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

WOTA

51CTO

出品人及主持人：

李鹏涛 京东商城
青龙研发部高级总监

大数据应用创新

WOTA

51CTO

数据驱动的决策辅助与产品智能化

王建强 twitter/stitch fix

WOTA

51CTO

王建强

Stitch Fix数据科学总监
前Twitter美国总部技术主管

分享主题：

数据驱动的决策辅助与产品智能化

关于我

- 数据分析 + 机器学习

- HP Labs : 定价模型
- Twitter: 广告排序 点击率预测
- Stitch fix : 产品推荐 人货匹配

1. 数据科学的心得体会 2.数据驱动的零售业公司stitch fix

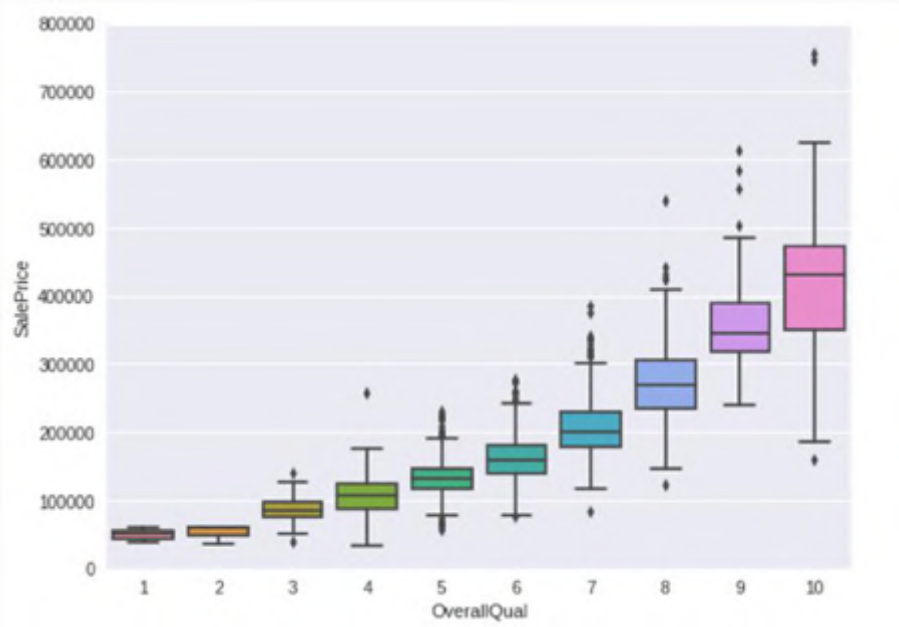
从Kaggle说起

- 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

kaggle

- **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)

