

WOTA

51CTO

World Of Tech 2017

全球架构与运维技术峰会

2017年4月14日-15日 北京富力万丽酒店

ARCHITECTURE



出品人及主持人：

**于雪**

51CTO WOT大会主编

---

网络性能优化实践

# 数据驱动 决策辅助与 产品智能化



**张宇石** 美团点评  
高级架构师

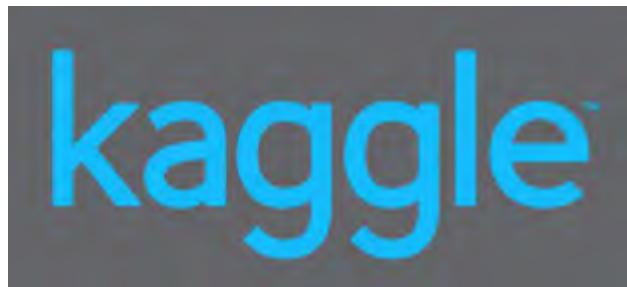
分享主题：

美团点评移动网络优化实践

# 关于我

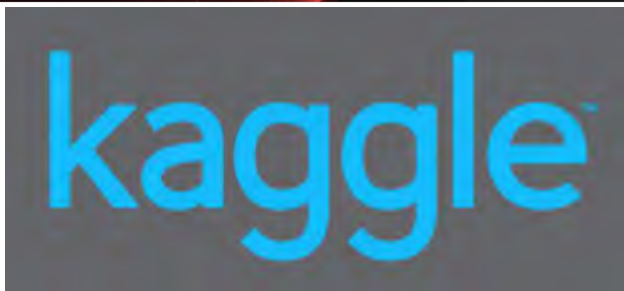


# 从Kaggle说起

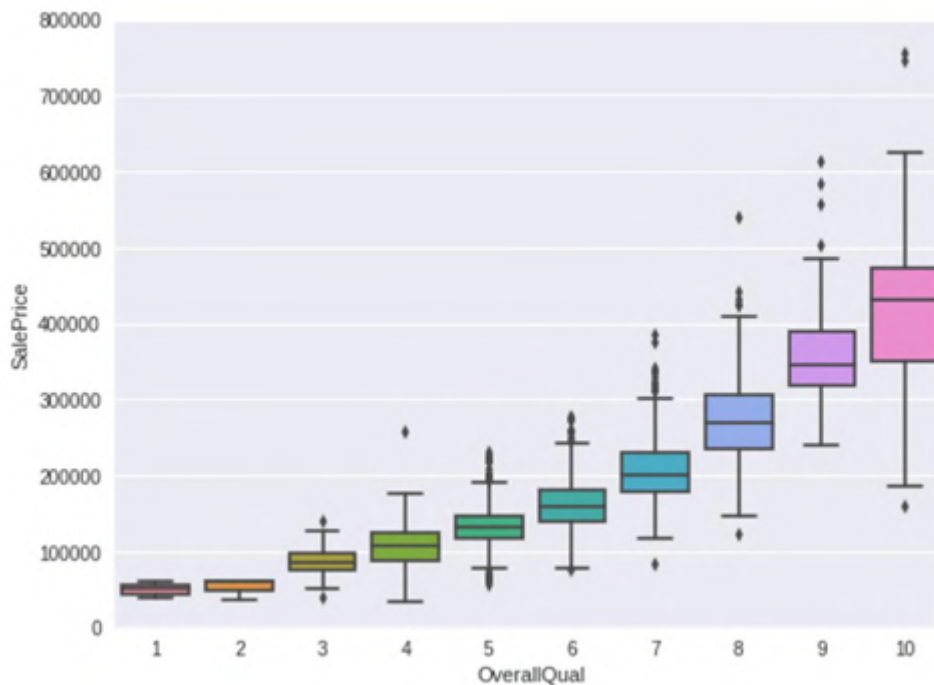


- 大数据竞赛平台
- 玩数据、ML 的开发者们展示功力、扬名立万的江湖
- 招聘服务
- 代码分享工具(Kaggle Kernels)

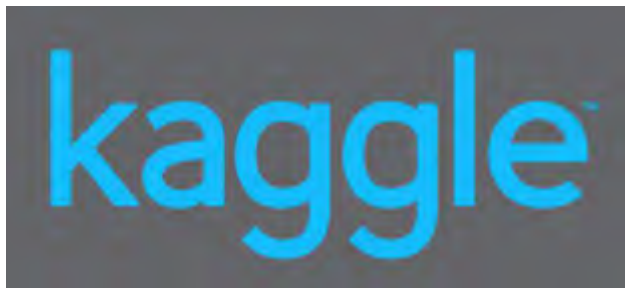
# 从Kaggle说起



- 回归分析预测房价
  - 79个解释变量
  - 质量打分, 形状(规则/不太规则/很不规则), 居住面积, 路面(铺碎石/柏油路)



