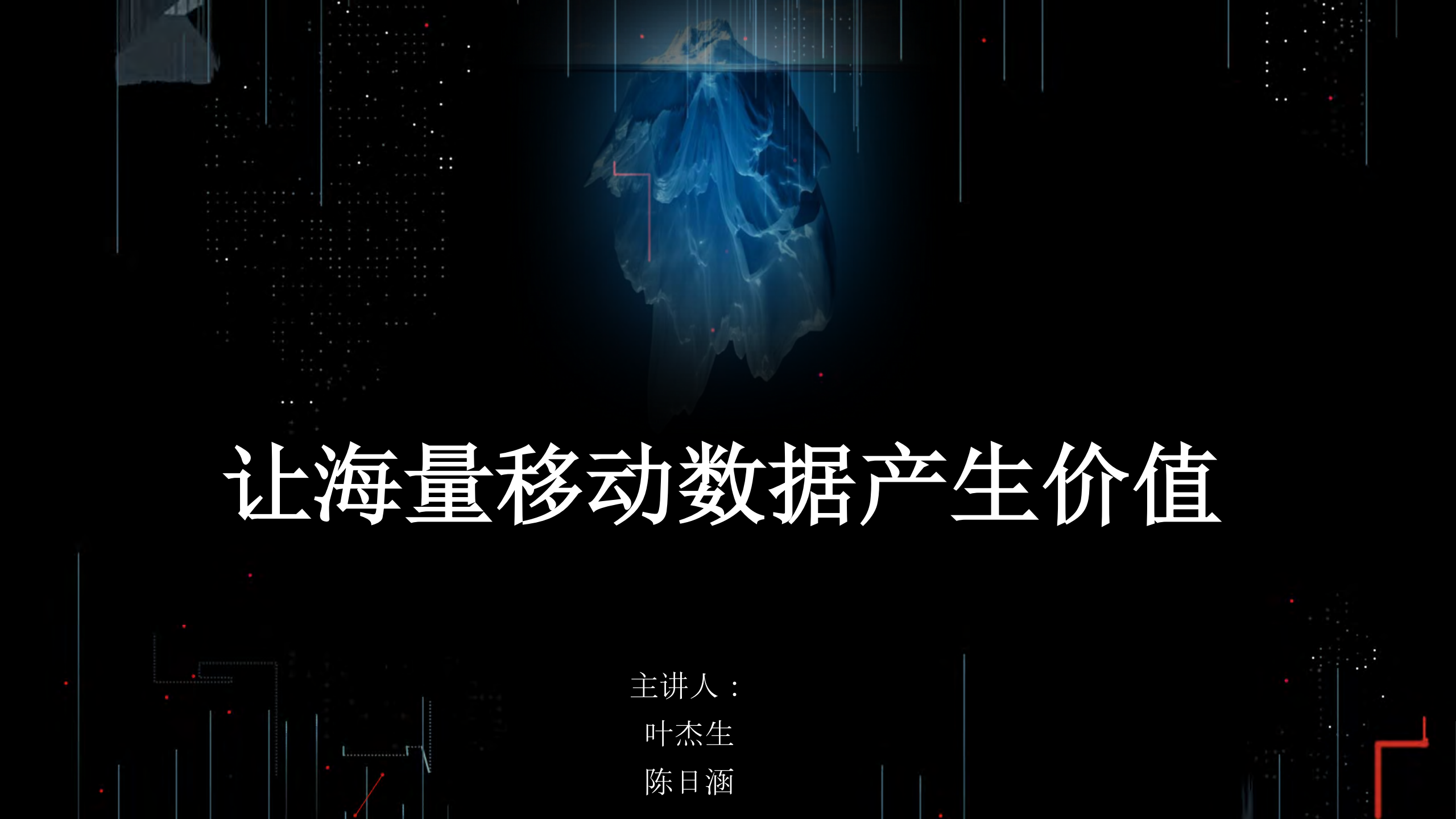


让海量移动数据产生价值

主讲人：
叶杰生
陈日涵

问题背景

基于海量的数据

如何处理与计算出
全面与有价值的结果？

数据背景

应用数据

行为数据

位置数据

事件数据

解决方案

数据处理

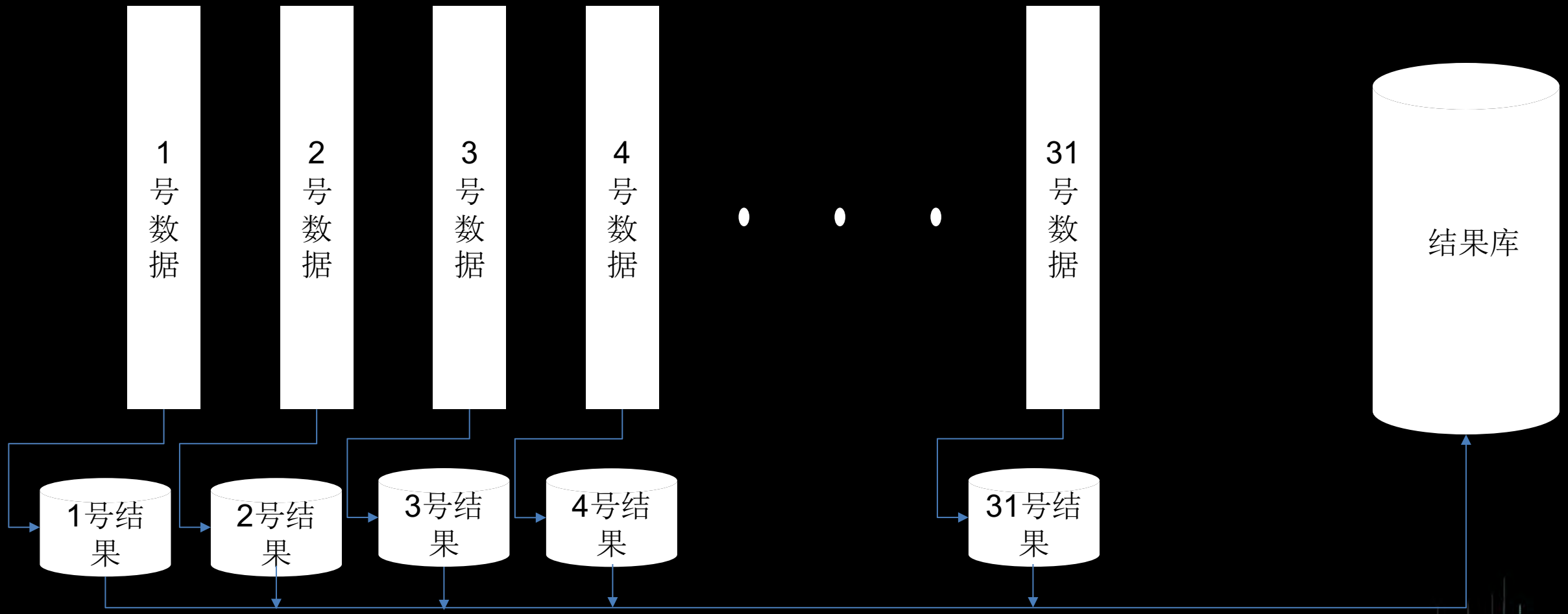
叶杰生

1.传统处理方案
2.传统方案的结果
3.改进处理方案
4.改进处理方案的结果