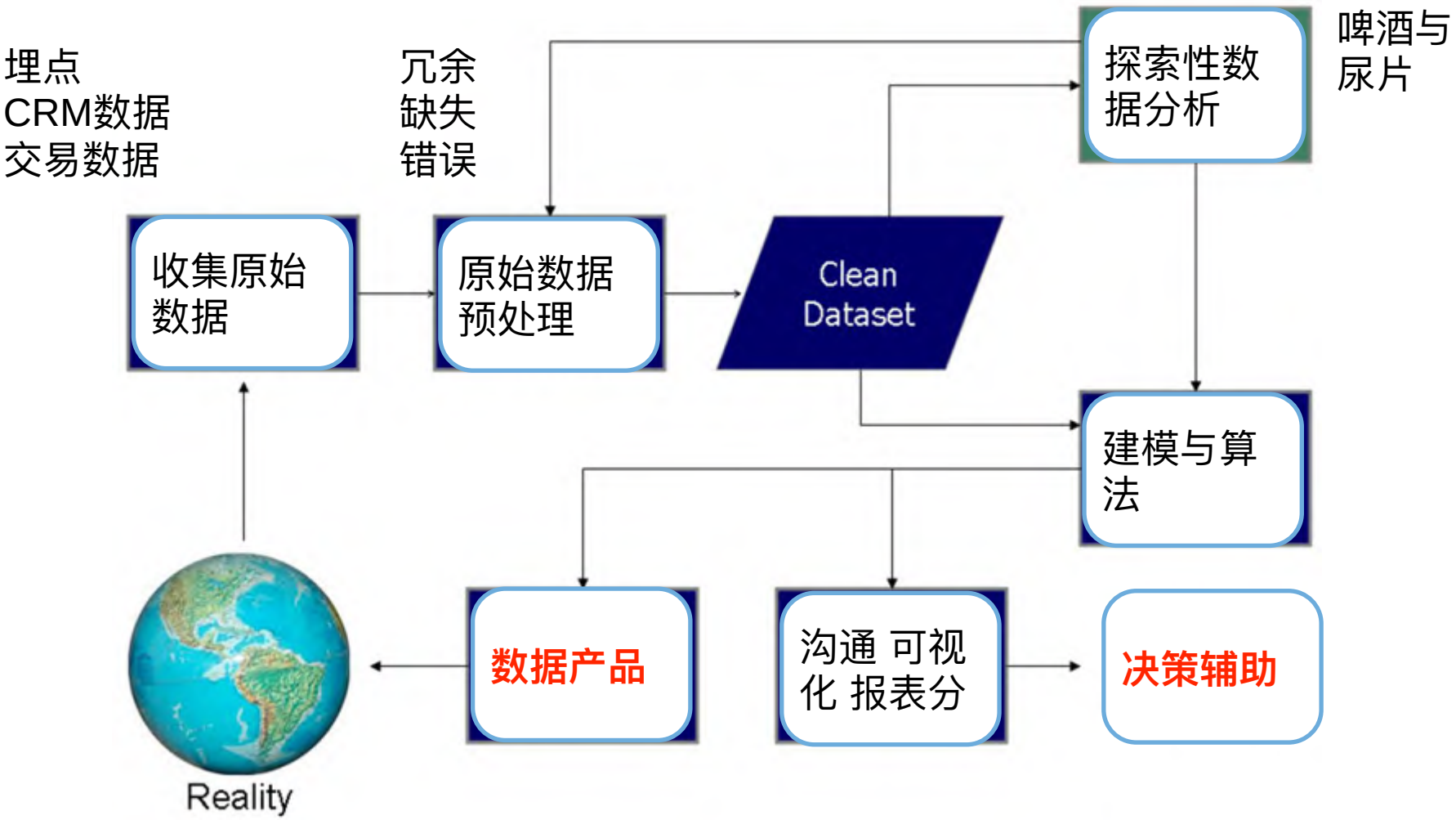


- **谷歌视频打标签**

视频数量：百万量级，
总标签数：4k~
训练集 vs 测试集

产出：标签及置信度

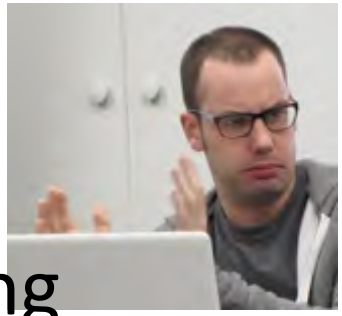
数据科学流程


埋点
CRM数据
交易数据
冗余
缺失
错误
啤酒与尿片
探索性数据分析
收集原始数据
原始数据预处理
Clean Dataset
建模与算法
数据产品
沟通 可视化 报表分
决策辅助
Reality

企业数据科学

- **企业的数据文化**
  - A/B 测试
  - 诺基亚 —> 苹果手机
- **沟通建立信任**
  - 失望的结果: 平台广告投放饱和
- **Actionable insights**
  - 活跃用户量周末大于周中, so what?
    - 用户变现?
    - 增长用户?

两类数据科学家



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

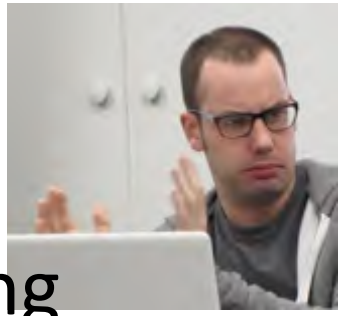
- **指标驱动**
- 提高广告平台用户转化率

- 智能化产品(e.g. 广告系统)
- 规模化，自动化

角色转换：戴两个帽子

Slide adapted from Josh Wills “From the Lab to the Factory”

两类数据科学家



Analytics

- 菜鸟
  - 分析任务明确具体
  - 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
  - 指引方向 资源调度

- **产出：自动采集数据+决策的智能产品**

Slide adapted from Josh Wills “From the Lab to the Factory”