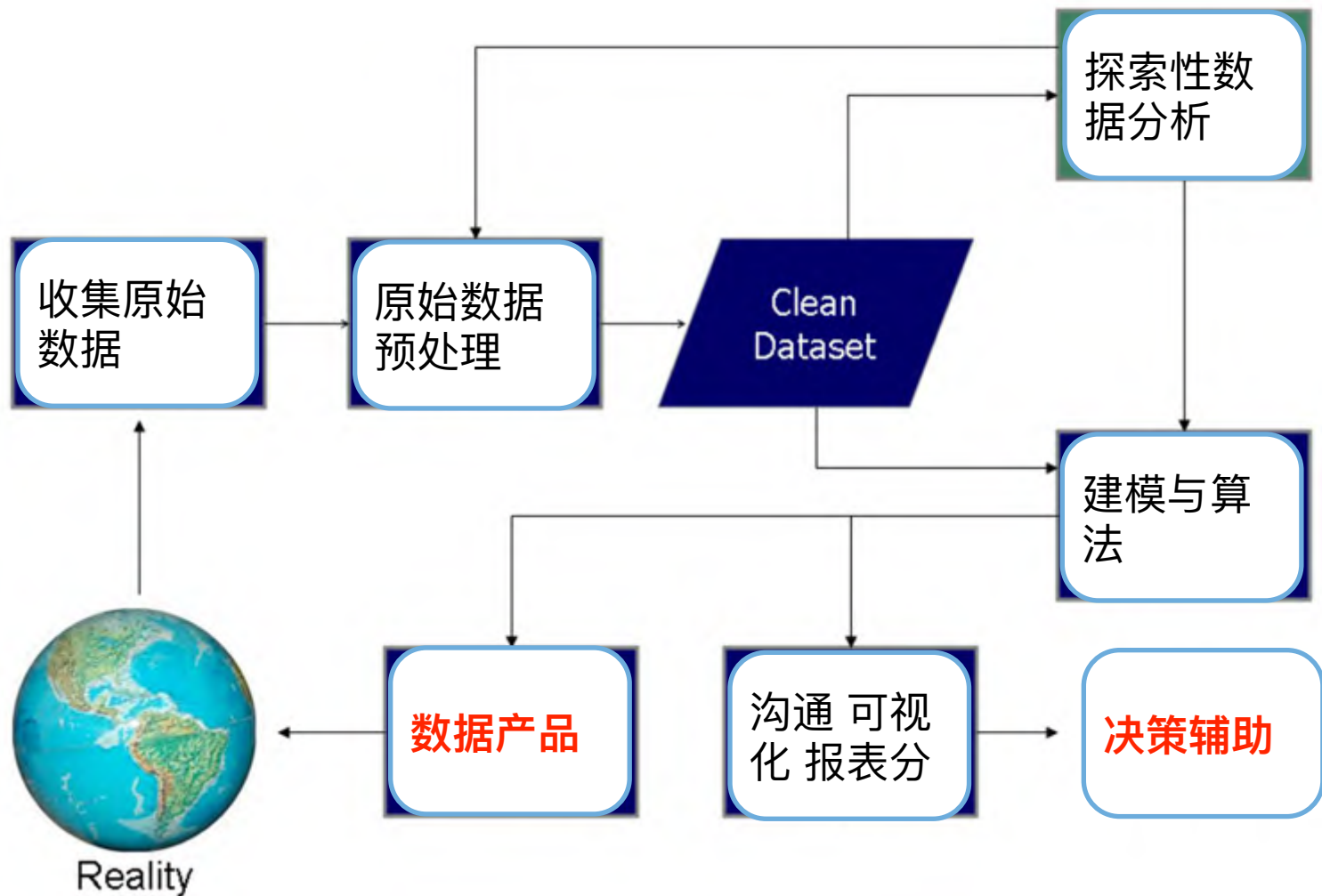
# 从Kaggle说起



- 泰坦尼克乘客二元分类
- 谷歌视频打标签



# 数据科学流程



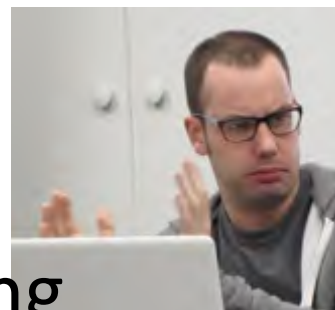
# 数据科学家在做什么？





## Analytics

- **问题导向**
- 购物平台用户工作时间 vs 下班后消费习惯差异
- 交互式分析
- 决策辅助
- Ad-hoc, 事后分析, 归因



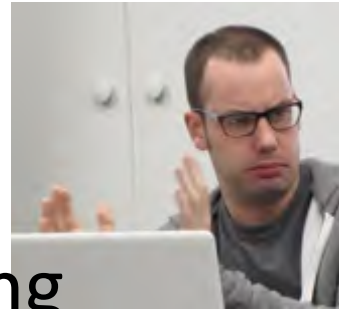
## Machine Learning

- **指标驱动**
- 提高广告平台用户转化率
- 智能化产品(e.g. 广告系统)
- 规模化, 自动化



## Analytics

- 菜鸟
  - 明确具体的分析任务
  - 产品局部
- 老司机
  - 提出问题并解决
  - 全局考量
- 产出：报告图表ppt



## Machine Learning

- 菜鸟
  - 特征工程
  - 渐进式改变
- 老司机
  - 系统架构设计
- 产出：自动采集数据+决策的智能产品

# 机器学习应用

搜索引擎

计算广告学

**推荐系统**

计算机视觉 语音书写识别 模式识别

机器翻译

医疗大数据

金融大数据

**精准营销**

风控

自然语言处理/理解/生成

自适应个性化网站

# 数据驱动的创业公司 Stitch Fix

- 在线个性化服装推荐
- 全职员工 + 兼职造型师 + 仓库工作人员
- **购物痛点**
  - 忙碌!
  - 发现
  - 追随时尚潮流



# 数据团队 @stitch fix

## 客户团队

- 精准营销
- 需求预测
- 用户画像
- 客服分析

## 推荐团队

- 推荐算法
- Human computation
- 造型师行为分析

## 库存团队

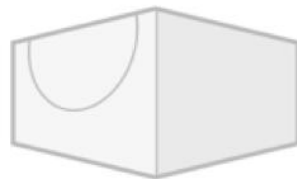
- 库存预测
- 基于算法清仓
- 打标签

**数据平台** (大数据架构 | 分析流程)

# Stitch fix业务模式



回答个人风格问卷



