

论文课 LESSON3

科研论文的数据的采集和分析 (一)

(刘佳妮, joeyliu1997@163.com)

论文课的计划

8-2: 科研论文写作的基本规范 (科研论文的结构)

8-6: 如何选择研究课题、写好文献综述?

8-9: 科研论文数据的采集和分析 (一) 

- 实证研究的含义及分类? 不同类型的研究选题与研究数据之间有什么关系?
- 优质数据集有哪些特征?
- 如果我们想要完成一篇实证科研论文, 我们可以通过哪些方式采集数据?

8-15: 科研论文数据的采集和分析 (二)

8-20: 如何写好学术论文的每个部分、论文的投稿、修改和发表

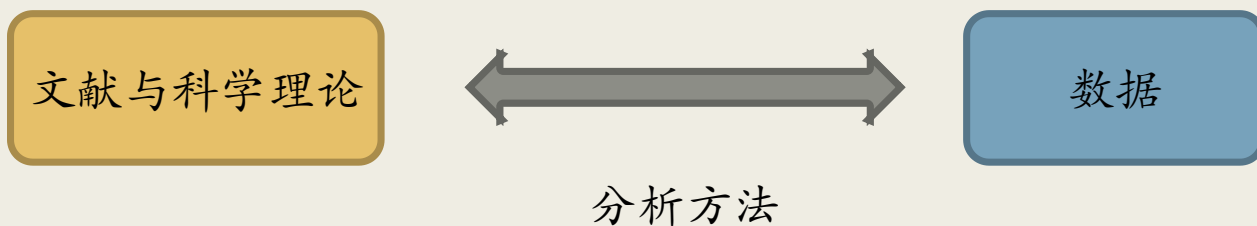
实证研究的含义及分类

- 实证研究（Empirical Research）的含义：是一种与规范研究方法（Normative Research）相对应的方法，它是基于观察和实验取得的大量事实、数据，利用统计推断的理论和**技术**，并经过严格的**经验检验**，而且引进数量模型，对社会现象进行数量分析的一种方法，其目的在于揭示各种社会现象的本质联系。相比规范研究方法，实证研究方法主要进行定量分析，依据数据说话，使其对社会问题的研究更精确、更科学。
- “揭示各种社会现象的本质联系”：
 - 要从“现象”或者“话题”出发，抽象后得到研究问题：
 - 现象或话题：音综中出现了越来越多的老歌手与新歌手的组合
 - 研究问题：对于新歌手来说，刚出道的时候与老歌手的合作，是否有利于他们之后**职业生涯取得成功**？
 - 提炼出研究问题，着眼于变量及变量之间的关系
 - 解释变量：X、被解释变量：Y、X与Y之间的关系
 - 例如：研究某一类广告会提升/降低产品销量；某一类信息会增加/抑制人们的资助意愿；某一类信息会唤起/消解人们的风险感知等等

实证研究的含义及分类

■ 实证研究的三要素：

- 科学理论：形成假设、提出问题、设计研究变量、构建模型、分析和解读结果
- 数据：数据越完整、越准确、质量越高，研究的可靠性与价值越高（本节课）
- 分析方法：计量经济学方法融合了统计推断、参数估计等统计学知识，可以对变量之间的统计关系展开度量（下节课）



实证研究的含义及分类

实证研究可以分成哪些类别？从数据获取方式出发：

- 基于*实验数据*的实证研究（实验数据：在实验中控制实验对象而收集到的数据）：
 - *实验室实验*：被试被随机地分入不同的组，控制两组的其他特征，仅保留他们在解释变量上水平的差异
 - *田野实验*：保留随机分组的特征，但将实验室的条件搬到更广阔的场景中
- 基于*观测数据*的实证研究（观测数据：通过调查或观察、而没有对事物进行人为控制、没有采用随机分组的条件下得到的数据）：
 - *以回归分析为主的计量模型*：利用计量模型，刻画变量间的统计关系
 - *自然实验*：借助于观测数据，巧妙地历史/地理/制度条件等制造“近似随机”的研究设计，常用于研究难以进行实验的问题（如：教育对于收入的影响）

实证研究的含义及分类

实证研究可以分成哪些类别？从数据获取方式出发：

注：

- 同一个研究问题可以使用不同的研究方法，多方法之间彼此互补、共同为研究问题提供论证，日益成为营销顶级期刊研究的主。
 - 例子：电影评论中的剧透内容会如何影响电影票房？

实证研究的含义及分类



Empirical Analysis

Model of Box Office Revenue

Let i denote movies and t denote the days after release. The dependent variable is $\ln(\text{DAILYREV})_{it}$, which represents the log-transformed daily box office revenue for movie i on day t . To examine the relationship between spoiler reviews and box office revenue, we considered the following model specification:

$$\begin{aligned} \ln(\text{DAILYREV})_{it} = & \beta_1 \ln(\text{DAILYREV})_{i,t-1} \\ & + \beta_2 \ln(\text{INTENSITY})_{i,t-1} + \beta_3 \text{PROP}_{i,t-1} \\ & + \beta_4 \ln(\text{CUMRATING})_{i,t-1} + \beta_5 \ln(\text{CUMVOL})_{i,t-1} \\ & + \beta_6 \ln(\text{ADVERT})_{i,t-1} + \beta_7 \ln(\text{THEATERS})_{it} \\ & + \beta_8 t + \beta_9 \text{HOLIDAY}_{it} \\ & + \sum_{d=1}^6 \gamma_j I\{\text{DAYOFWEEK}_{it} = d\} + \omega_i + \epsilon_{it} \end{aligned}$$