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

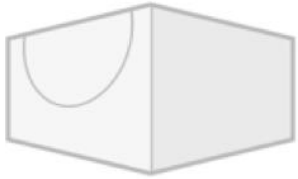
自适应个性化网站

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

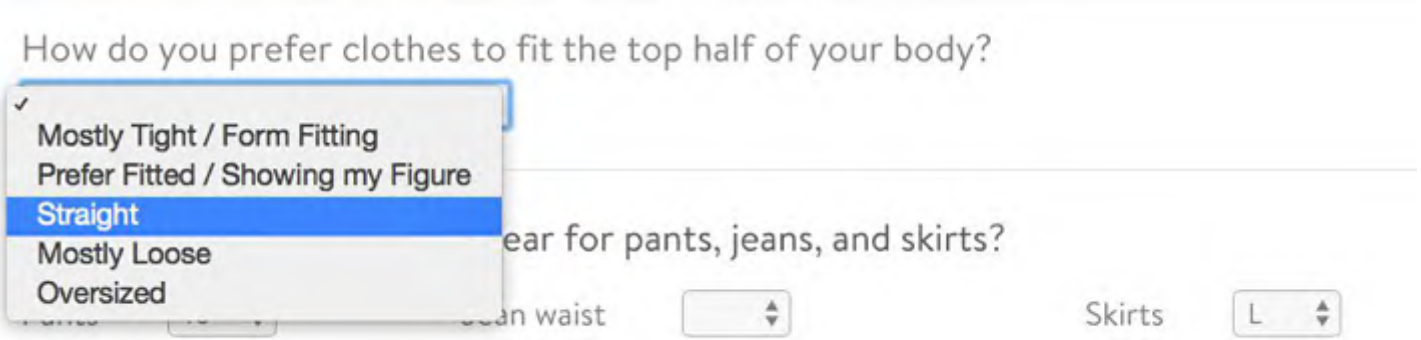
Stitch fix业务模式

回答个人风格问卷



收到5件服装快递

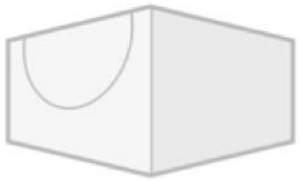
留下喜欢的免费退其他



你喜欢上身的fit 怎样?
紧身 / 合身的 / 直筒 / 宽松款等

Stitch fix业务模式

回答个人风格问卷

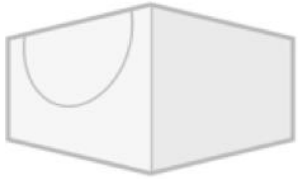


收到5件服装快递

留下喜欢的免费退其他

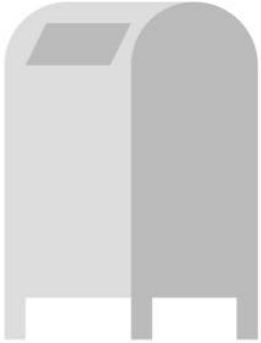
Stitch fix业务模式

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



人机协同推荐衣服：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

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
- Context sensitivity/nuance