收到5件服装快递



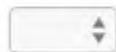
留下喜欢的免费退其他

How do you prefer clothes to fit the top half of your body?

- Mostly Tight / Form Fitting
- Prefer Fitted / Showing my Figure
- Straight
- Mostly Loose
- Oversized

How do you prefer clothes to fit the bottom half of your body?

Jeans waist



Skirts

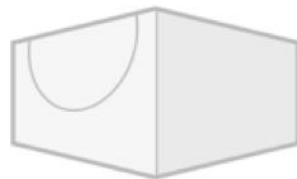




# Stitch fix业务模式



回答个人风格问卷



收到5件服装快递



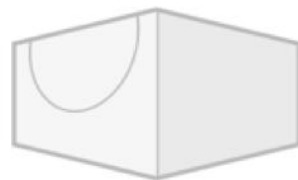
留下喜欢的免费退其他



# Stitch fix业务模式



回答个人风格问卷



收到5件服装快递



留下喜欢的免费退其他



# 人机协同 : $1 + 1 > 2$



繁重的重复性计算

大量工作记忆

大量长期记忆



非结构化数据

美学评估

营造人际关系

Novel inferences

Context sensitivity/nuance

# 人机协同 : 1 + 1 > 2



- 对大量库存SKU筛选和排序
- 从大规模数据中找到Pattern
- 降噪

# 人机协同 : 1 + 1 > 2



- 处理非结构化数据
- 情感沟通
- 创造性
- 算法开发免于考虑边缘情况

*I am going on a  
cruise next month  
- send me some  
dresses for sunny  
weather!*

*Dear Molly,*

*It was such a pleasure  
to pick out some items that  
I know will be perfect for  
you!*