(6)

不确定性高，
剧透程度高

不确定性高，
剧透程度低

不确定性低，
剧透程度高

不确定性低，
剧透程度低

营销期刊 (JM) 2020, 评论中的剧透信息可以消减消费者的风险感知, 进而增加电影票房

实证研究的含义及分类

实证研究可以分成哪些类别？从数据获取方式出发：

注：

- 同一个研究问题可以使用不同的研究方法，多方法之间彼此互补、共同为研究问题提供论证，日益成为营销顶级期刊研究的主。
 - 例子：电影评论中的剧透内容会如何影响电影票房？
- 本课程主要侧重的是基于观测数据开展回归分析的实证研究方法
- 互联网及大数据为我们探索各类选题、获取各种数据提供了便利的途径，但是要注意判断研究问题的可行性！
 - 网络开放数据（本课程的主要内容）
 - 数据商提供的数据：如，Nielsen Scan
 - 平台内部数据：难以通过开放渠道获取、依赖于与平台和项目方的数据合作
 - 如，支付宝消费券的投放

实证研究的含义及分类

informs

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE

Vol. 67, No. 12, December 2021, pp. 7291–7307
ISSN 0025-1909 (print), ISSN 1526-5801 (online)

通过低成本刺激消费：来自COVID-19疫情期间的大规模政策实验

Stimulating Consumption at Low Budget: Evidence from a Large-Scale Policy Experiment Amid the COVID-19 Pandemic

Qiao Liu,^a Qiaowei Shen,^a Zhenghua Li,^b Shu Chen^b

^aGuanghua School of Management, Peking University, 100871 Beijing, China; ^bResearch Institute, Ant Group, 310013 Hangzhou, China

Contact: qiao_liu@gsm.pku.edu.cn, <https://orcid.org/0000-0003-0007-7190> (QL); qshen@gsm.pku.edu.cn, <https://orcid.org/0000-0003-3269-4003> (QS); sunny.lzh@antgroup.com (ZL); emily.cs@antgroup.com (SC)

Received: November 4, 2020

Revised: February 11, 2021

Accepted: March 12, 2021

Published Online in Articles in Advance:

Abstract. We use a novel panel with detailed transaction records of more than one million de-identified individuals to study the effect of a large-scale Chinese government-issued digital coupon program on consumer spending. Exploiting a difference-in-differences approach, we find that the digital coupon is highly effective in stimulating consumption. An effective government subsidy of RMB 1 can drive excess spending of RMB 3.4 to RMB 5.8, and the effect persists across multiple coupon issuance waves. In explaining the results, we find that a behavioral model with mental accounting and loss aversion can match the empirical evidence from the field. Our analysis, by illustrating the importance of embedding behavioral factors into the design and implementation of public policy, informs the current debate about cost-effective policy tools to recover the economy.

3.2. Data

支付宝中一百万用户的转账记录（时间、金额、对象、是否使用消费券、消费内容等）、用户个人特征

We use individual-level data from Alipay for the analysis. The basic function of Alipay is an e-wallet, which allows users to transfer money and make payments for both online and offline transactions. Users can link their major bank accounts to the service. In addition, Alipay offers other financial services, including virtual credit card services, *Huabei*, and financial management tools in the app. We have access to the de-identified account level transaction details for a total number of 1 million individuals sampled in this study.⁵

For each individual in the sample, we observe the complete transaction information of the account and the account holder's personal attributes, such as gender and age. The transaction-level data include the transaction time, transaction amount, the usage of coupon (if any), merchant category, and whether the transaction is online or offline. The advantage of this

数字消费券可以显著刺激消费，
1元消费券带来3.4~5.8元的消费



管理科学 (MASC) 2021, 数字消费券是否真的可以带动消费? 与支付宝平台的合作

数据与变量

■ 统计数据、统计变量与理论变量之间的关系：

- 统计数据：是对现象进行测量的结果

■ 总体与样本之间的关系：是否有代表性？

- 总体的不可得性使得我们需要借助样本来进行研究
- 例如，为了研究电影评论中的剧透程度对于电影票房的影响，作者选择了“993部发行于2013~2017年间的电影的特征、评级、影片类型、上线每日的票房、及其在IMDb网站上的评论”作为样本（代表性？）

■ 统计数据类型及举例：

- 分类数据：是否全球发行（是或否）、影片类型（科幻、悬疑等）
- 定序数据（顺序型数据）：电影评级（PG、PG-13、R）
- 定量数据（数值型数据）：制作成本（百万）、上线每日的票房
- 非结构化数据：评论文本、电影剧情简介、社交媒体平台上的视频等