算法设计
陈日涵

1.维数灾难
2.降维措施
3. 设计
4. 结果

传统处理流程



实例

9月1号—9月3号

活跃

安装

实例

9月4号—9月11号

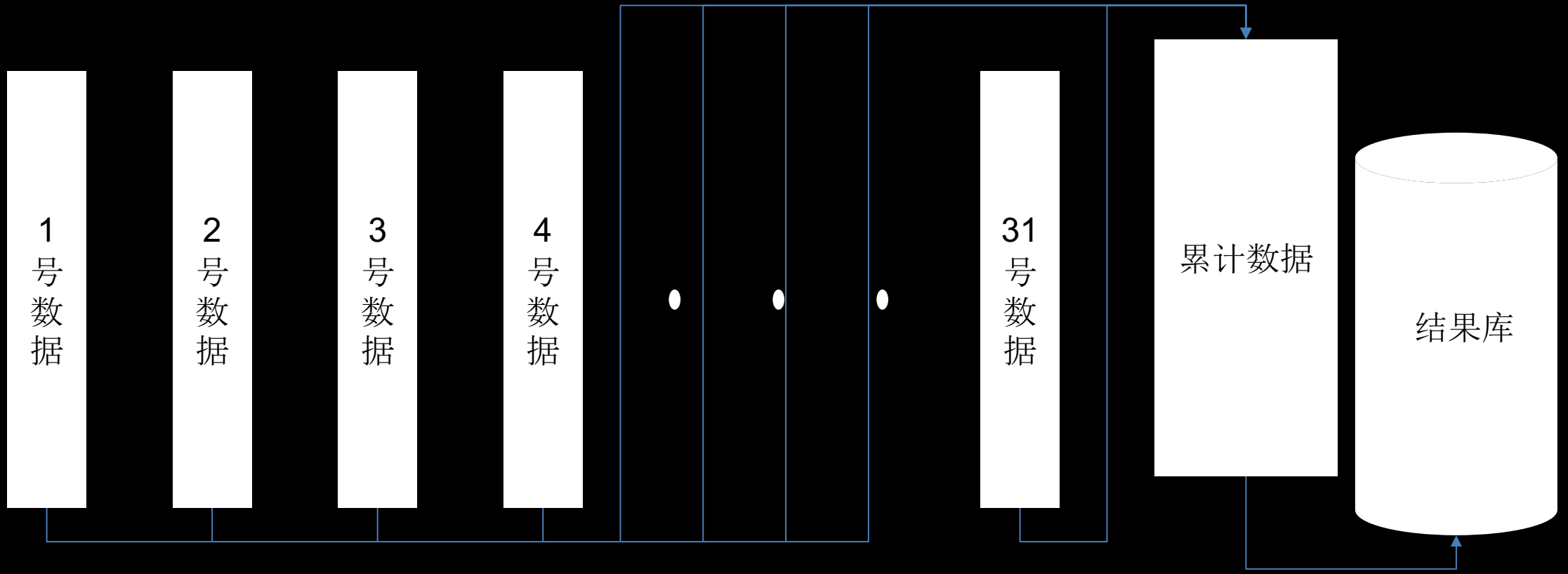
活跃

安装

结果

ID	App	Label	Point
4297	＊＊单词	应用兴趣/应用类别/教育/外语 02010204	0.4
4297	＊＊＊课堂	应用兴趣/应用类别/教育/课程管理 02010211	0.5
4297	王者＊＊	游戏偏好/游戏类型/角色扮演/即时 01011002	0.4

改进流程



9月4号－9月11号

9月1号－9月3号

结果

ID	App	Label	Point
4297	＊＊单词	应用兴趣/应用类别/教育/外语　02010204	0.2
4297	＊＊＊课堂	应用兴趣/应用类别/教育/课程管理　02010211	0.3
4297	王者＊＊	游戏偏好/游戏类型/角色扮演/即时　01011002	0.5