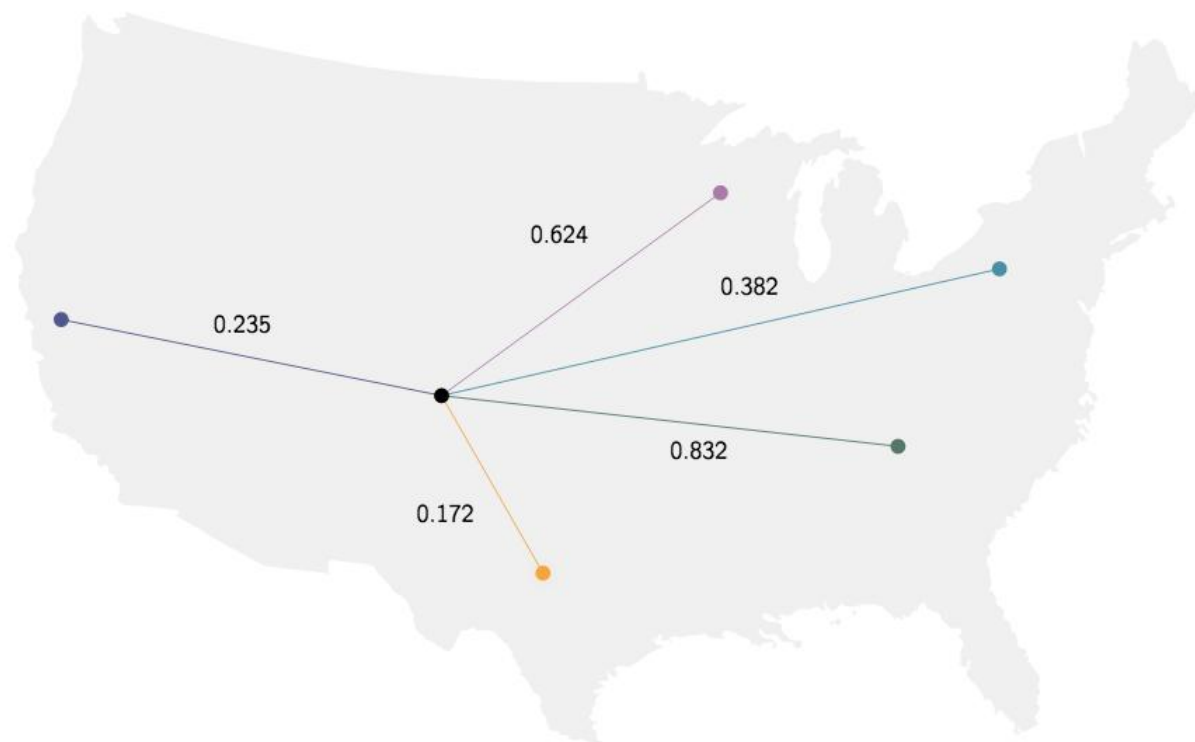
*I hope you love your  
Fix.*

*Best Wishes,  
Margaret*

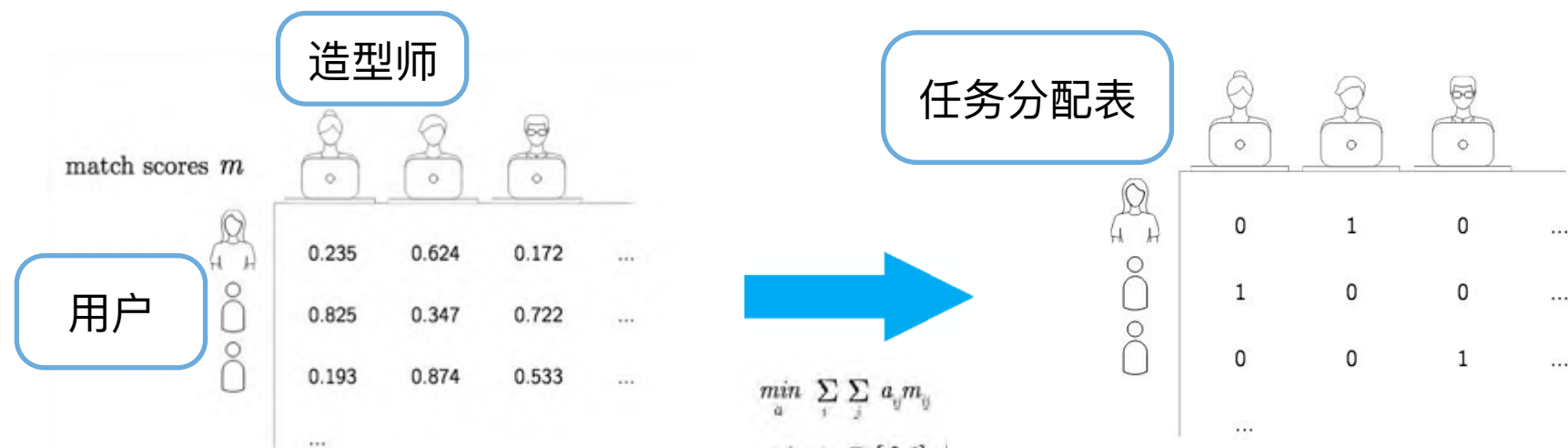


# 智能化物流 — 仓库分配

- 单一仓库发货 单一包裹
- 选仓发货优化问题  
(运费, 投递时间, 库存匹配, ...)



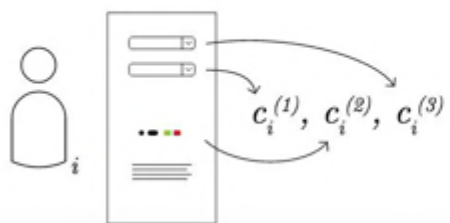
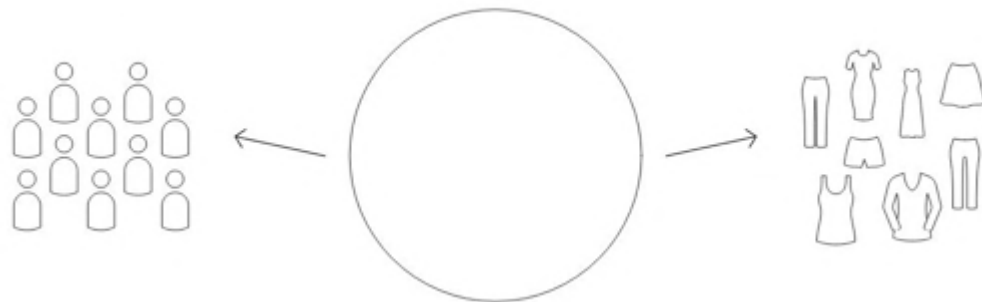
# 智能化物流 — 造型师匹配



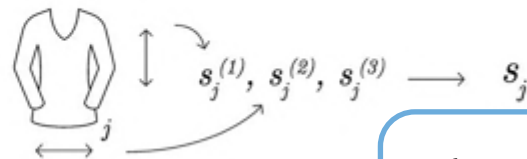
用户造型师匹配表:

(交易历史, 用户打分, 资料匹配, ..)