Q：请大家结合之前科研课的内容以及此处数据类型的分类，思考：针对非结构化数据，我们应该如何处理？处理成为结构化数据之后它会变成分类、定序、定量数据中的哪一种？

数据与变量

■ 统计数据、统计变量与理论变量之间的关系：

- 统计变量：是对统计数据集其中信息的组织和概括

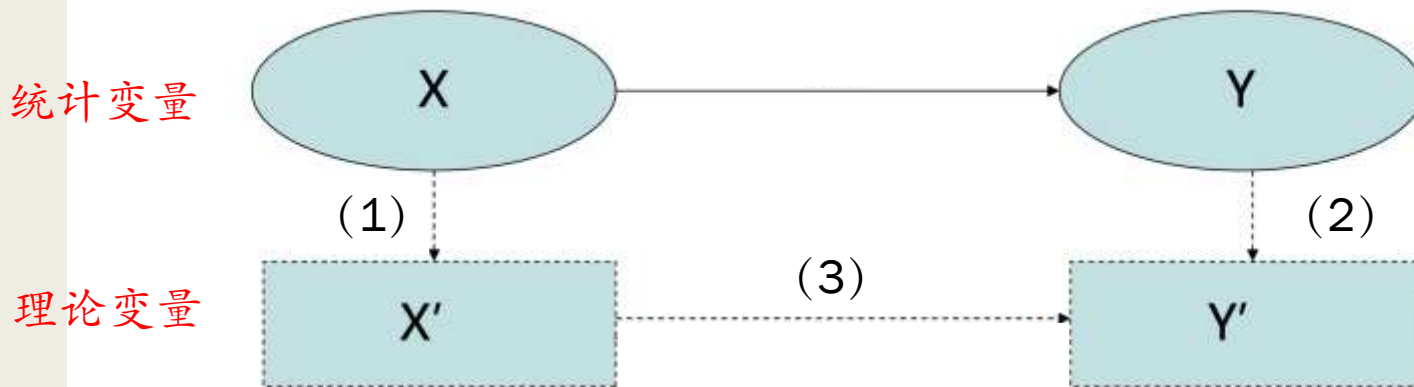
■ 常见的组织概括方式：

- 取值本身：分类变量、顺序变量、数值变量，包括离散型（电影上榜的天数；只能取有限个整数）与连续型（票房的金额；可以在整数区间中任意取值）
- 将某一类型的数据转化为另一类型的变量：“气温”与“气温是否超过40度”
- 根据数据构造统计量，来描述样本数据的概括性趋势与特征：
 - 集中趋势：平均数、众数、中位数
 - 离散趋势：方差（偏离中心的程度）、极差（最大值 - 最小值）

数据与变量

■ 统计数据、统计变量与理论变量之间的关系：

- 统计变量与理论变量之间的关系



• 4.实验效度

➤ 构念效度construct validity 关系 (1) (2)

测量的准确性，即变量测量的内容和构念的含义是否一致（首要指标）

➤ 统计结论效度statistical conclusion validity 关系 (3)

以统计检验对假设关系进行解释的可信度

数据分析的一般流程

数据获取

获取网络开放数据资源

数据处理

1. 数据预处理
2. 变量构造

分析与汇报

1. 建模与分析
2. 使用图表等方式展示结果

本节课

下节课

获取数据：优质数据集有哪些特征

小调查：

Q：请大家回忆一下自己之前阅读的论文，觉得如果想要利用网络开放的观测数据展开学术研究，大家想要一个满足何种标准的数据集？或者说，你觉得优质的开放来源的数据集应该具有哪些特征？

优质数据集有哪些特征

1. 可获得 Accessible
2. 有丰富的信息量 Informative:
 - 观察的维度多、样本量大
2. 与个体有关 Individual-level: 可以用于揭示个体的行为与决策

通过低成本来刺激消费：来自COVID-19疫情期间的大规模政策实验

Stimulating Consumption at Low Budget: Evidence from a Large-Scale Policy Experiment Amid the COVID-19 Pandemic

Qiao Liu,^a Qiaowei Shen,^a Zhenghua Li,^b Shu Chen^b

数字消费券可以显著刺激消费，
1元消费券带来3.4~5.8元的消费

^aGuanghua School of Management, Peking University, 100871 Beijing, China; ^bResearch Institute, Ant Group, 310013 Hangzhou, China

Contact: qiao_liu@gsm.pku.edu.cn, <https://orcid.org/0000-0003-0007-7190> (QL); qshen@gsm.pku.edu.cn, <https://orcid.org/0000-0003-3269-4003> (QS); sunny.lzh@antgroup.com (ZL); emily.cs@antgroup.com (SC)

Received: November 4, 2020

Revised: February 11, 2021

Accepted: March 12, 2021

Published Online in Articles in Advance:
October 7, 2021

Abstract. We use a novel panel with detailed transaction records of more than one million de-identified individuals to study the effect of a large-scale Chinese government-issued digital coupon program on consumer spending. Exploiting a difference-in-differences approach, we find that the digital coupon is highly effective in stimulating consumption. An effective government subsidy of RMB 1 can drive excess spending of RMB 3.4 to RMB 5.8, and the effect persists across multiple coupon issuance waves. In explaining the results, we find that a behavioral model with mental accounting and loss aversion can match the empirical evidence from the field. Our analysis, by illustrating the importance of embedding behavioral factors into the design and implementation of public policy, informs the current debate about cost-effective policy tools to recover the economy.

