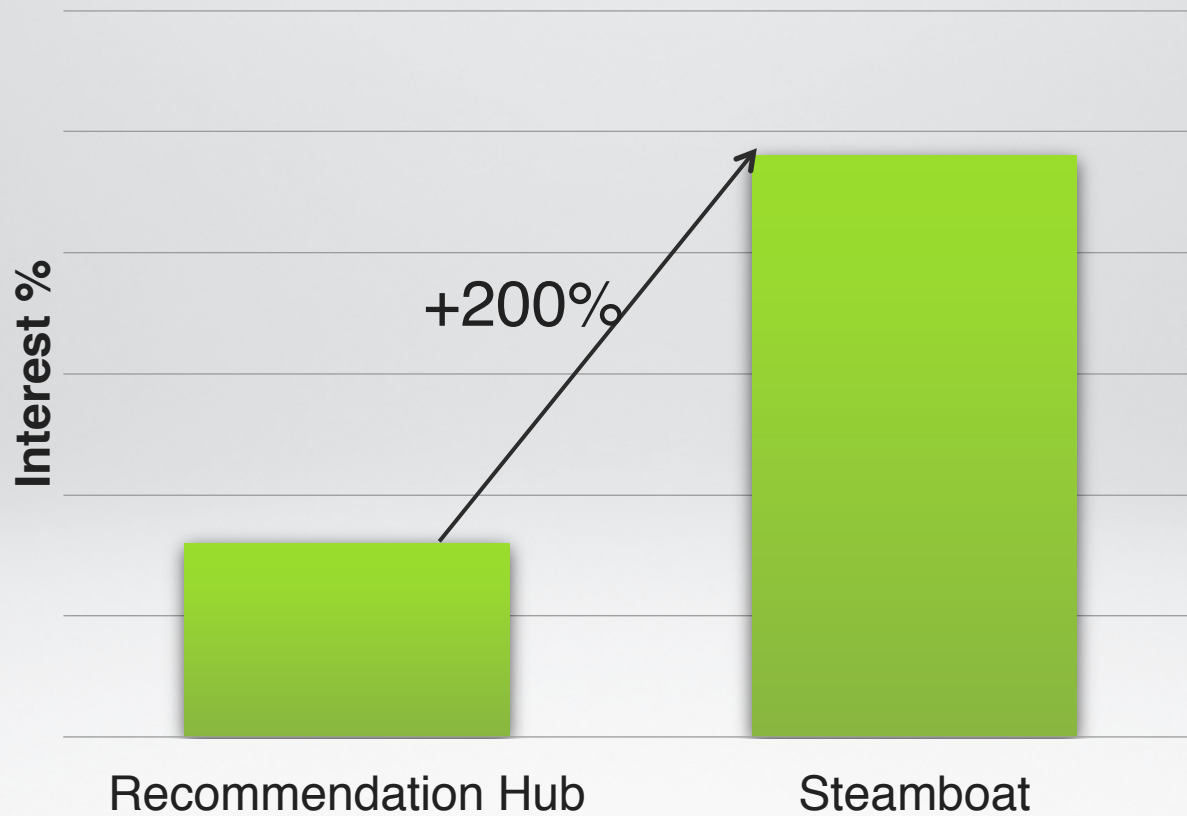
CTR



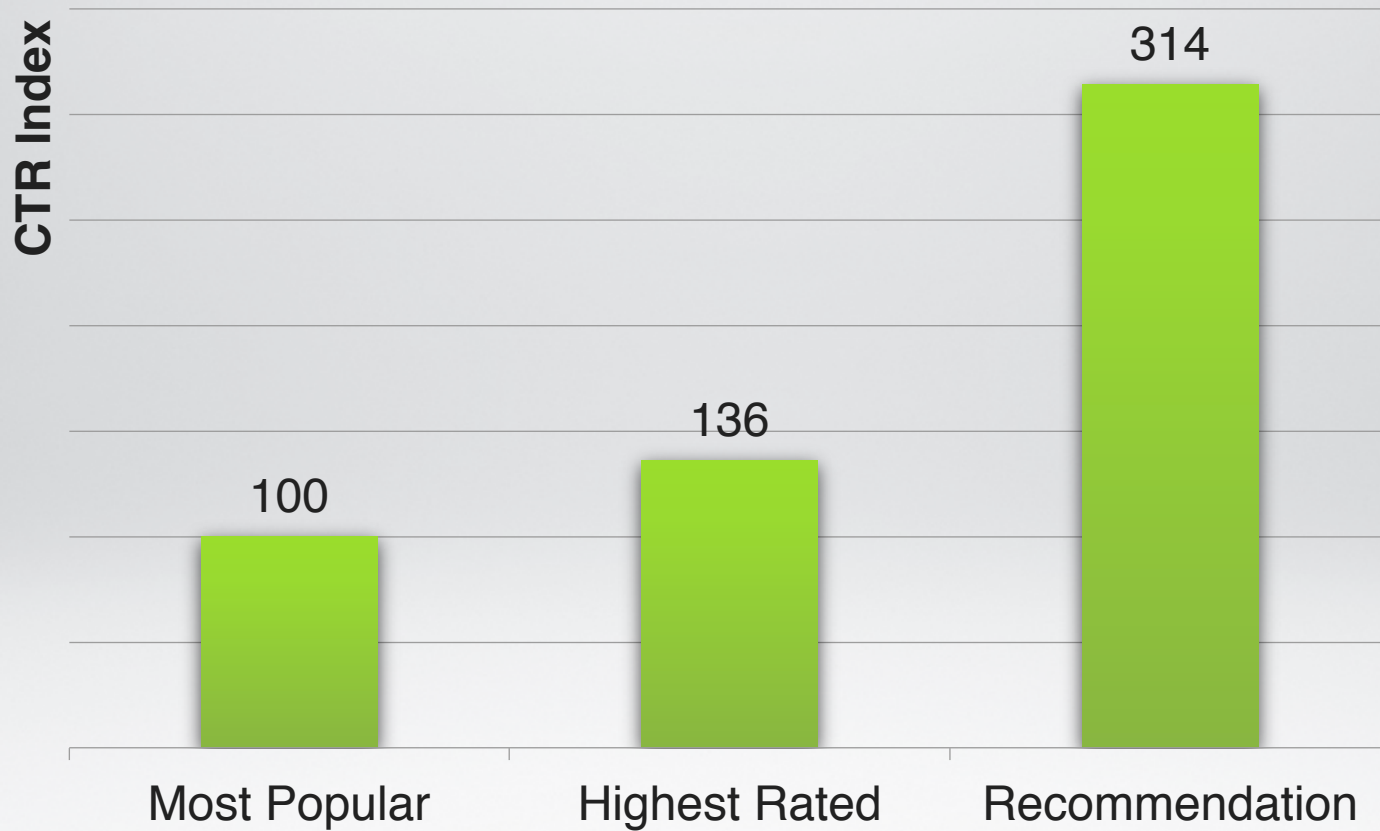
# 新鲜度



# 产品设计



## 推荐效果





# 总结

- 了解你的产品和用户
- 设计不同的功能或产品满足用户不同的需求
- 数据的规模和质量很重要
- 用户的各种反馈都可以想办法使用提高推荐效果
- 算法不能盲目上线，多听领域专家、产品经理和用户的反馈，多做离线评估和A/B测试
- 目前Hulu 大约**50%**的流量来自于推荐系统
- 下一步：
  - 个性化在移动设备和客厅设备上的优化(产品，算法)
  - 对内容更深入的理解

# 谢谢大家！

email: [zhenghua@hulu.com](mailto:zhenghua@hulu.com)

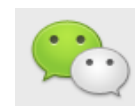
weibo: [@zhenghua00](#)

www.infoq.com/cn

# InfoQ<sup>new</sup>



@InfoQ



infoqchina

软件  
正在改变世界!