# 智能化物流 — 人货匹配



用户特征



产品特征

向量点积

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3.$$

用户特征

产品特征



# 用户问卷特征

- 年龄, 位置, 职业
- 身材尺寸
- 颜色价格偏好
- 样式彩虹(Style rainbow)
- 位置
- 隐式尺寸

Erin Sturm, 30

Mom Petites

Richfield, MN [\(weather\)](#)

## Profile Notes

I'm looking for more summer clothes.

Classic	Rom.	Boho	Edgy	Glam	Casual	Preppy
3	4	4	4	2	4	2

[Pinterest](#)

5'2" 120 lbs. Occupation: Client Services

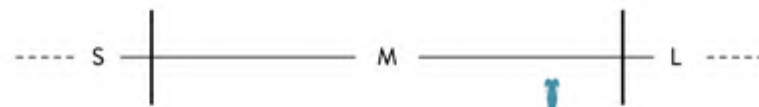
Bra	Tops	Dresses	Bottoms
34 DD	4 S	2 S	2 S

Top Fit: **Tight**

Bottom Fit: **Tight**

## Petite Preferences

Tops:	<b>Regular</b>	Pants/Denim:	<b>Petite</b>
Dresses:	<b>Regular</b>	Skirts:	<b>Petite</b>



# 产品特征

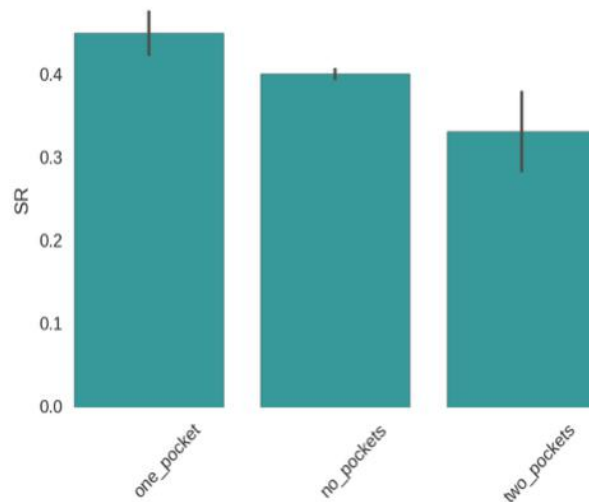
Inseam



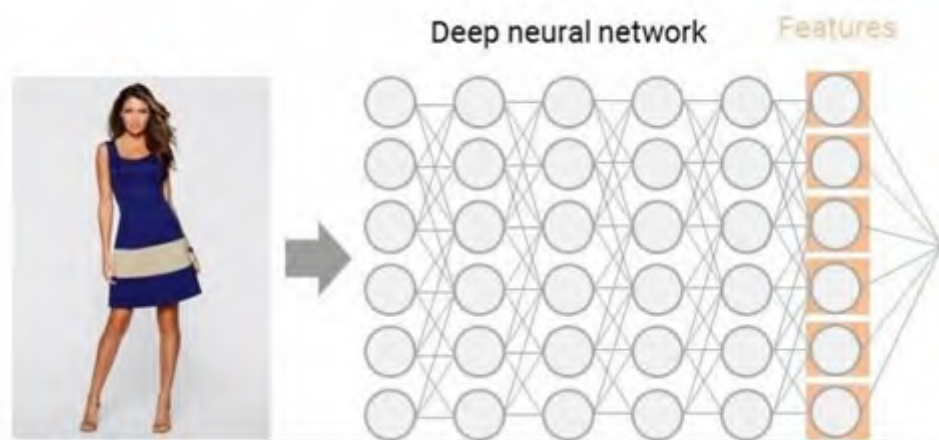
颜色



打标签



深度神经网络特征



产品相似  
度矩阵

k近邻法  
推荐

# 推荐算法



- Mixed-effect logistic regression 混合效应逻辑回归

匹配反馈



用户特征



?	0.83	?	0.54	?	0.47	0.23	...
---	------	---	------	---	------	------	-----



0.27	?	0.92	?	0.13	0.59	0.14	...
------	---	------	---	------	------	------	-----



?	?	0.85	0.76	?	0.62	0.90	...
---	---	------	------	---	------	------	-----

产品特征

0.21	0.74	0.53	0.26	0.85
⋮	⋮	⋮	⋮	⋮

$$\log \frac{p}{1-p} = a + \underbrace{X\beta}_{\text{固定效应}} + \underbrace{Zb}_{\text{随机效应}}, \quad b \sim N(0, \Sigma)$$

固定效应

随机效应

# 推荐算法的挑战

- 以什么指标对库存排序？

**Naive**方案：忽略造型师选择，对交易数据建模

$$\mathbb{P}(Y_i = 1) \approx \mathbb{P}(Y_i = 1 | S_i = 1)$$

好处

购买

选中

- 传统机器学习问题
- 交易数据

# 推荐算法的挑战

- 删失数据

Arms

Less is more, keep it covered

用户偏好



购买

$$Y_i | S_i = 1$$

产品：无袖

是 否

是

?

p

否

?