3.2. Data

支付宝中一百万用户的转账记录（时间、金额、对象、是否使用消费券、消费内容等）、用户个人特征

We use individual-level data from Alipay for the analysis. The basic function of Alipay is an e-wallet, which allows users to transfer money and make payments for both online and offline transactions. Users can link their major bank accounts to the service. In addition, Alipay offers other financial services, including virtual credit card services, *Huabei*, and financial management tools in the app. We have access to the de-identified account level transaction details for a total number of 1 million individuals sampled in this study.⁵

For each individual in the sample, we observe the complete transaction information of the account and the account holder's personal attributes, such as gender and age. The transaction-level data include the transaction time, transaction amount, the usage of coupon (if any), merchant category, and whether the transaction is online or offline. The advantage of this



管理科学 (MASC) 2021, 数字消费券是否真的可以带动消费? 与支付宝平台的合作

优质数据集有哪些特征

1. 可获得 **Accessible**
2. 有丰富的信息量 **Informative**:
 - 观察的维度多、样本量大
2. 与个体有关 **Individual-level**: 可以用于揭示个体的行为与决策
3. 包含你的听众感兴趣的指标 **Constructive**:
 - 听众可以是企业、非营利性组织、项目发起人、广告投放者等等
 - 让你的研究有实践意义

Stray

游戏页面

全部 讨论 截图 艺术作品 实况直播 视频 新闻 指南 评测

12,483 人游玩中 / 423 人在组队游玩 | 查看统计



一只孤身离群的流浪猫迷失在了了一座被遗忘的网络城市里。它必须解开一个古老的谜题才能逃离出去。找到回家的路。


HK\$ 139.00

访问商店页面

思考:

1. 请问大家，如果现在要基于Steam评论数据展开研究，你认为这一研究的“听众”是谁？
2. 在Steam评论页面展现的这些数据中，最有价值的是哪些？你的听众感兴趣的指标有哪些？

有 382 人觉得这篇评测有价值
有 65 人觉得这篇评测没价值  3


 推荐
总时数 1.2 小时

发布于: 7 月 21 日
免费获取的产品

艾尔登之王退休来当猫猫了

 垃圾站
帐户内拥有 39 项产品

有 292 人觉得这篇评测有价值
有 120 人觉得这篇评测没价值  4


 推荐
总时数 0.7 小时

发布于: 7 月 20 日

圣火昭昭，圣光耀耀，凡我弟子，喵喵喵喵。
怜我猫猫，飘零无助，悲凉万物，喵喵喵喵。
猫猫慈父，知义知情，自救理心，苏我明性。
怜我世间，红尘尘染，除恶扬善，喵喵喵喵。
生亦何欢，死亦何苦，熊熊圣火，焚我残躯。
十二寒宝，碧台清明，妙音引路，无量净土。

 yukisakura
帐户内拥有 3,522 项产品

有 316 人觉得这篇评测有价值
有 8 人觉得这篇评测没价值  16

 推荐
总时数 13.0 小时

发布于: 7 月 21 日

希望这篇评测对你购买这款游戏，或者推荐朋友购买时，有所帮助。

注:
制作组最近更新已经修复了部分剧情闪退bug，下次更新会提升GPU优化
本评测结尾处带有整个游戏的故事梗概，非剧透党慎入
帧数选项往左选才可以超过60帧，该游戏其实不错60FPS!
对克鲁苏相关内容敏感的玩家不要购买!

前言

Annapurna在Stray获得的战绩，注定是一个必然。
一个上来就卖给你《艾迪芬奇的记忆》的公司，世界上能有几家？
他还说没事，又要给你Journey, Outer Wilds, Neon White, Donut Count, The Pathless, GOROGOIA, Florence等。天啊兄弟，你生错年了吧？你确定今天的电子游戏行业值得你这么卷吗？

当然，今天的主角，Stray (迷失)，一款从2015年就开始开发的游戏，也加入了这奇妙的行列。

(12分钟? 谁记得那个游戏?)

Stray和之前Annapurna又是截然不同的体验，而且在部分方面更具野心。赛博朋克的世界观? 猫猫为主角!!! 中小作坊最高端的价格能给你带来优秀的手柄适配和还算相当不差的画面? 画面或许没有索尼圣火王的一流水准，但是其它方面确在一流。更别提那轻快的音乐，更加锦上添花。

 垃圾
帐户内拥有 1,382 项产品



Articles

About 485 results (0.06 sec)

- Any time
- Since 2022
- Since 2021
- Since 2018
- Custom range...