数据团队 @stitch fix

- 公司~300人, 数据团队80人

客户团队
- 精准营销
- 需求预测
- 用户画像
- 客服分析

推荐团队
- 人货匹配
- 用户造型师匹配
- 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）

需求预测：用户稳定增长，需求的季节性，订阅式用户

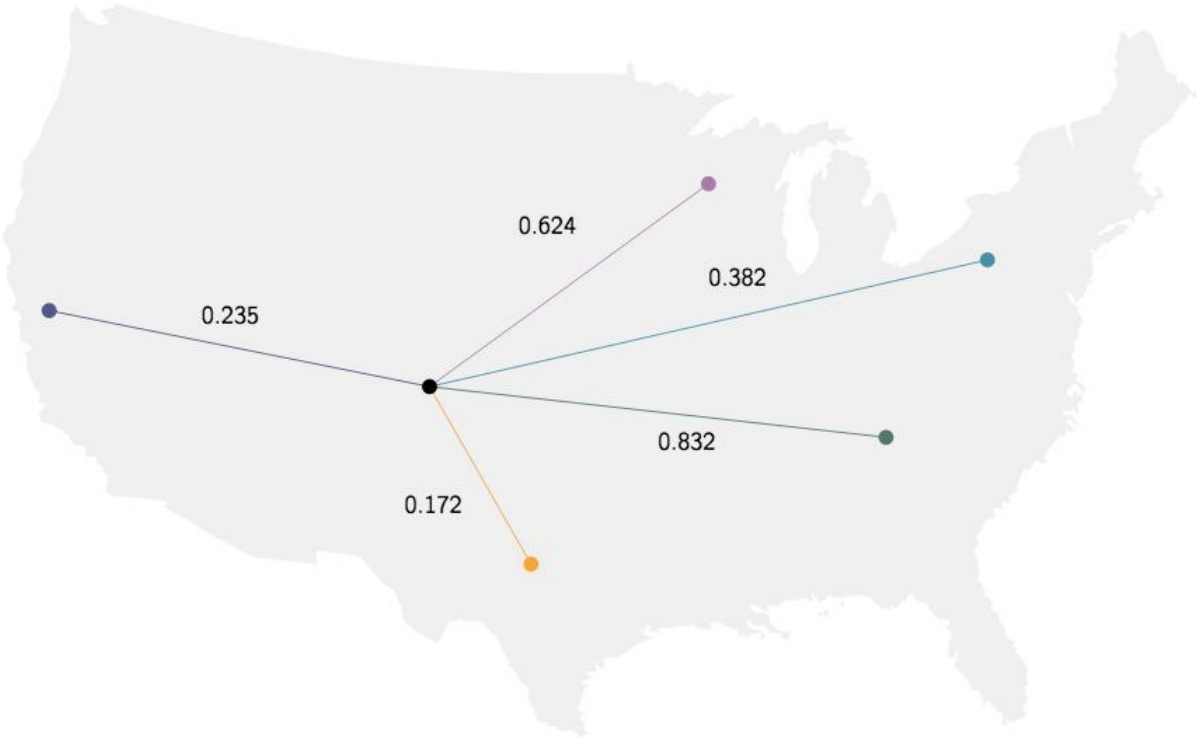
Human computation : 人机协同，虚拟环境下研究造型师行为

行为分析： 日志数据,造型师行为模式

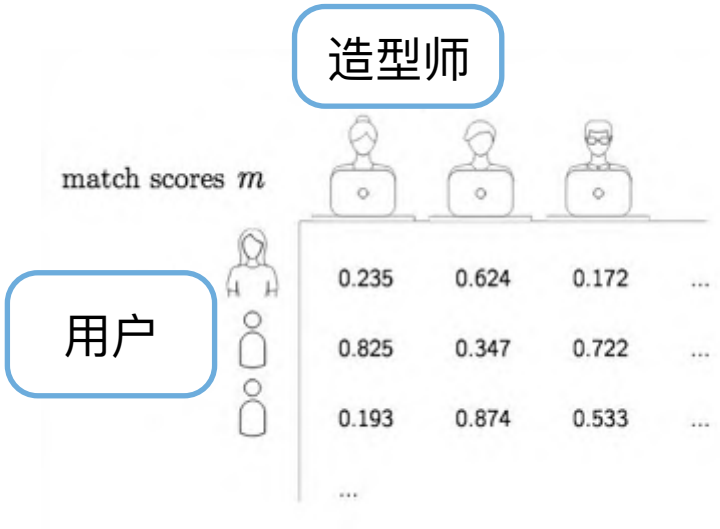
智能化物流 — 仓库分配

- 单一仓库发货 单一包裹
- 用户请求 —> 选仓发货

  （运费，投递时间，库存匹配，…）



智能化物流 – 造型师匹配



$$\min_{a} \sum_{i} \sum_{j} a_{ij} m_{ij}$$

$$s.t.\ a_{ij} \in \{0,1\},$$

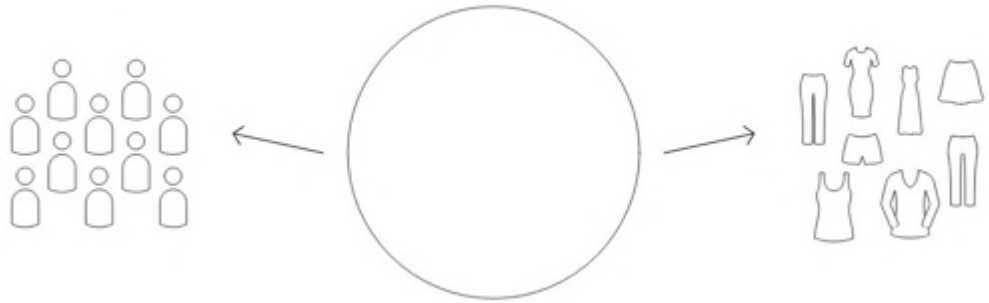
$$\sum_{j} a_{ij} = 1 \quad \forall i$$

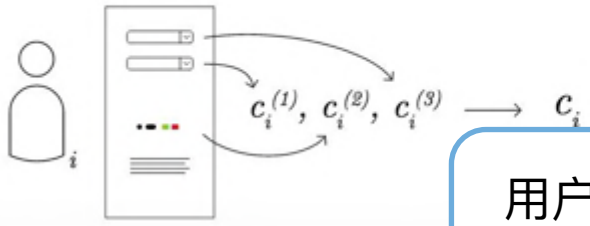
$$\sum_{i} a_{ij} < k_j \quad \forall j$$

用户造型师匹配表:

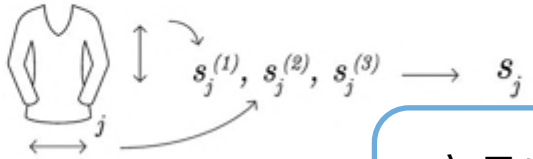
(交易历史，用户打分，资料匹配，..)