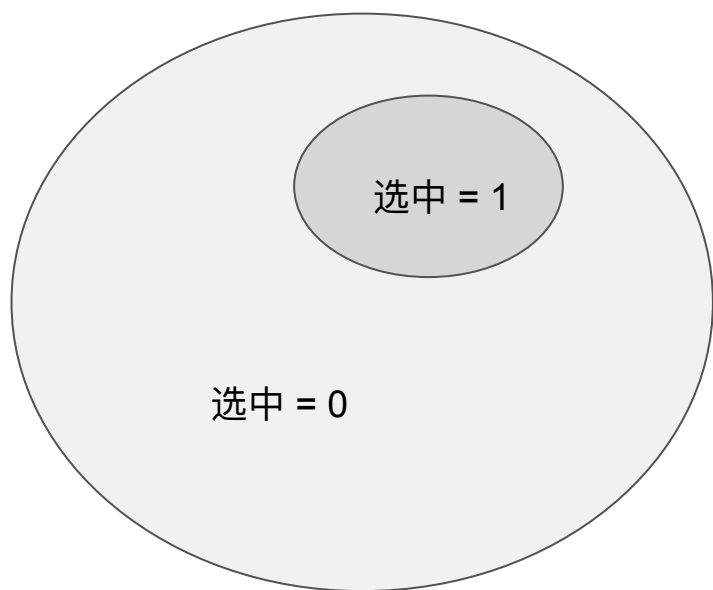
1-p

	是	否
是	?	p
否	?	1-p

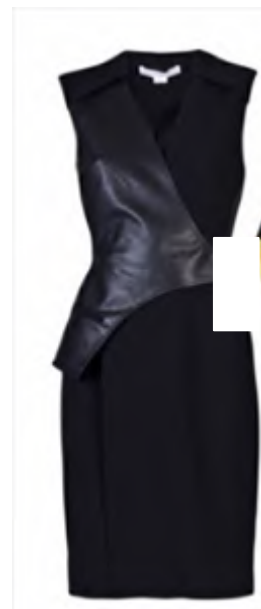
# 推荐算法的挑战

- 购买率不一定是好的排序指标

低覆盖率产品



用户群



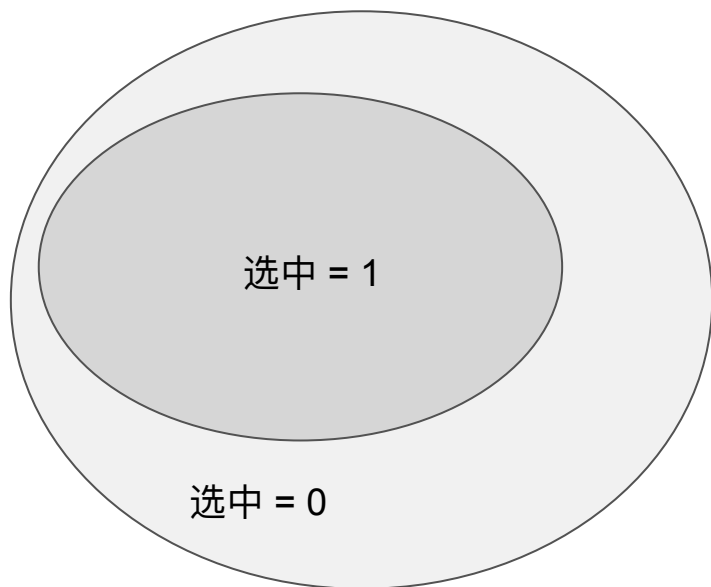
一条前卫连衣裙

高购买率

# 推荐算法的挑战

- 购买率不一定是好的排序指标

高覆盖率产品



中性产品

低购买率

# 推荐算法的挑战

- 购买率不一定是好的排序指标

$$\mathbb{P}(Y_i = 1 | S_i = 1)$$



排前



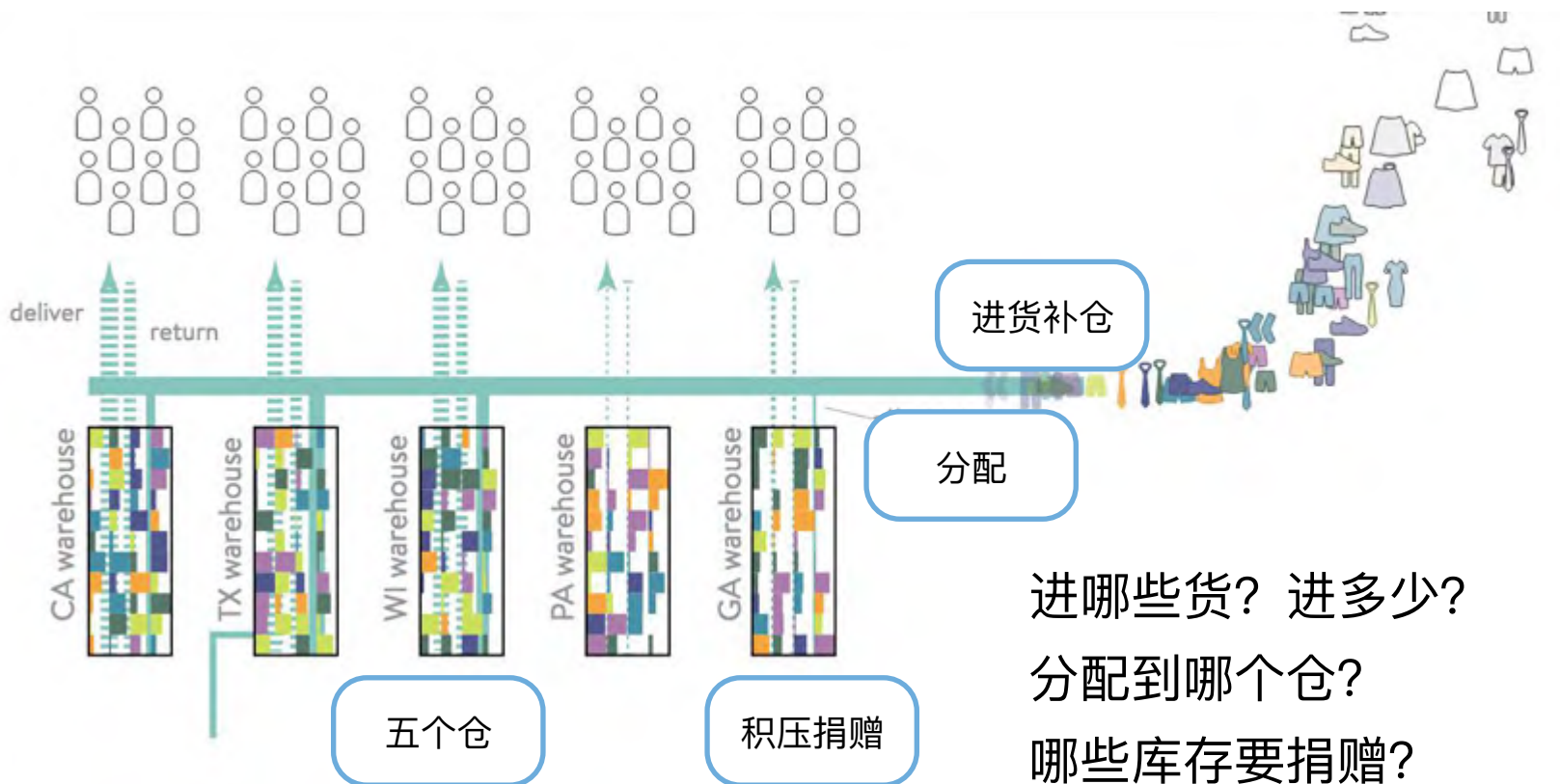
排后

造型师选择有针对性

建模预测造型师选择



# 智能化物流 — 库存管理



# 数据团队 @stitch fix

## 客户团队

- 精准营销
- 需求预测
- 用户画像
- 客服分析

## 推荐团队

- 推荐算法
- Human computation
- 造型师行为分析

## 库存团队

- 库存预测
- 基于算法清仓
- 打标签

**数据平台** (大数据架构 | 分析流程)

# 结语

- 兴趣

高山仰止，虽不能至，然心向往之

- 实战

千里之行始于足下

- 分享

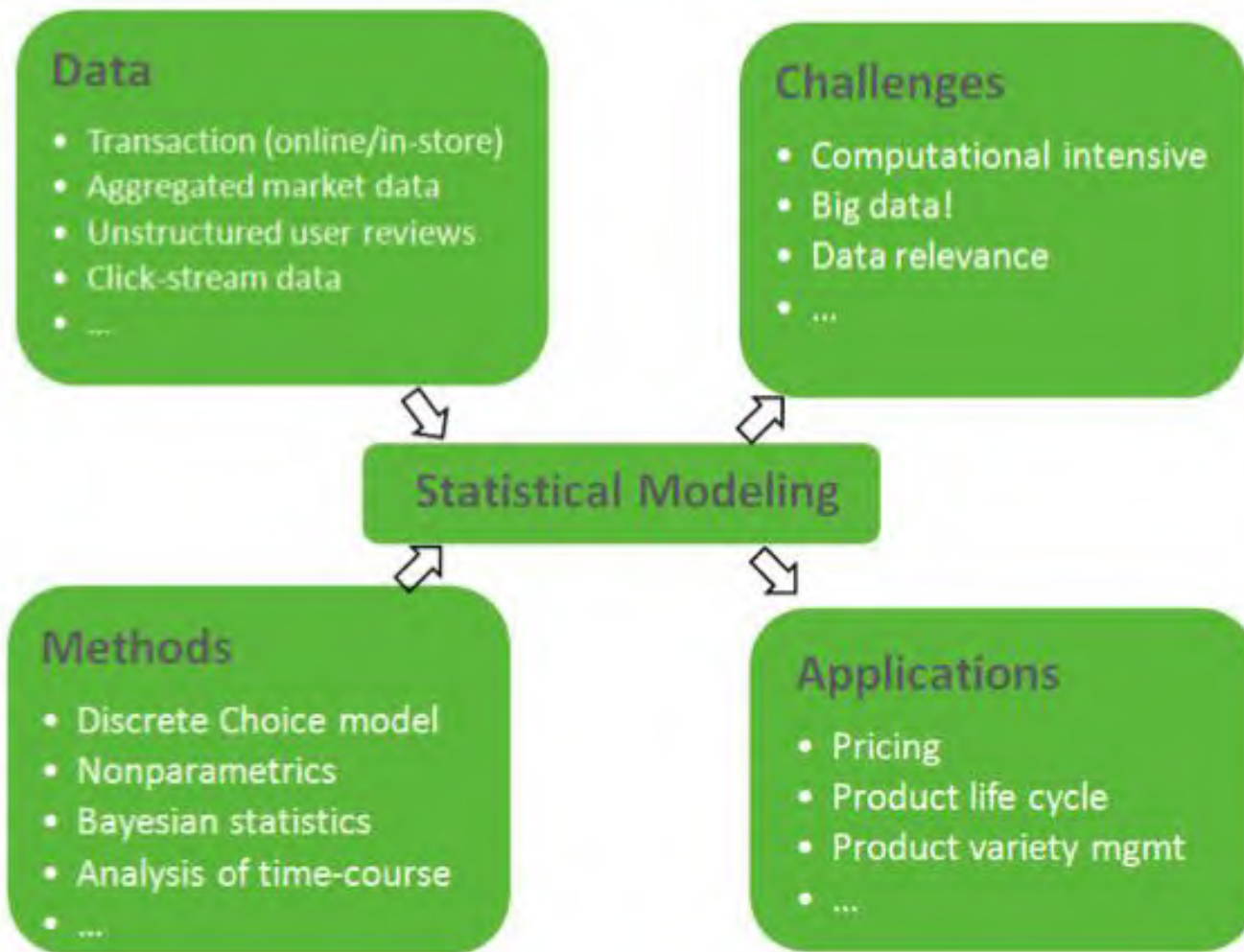
独乐乐不如众乐乐