- Sort by relevance
- Sort by date

- Any type
- Review articles

- include patents
- include citations

The impact of social influence on the perceived **helpfulness** of online consumer reviews [PDF] rug.nl

[H Risselada](#), [L de Vries](#), [M Verstappen](#) - ... **Journal of Marketing**, 2018 - emerald.com

... This study aims to study to what extent the **helpfulness** votes others attach to a **review** affect ... votes about the **review's helpfulness** have on the perceived **helpfulness** of the **review**. That is...

☆ Save Cite Cited by 41 Related articles All 5 versions

European Journal of Marketing

Why should I believe this? Deciphering the qualities of a credible online customer review [PDF] tandfonline.com

[CJ Clare](#), [G Wright](#), [P Sandiford](#)... - **Journal of Marketing** ..., 2018 - Taylor & Francis

... that are important in the context of this study, '**review helpfulness**' and '**review credibility**'. ... between the constructs of **review consultation**, **review helpfulness** and **review credibility**. ...

☆ Save Cite Cited by 33 Related articles All 9 versions

Journal of Marketing Management

Third-party product **review** and firm marketing strategy [PDF] informs.org

[Y Chen](#), [J Xie](#) - **Marketing science**, 2005 - pubsonline.informs.org

... /market/**review**/media ... **review**-endorsed advertising format (ie, advertisements containing third-party award logos) to broadcast its victory can hurt the winning product of a product **review**...

☆ Save Cite Cited by 494 Related articles All 19 versions

Marketing Science

Exploring effects of source similarity, message valence, and receiver regulatory focus on yelp **review** persuasiveness and purchase intentions [PDF] tandfonline.com

[I Pentina](#), [AA Bailey](#), [L Zhang](#) - **Journal of Marketing** ..., 2018 - Taylor & Francis

... of **helpfulness**, trustworthiness, and credibility of the **review**. The study also tests moderating effects of perceived source similarity and user regulatory focus on the relationships among...

☆ Save Cite Cited by 96 Related articles All 2 versions

Journal of Marketing Management

以“评论有用性”作为被解释变量的研究的管理启示较弱，很难发表于近年的顶级期刊（Journal of Marketing Research, Journal of Marketing, Marketing Science）

优质数据集有哪些特征

1. 可获得 Accessible
2. 有丰富的信息量 Informative:
 - 观察的维度多、样本量大
3. 与个体有关 Individual-level: 可以用于揭示个体的行为与决策
4. 包含你的听众感兴趣的指标 Constructive:
 - 听众可以是企业、非营利性组织、项目发起人、广告投放者等等
 - 让你的研究有实践意义
5. 独特 Unique: 独特的数据集往往来源于全新的商业实践，能帮助我们:
 - 关注全新领域的研究问题

优质数据集有哪些特征

PNAS PNAS PNAS



Structured, uncertainty-driven exploration in real-world consumer choice

Eric Schulz^{a,1,2}, Rahul Bhui^{a,1}, Bradley C. Love^{b,c}, Bastien Brier^d, Michael T. Todd^d, and Samuel J. Gershman^a

^aDepartment of Psychology, Harvard University, Cambridge, MA 02138; ^bDepartment of Experimental Psychology, University College London, London WC1H 0AP, United Kingdom; ^cThe Alan Turing Institute, London NW1 2DB, United Kingdom; and ^dData Science Team, Deliveroo, London EC4R 3TE, United Kingdom

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 23, 2019 (received for review December 10, 2018)

Making good decisions requires people to appropriately explore their available options and generalize what they have learned. While computational models can explain exploratory behavior in constrained laboratory tasks, it is unclear to what extent these models generalize to real-world choice problems. **We investigate the factors guiding exploratory behavior in a dataset consisting of 195,333 customers placing 1,613,967 orders from a large online food delivery service.** We find important hallmarks of adaptive exploration and generalization, which we analyze using computational models. In particular, customers seem to engage in uncertainty-directed exploration and use feature-based generalization to guide their exploration. Our results provide evidence that people use sophisticated strategies to explore complex, real-world environments.

exploration | generalization | reinforcement learning | decision making

it is unclear whether these theories can successfully predict real-world choices.

Our results suggest that customers explore (i.e., order from unexperienced restaurants) adaptively based on signals of restaurant quality and make better choices over time. Exploration is indeed risky and leads to worse outcomes on average, but people are more likely to explore in cities where this downside is lower due to higher mean restaurant quality. Moreover, we show that customers' exploratory behavior might take into account not only the prospective reward from choosing a restaurant, but also the degree of uncertainty in their reward estimates. Consistent with an optimistic uncertainty-directed exploration policy, they preferentially sample lesser-known options and are more likely to reorder from restaurants with higher uncertainties.