智能化物流 — 人货匹配





用户特征



产品特征

向量点积

$$\begin{bmatrix} a_1 \\ a_2 \\ \\ a_3 \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ \\ b_3 \end{bmatrix} = a_1b_1 + a_2b_2 + a_3b_3.$$

用户特征

产品特征

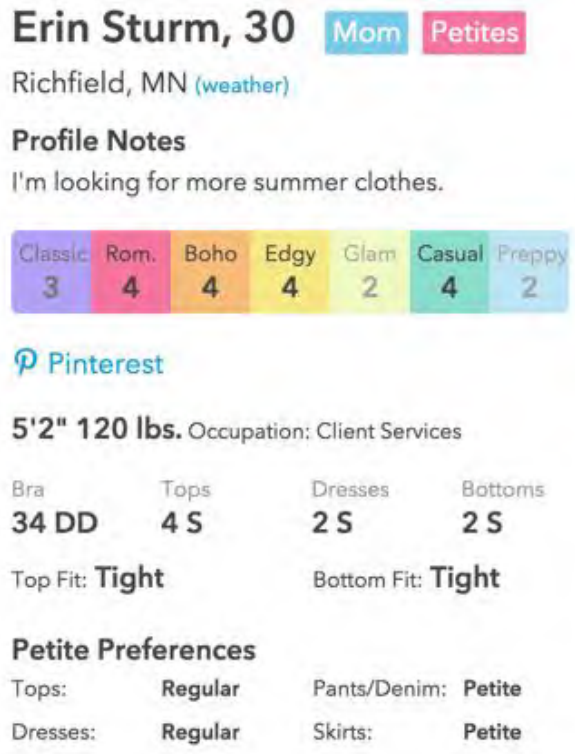
用户问卷特征

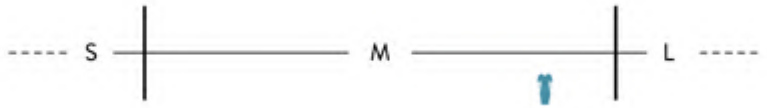
- 年龄, 位置, 职业
- 身材尺寸, 颜色价格偏好
- 样式彩虹(Style rainbow)

经典， 浪漫， 波西米亚风， 前卫，闪亮，休闲，**preppy**

制服风格,简约精致,带点叛逆+颓废

- 隐式尺寸



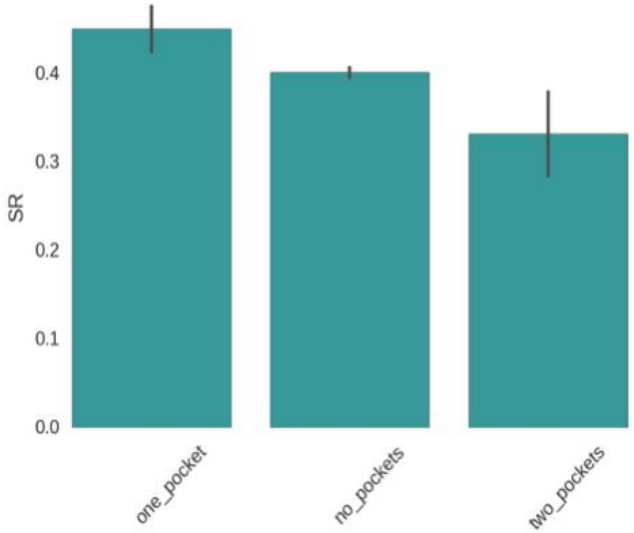


产品特征

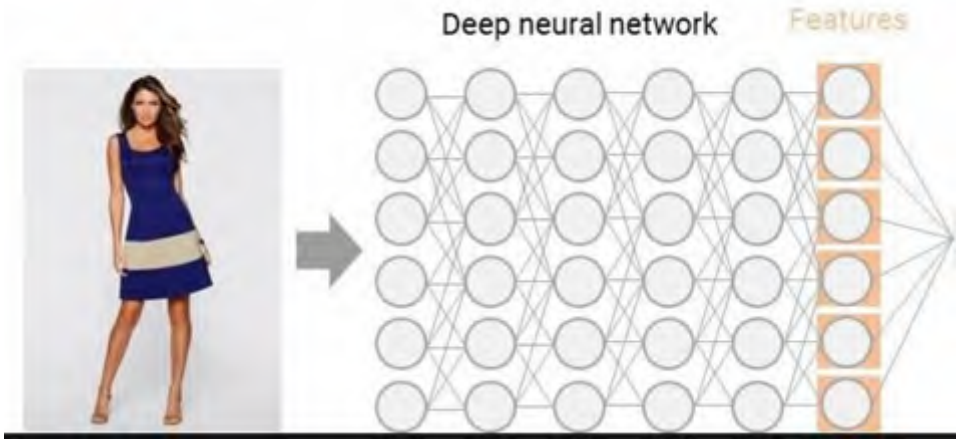
Inseam

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
⋮	⋮	⋮	⋮	⋮

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

推荐算法的挑战

- 排序指标?