Thank you !

Jay Wang (王建强)

# 线下到线上数据



# 大数据既是挑战也是机遇

- 挑战：存储、处理、建模
- 机遇：
  - 数据质与量决定了学习的上限
- 智能中枢 vs 运动中枢
  - (rules learned from data or humans, learning from data)



# 数据驱动的创业型公司

<b>Jet.com</b>	Amazon killer: subscription-based retail, Marc Lore (Diapers.com), \$50/yr, 5-10% lower price
<b>Thumbtacks</b>	Service provider referral (how to monetize?)
<b>SpotTrender</b>	Pre-test video commercials
<b>Sano</b>	Realtime news discovery from social networks (twitter, instagram, weibo, VK, ..)
<b>Common crawl</b>	(non-profit) Open repo of web crawl data, billions of pages each month

# 新加一个创新型公司

背景 太白

增加国内栗子



# 数据驱动的创业型公司

<b>Jet.com</b>	Amazon killer: subscription-based retail, Marc Lore (Diapers.com), \$50/yr, 5-10% lower price
<b>Thumbtacks</b>	Service provider referral (how to monetize?)
<b>SpotTrender</b>	Pre-test video commercials
<b>Sano</b>	Realtime news discovery from social networks (twitter, instagram, weibo, VK, ..)
<b>Common crawl</b>	(non-profit) Open repo of web crawl data, billions of pages each month

# Analytics

漏斗分析

用户留存及LTV

# 精准营销 — 用户画像



- **固定特征**
  - 年龄，生日，性别，教育水平，职业等
- **兴趣特征**
  - 兴趣爱好，使用APP，网站，浏览/评论内容，品牌偏好，产品偏好等
- **社会特征**
  - 生活习惯，婚恋情况，社交/信息渠道偏好，宗教信仰等
- **消费特征**
  - 收入状况，购买力水平，商品种类，购买渠道喜好，购买频率等
- **动态特征**
  - 当下时间，正在前往的地方，当下的需求，周围环境等

# 大数据既是挑战也是机遇

- 挑战：存储、处理、建模
- 机遇：数据质与量决定了学习的上限
  - 零售业
  - 图像识别