Importantly, we apply cognitive and statistical modeling to customers' choice behavior and find that their choices are best

美国科学院院刊 PNAS, 2019

优质数据集有哪些特征

1. 可获得 **Accessible**
2. 有丰富的信息量 **Informative**:
 - 观察的维度多、样本量大
2. 与个体有关 **Individual-level**: 可以用于揭示个体的行为与决策
3. 包含你的听众感兴趣的指标 **Constructive**:
 - 听众可以是企业、非营利性组织、项目发起人、广告投放者等等
 - 让你的研究有实践意义
4. 独特 **Unique**: 独特的数据集往往来源于全新的商业实践，能帮助我们:
 - 关注全新领域的研究问题
 - 提出**全新的变量**，研究以往无法回答或者难以研究的问题（例：音乐风格、偏见）

优质数据集有哪些特征

What Makes Popular Culture Popular? Product Features and Optimal Differentiation in Music

Noah Askin^a and Michael Mauskapf^b

美国社会学评论 ASR, 2017

如何利用音频分析技术 (Musical Information Retrieval) 量化音乐作品旋律的风格? 并讨论当前歌曲的风格与同时代其他作品风格的相似性, 及其对于歌曲受欢迎程度的影响

Table 1. The Echo Nest Sonic Features

Attribute	Scale	Definition
Acousticness	0–1	Represents the likelihood that the song was recorded solely by acoustic means (as opposed to more electronic/electric means).
Danceability	0–1	Describes how suitable a track is for dancing. This measure includes tempo, regularity of beat, and beat strength.
Energy	0–1	A perceptual measure of intensity throughout the track. Think fast, loud, and noisy (i.e., hard rock) more than dance tracks.
Instrumentalness	0–1	The likelihood that a track is predominantly instrumental. Not necessarily the inverse of speechiness.
Key	0–11 (integers only)	The estimated, overall key of the track, from C through B. We enter key as a series of dummy variables.
Liveness	0–1	Detects the presence of a live audience during the recording. Heavily studio-produced tracks score low on this measure.
Mode	0 or 1	Whether the song is in a minor (0) or major (1) key.
Speechiness	0–1	Detects the presence of spoken word throughout the track. Sung vocals are not considered spoken word.
Tempo	Beats per minute (BPM)	The overall average tempo of a track.
Time Signature	Beats per bar/measure	Estimated, overall time signature of the track. 4/4 is the most common time signature by far and is entered as a dummy variable in our analyses.
Valence	0–1	The musical positiveness of the track.

Note: This list of features includes all but one of the attributes provided by The Echo Nest's suite of algorithms: loudness. We cut this variable from our final analysis at the suggestion of the company's senior engineer, who explained that loudness is primarily determined by the mastering technology used to make a particular recording, a characteristic that is confounded through radio play and other forms of distribution.

优质数据集有哪些特征

Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg^{a,1}, Londa Schiebinger^b, Dan Jurafsky^{c,d}, and James Zou^{a,e,1}

^aDepartment of Electrical Engineering, Stanford University, Stanford, CA 94305; ^bDepartment of History, Stanford University, Stanford, CA 94305; ^cDepartment of Linguistics, Stanford University, Stanford, CA 94305; ^dDepartment of Computer Science, Stanford University, Stanford, CA 94305; ^eDepartment of Biomedical Data Science, Stanford University, Stanford, CA 94305; and ^fChan Zuckerberg Biohub, San Francisco, CA 94158

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

Word embeddings are a powerful machine-learning framework that represents each English word by a vector. The geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words. In this paper, we develop a framework to demonstrate how the temporal dynamics of the embedding helps to quantify changes in stereotypes and attitudes toward women and ethnic minorities in the 20th and 21st centuries in the United States. We integrate word embeddings trained on 100 y of text data with the US Census to show that changes in the embedding track closely with demographic and occupation shifts over time. The embedding captures societal shifts—e.g., the women's movement in the 1960s and Asian immigration into the United States—and also illuminates how specific adjectives and occupations became more closely associated with certain populations over time. Our framework for temporal analysis of word embedding opens up a fruitful intersection between machine learning and quantitative social science.

word embedding | gender stereotypes | ethnic stereotypes

in the large corpora of training texts (20–23). For example, the vector for the adjective honorable would be close to the vector for man, whereas the vector for submissive would be closer to woman. These stereotypes are automatically learned by the embedding algorithm and could be problematic if the embedding is then used for sensitive applications such as search rankings, product recommendations, or translations. An important direction of research is to develop algorithms to debias the word embeddings (20).