Naive方案：忽略造型师选择，对交易数据建模

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

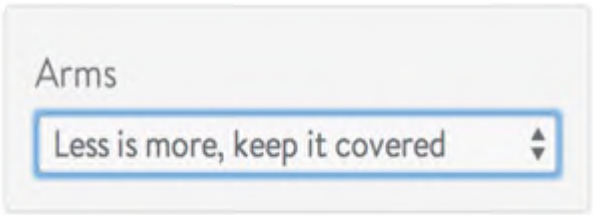
购买　　选中

好处

- 传统机器学习
- 交易数据

推荐算法的挑战

- 删失数据 censored data



用户偏好

$$Y_i|S_i = 1$$

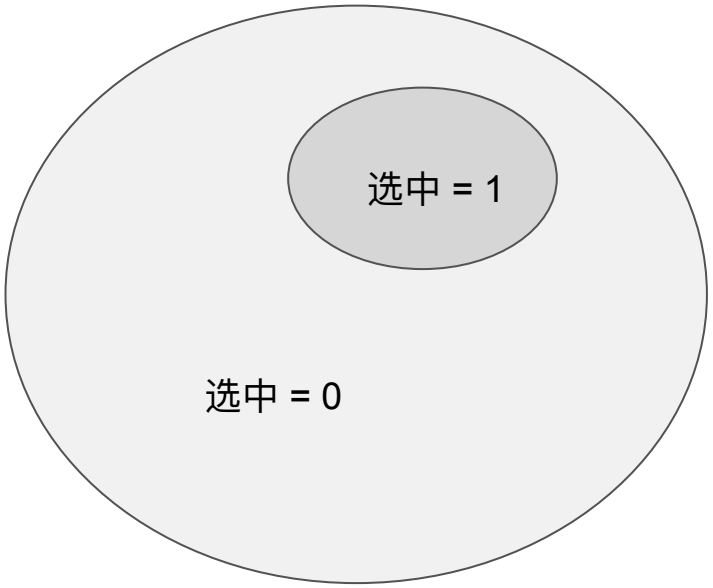
衣服：无袖