算法设计

陈日涵

目录

维数灾难

- 高维特征
- 特征信息
- 样本稀缺
- 模型限制

降维

- 降维目的
- 方法选择
- 方法及框架

未来趋势

- 深度学习
- 稀疏表达

维数灾难

- 目前TalkingData的数据中包含大量的Categorical Data

- One Hot Encoding的做法虽然直观但是会导致很多问题：

1、维数灾难
2、特征信息量很少，甚至有的可以当作是噪声。
3、模型限制

降维

目的

- 增强每维特征包含的信息
- 增加模型选择的灵活性
- 方便调参

选择

- 稀疏表达
- 支持大规模

方法

- 基于Metropolis Hashing的WarpLda方法
- Parameter Server框架

方法探讨

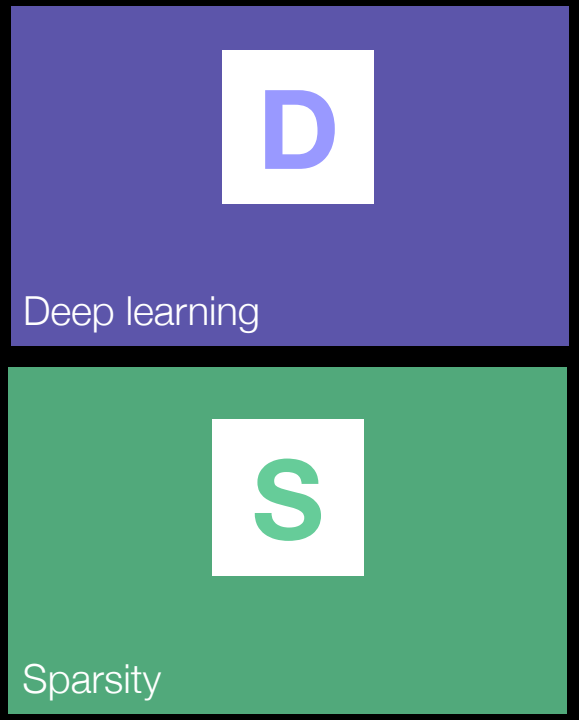
- LDA（Latent Dirichlet Allocation）是一种基于贝叶斯框架的生成模型，其目的是学习出隐含在文本中的主题。
- 将One-Hot-Encoding的特征当作Bag-of-Words，训练出主题，并用主题去代表一个样本的特征，达到降维的目的。
- WarpLda是一种基于Metropolis-Hashings的LDA方法，其分步的采样方法，使得WarpLda不仅从理论上降低了采样的复杂度，从工程实现上也降低了在优化LDA时random access的频率。
- Parameter Server的框架有着较细的通信粒度，且异步更新的方式可以很大程度上提高机器学习算法训练的效率。

结果

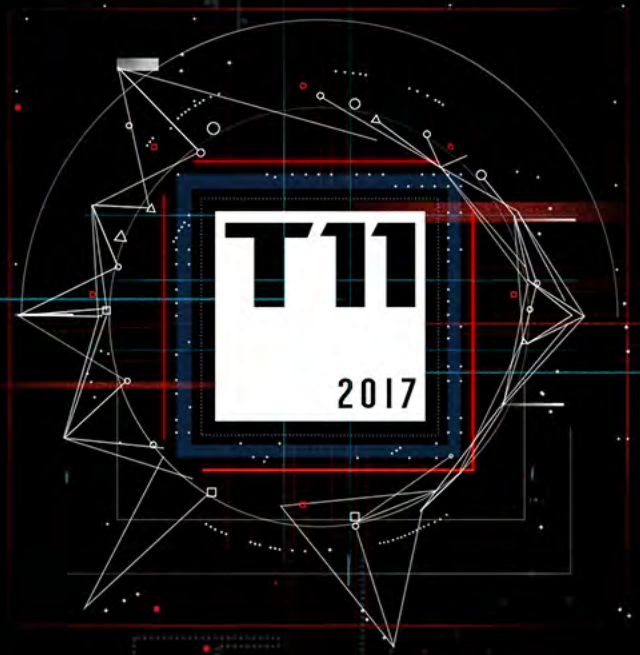
- 利用App数据建立模型
- 300万App、1000个主题、30亿参数

Topic	Apps
312	＊＊视频:0.1257256242895156 ＊＊视频:0.1121853733749541 ＊＊视频:0.1050368852539078 优＊:0.1031043142578524 芒果＊＊:0.090655374570 ...
72	＊＊手机银行:0.3137787934789783 ＊＊融e联:0.2836321098378883 中国＊＊银行:0.07132113645710109 ＊＊宝钱包:0.05142653696379523 ＊＊掌上银行:0.04880802485746054 ...

未来趋势



- Deep learning多基于频率学派的观点
- 基于梯度的优化效率高
- 主流模型稀疏性较差
- Word2vec: what’s next?
  Tomas Mikolov,Facebook



THANKS

TalkingData