In this paper, we take another approach. We use the word embeddings as a quantitative lens through which to study historical trends—specifically trends in the gender and ethnic stereotypes in the 20th and 21st centuries in the United States. We develop a systematic framework and metrics to analyze word embeddings trained over 100 y of text corpora. We show that temporal dynamics of the word embedding capture changes in gender and ethnic stereotypes over time. In particular, we quantify how specific biases decrease over time while other stereotypes increase. Moreover, dynamics of the embedding strongly correlate with quantifiable changes in US society, such as demographic and occupation shifts. For example, major transitions in

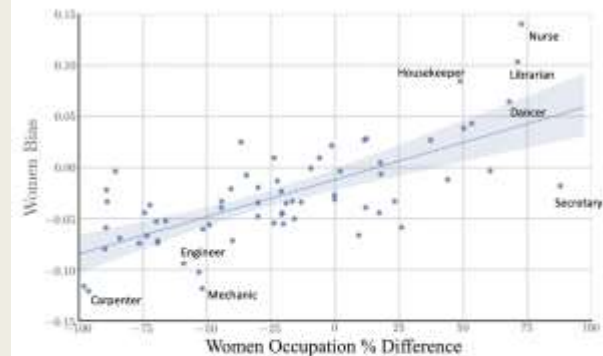
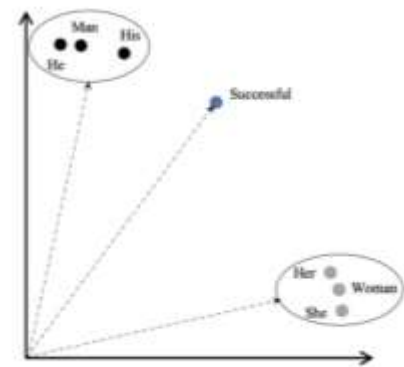


Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.



美国科学院院刊 PNAS, 2019

利用词嵌入技术，通过分析Google Books文本，量化美国过去100年间关于性别和种族的偏见

优质数据集有哪些特征

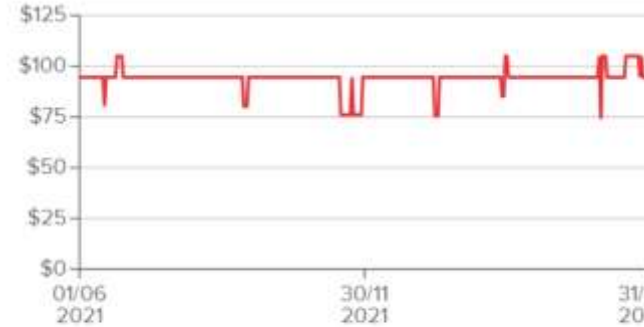
1. 可获得 **Accessible**
2. 有丰富的信息量 **Informative**:
 - 观察的维度多、样本量大
2. 与个体有关 **Individual-level**: 可以用于揭示个体的行为与决策
3. 包含你的听众感兴趣的指标 **Constructive**:
 - 听众可以是企业、非营利性组织、项目发起人、广告投放者等等
 - 让你的研究有实践意义
4. 独特 **Unique**: 独特的数据集往往来源于全新的商业实践，能帮助我们：
 - 关注全新领域的研究问题
 - 提出全新的变量，研究以往文献无法回答的问题
 - 创造独特的“**实验情景**”，弥补以往文献囿于数据而难以厘清的问题（例：历史价格透明与信息不对称）

优质数据集有哪些特征

Check for updates



Journal of Marketing
2022, Vol. 59(1)
© American Marketing Association
Article reuse
sagepub.com/journals-permissions
DOI: 10.1177/0022243721106111
journals.sagepub.com



历史价格变化的方向 × 频率 —— 决策延迟

发现:

1. 方向: 控制变动程度, 当价格上涨时消费者更愿意延迟购买;
2. 频率: 相比于多次小幅变动, 变动大的时候效应更强

解释机制: 消费者对于未来价格的预期

顯示:

前 7 天	前 30 天	前 90 天
前 180 天	前 365 天	

Article

The Impact of Historical Price Information on Purchase Deferral

Manissa P. Gunadi and Ioannis Evangelidis

Abstract

营销研究期刊 JMR, 2022

优质数据集有哪些特征



其他创设独特研究情景的数据例子：如，腾讯公益小程序：

- 1 捐赠者需要在多个项目中选择捐赠项目，项目之间可能会互相影响；
- 2 可以实时地浏览他人的捐款数据；
- 3 个体的捐赠金额与捐赠行为会被朋友知道；

依然要注意！
这些数据是否
可得的问题

优质数据集有哪些特征

1. 可获得 **Accessible**
2. 有丰富的信息量 **Informative**:
 - 观察的维度多、样本量大
2. 与个体有关 **Individual-level**: 可以用于揭示个体的行为与决策
3. 包含你的听众感兴趣的指标 **Constructive**:
 - 听众可以是企业、非营利性组织、项目发起人、广告投放者等等
 - 让你的研究有实践意义
4. 独特 **Unique**: 独特的数据集往往来源于全新的商业实践，能帮助我们：
 - 关注全新领域的研究问题
 - 提出全新的变量，研究以往文献无法回答的问题
 - 创造独特的“实验情景”，弥补以往文献囿于数据而难以厘清的问题（例：历史价格透明与信息不对称）

数据的采集

Q: 大家觉得对于完成一篇学术论文而言, 学会采集数据是不是必备的技能?

数据的采集

Q: 大家觉得对于完成一篇学术论文而言, 学会采集数据是不是必备的技能?

A: 并不是! 相比于亲自动手编写程序爬数据而言, 更重要的是知道:

应该爬取哪些数据? 数据集应该具备何种特征? 从数据中可以挖掘出哪些有价值的洞察? 如何将这些洞察翻译表达成学术问题、使用科学的方法分析这些问题并得出结论?