购买		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

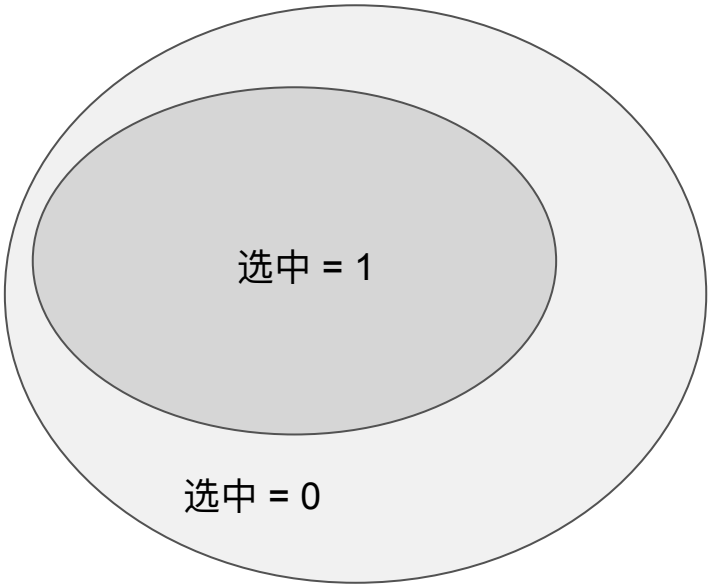
一条前卫连衣裙

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

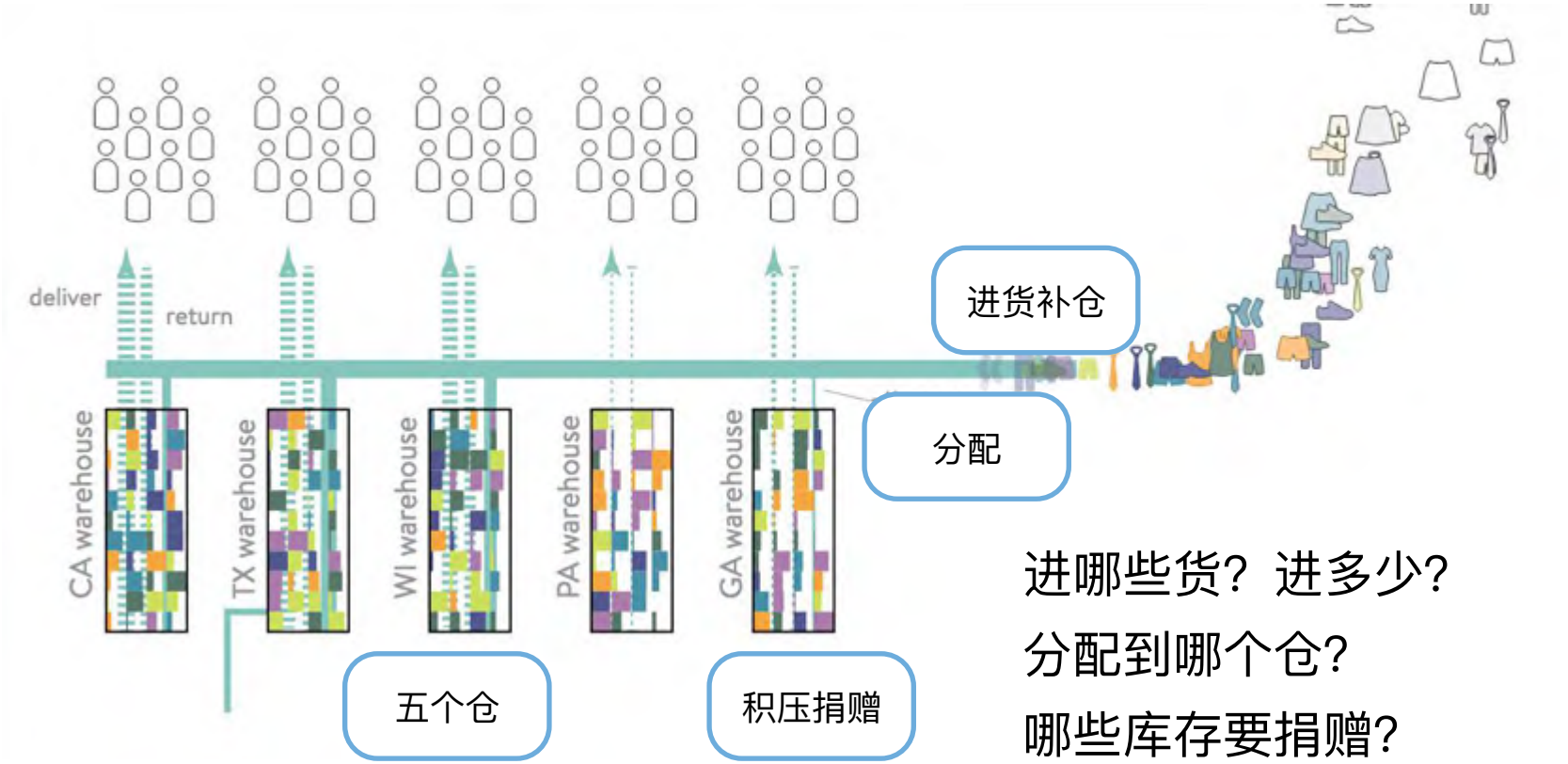
排后

造型师选择有针对性

建模处理 选择性偏差 (selection bias)

Heckman 两阶段模型

智能化物流 — 库存管理



进哪些货？进多少？

分配到哪个仓？

哪些库存要捐赠？

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

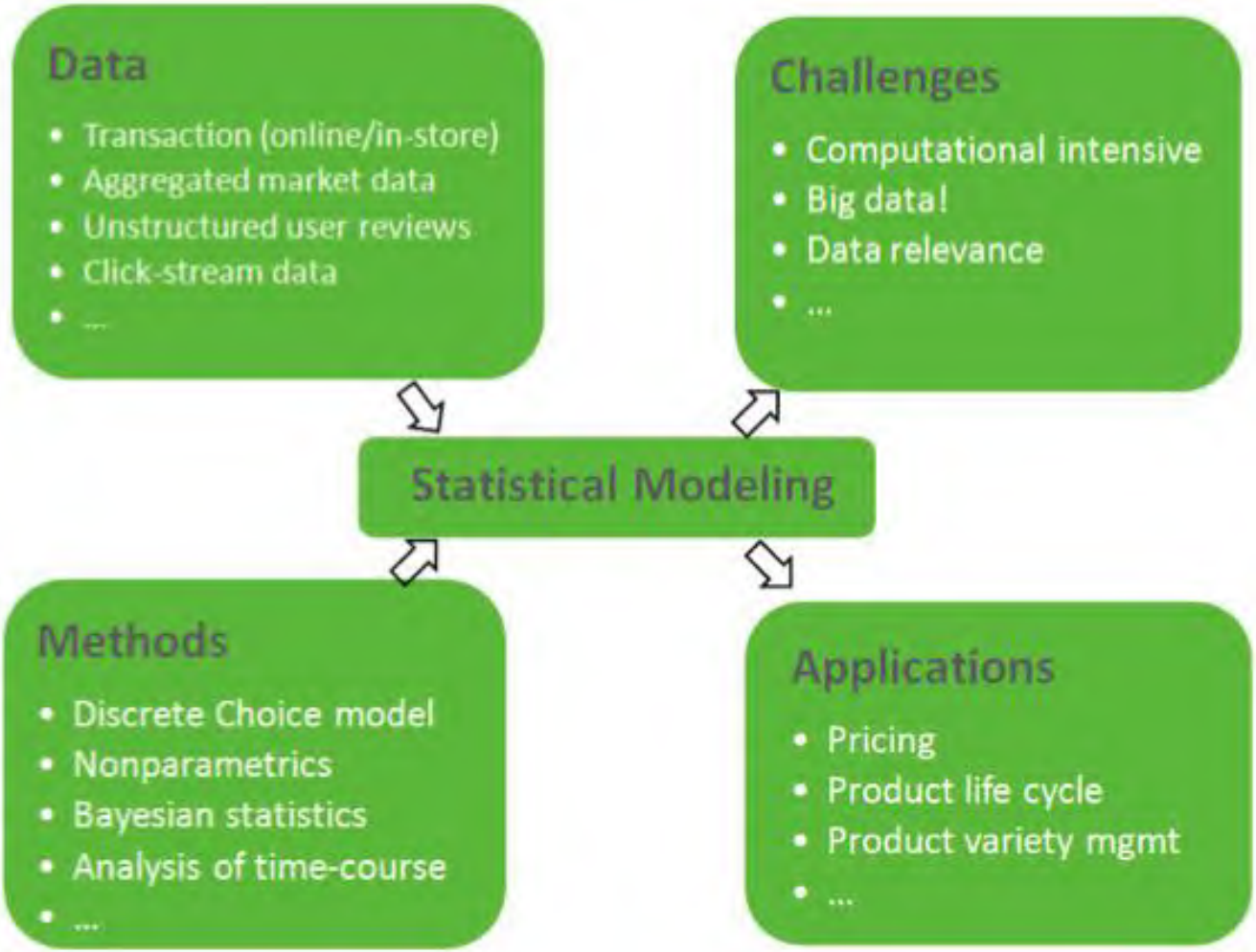
WOTA

51CTO

Thank you !

Jay Wang （王建强）

线下到线上数据


Data
Transaction (online/in-store)
Aggregated market data
Unstructured user reviews
Click-stream data
...
Challenges
Computational intensive
Big data!
Data relevance
...
Statistical Modeling
Methods
Discrete Choice model
Nonparametrics
Bayesian statistics
Analysis of time-course
...
Applications
Pricing
Product life cycle
Product variety mgmt
...

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)

Slide adapted from Josh Wills “From the Lab to the Factory”

数据驱动的创业型公司

Jet.com	Amazon killer: subscription-based retail, Marc Lore (Diapers.com), $50/yr, 5-10% lower price
Thumbtacks	Service provider referral (how to monetize?)
SpotTrender	Pre-test video commercials
Sano	Realtime news discovery from social networks (twitter, instagram, weibo, VK, ..)
Common crawl	(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

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

Analytics

漏斗分析

用户留存及LTV

数据科学家在做什么?


Sentiment Analysis
Statistics
Recommender Systems
Data Visualization
Computer Science
Pattern Recognition
Python
Distributed Systems
Statistical Modeling
Apache Pig
R
Java
Big Data
Natural Language Processing
Algorithms
Business Analytics
Weka
Decision Trees
Perl
Predictive Modeling
Machine Learning
SAS
Data Analysis
C++
Lucene
Predictive Analytics
Text Classification
Analytics
Computer Vision
Artificial Intelligence
Data Mining
Text Mining
Information Extraction
Neural Networks
Web Mining
Hive
Time Series Analysis
Computational Linguistics
Information Retrieval
Logistic Regression
Text Analytics
Matlab
MapReduce
LaTeX
Hadoop

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