早在18世纪, 亚当斯密在其著作《国富论》中就已经观察到了“分工”对于“提高生产效率”的重要意义!

如何采集数据 —— 如何利用多方资源获取数据

如何利用多方资源获取数据

■ 传统渠道公开数据库：

- 国家统计局年鉴、IMF、世界银行、WHO等

■ 互联网公开数据库、数据源

- *Kaggle*: https://www.youtube.com/watch?v=TNzDMOg_zsw
- 亚马逊商品评论数据: <http://jmcauley.ucsd.edu/data/amazon/>
- Yelp 评论数据: <https://www.yelp.com/dataset>
- 斯坦福大规模网络数据集: <http://snap.stanford.edu/data/index.html>
- 哈佛大学大规模数据库: <https://dataverse.harvard.edu/>
- 其他数据渠道: <https://awesomeopensource.com/>

如何利用多方资源获取数据

- 如何自己爬数据？

- 利用API资源 (*Application Programming Interface*) :

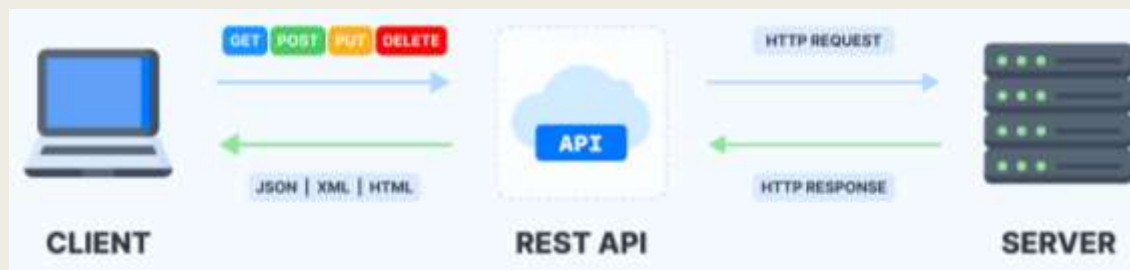
- API是什么？是一种计算接口，定义了多个软件之间的交互，以及可以进行的调用 (call) 或请求 (request) 的种类，如何进行调用或发出请求，应使用的数据格式，应遵循的惯例等。它还可以提供扩展机制，以便用户可以通过各种方式对现有功能进行不同程度的扩展。

如何利用多方资源获取数据

■ 如何自己爬数据？

- 利用API资源 (*Application Programming Interface*) :

■ API是什么？



- **简单来说**，API允许我们在获取权限后，按照一定的对话方式，通过编写非常简单的程序，在目标爬取软件列示出的、愿意分享的数据源中，告诉他们我们需要的数据，并且接收、存储。
 - 获取权限：*按照API documents中的要求获取权限*
 - 目标爬取软件愿意分享哪些数据？*API documents列出来的字段*
 - 一定的对话方式？*API documents规定的检索格式*
 - 告知、接收、存储？*利用Python等编程语言帮助自己迅速重复发送请求、接收、存储数据*

如何利用多方资源获取数据

■ 如何自己爬数据？

- 利用API资源 (*Application Programming Interface*) :

■ 利用API爬数据是一种行之有效的方式吗？

The screenshot shows a Google Scholar search for the query "(*API AND *empirical*) (source: *Marketing Science*)". The search results are displayed in a list format with the following entries:

- Frontiers: virus shook the streaming star: estimating the COVID-19 impact on music consumption**
J.Sim, D.Cho, Y.Hwang, R.Telang - *Marketing Science*, 2022 - pubsonline.informs.org
... To the best of our knowledge, this study provides the first **empirical** evidence of the impact of ... Using data collected from Spotify's **API**, we reveal that top-tier artists have not significantly ...
☆ Save 📄 Cite Cited by 32 Related articles All 5 versions [PDF] informs.org
- Regulatory spillovers and data governance: Evidence from the GDPR**
C.Pevzner, S.Bechtold, M.Baliga - *Marketing Science*, 2022 - pubsonline.informs.org
... Although it is difficult to measure data minimization in general, our **empirical** context lets us ... mean of about 75.8% and in CDN/**API** by about 0.4 percentage points from mean of about ...
☆ Save 📄 Cite Cited by 19 Related articles All 2 versions [PDF] informs.org
- Advertising strategy in the presence of reviews: An **empirical** analysis**
B.Hollenbeck, S.Moorthy, D.Proserpio - *Marketing Science*, 2019 - pubsonline.informs.org
... Then, using the SpyFu **API**, we obtained search advertising spending information for all the independent hotels for which SpyFu had this information. This procedure yielded monthly ...
☆ Save 📄 Cite Cited by 46 Related articles All 14 versions [PDF] informs.org
- Digital piracy, creative productivity, and customer care effort: Evidence from the digital publishing industry**
X.Li, C.Liao, Y.Xie - *Marketing Science*, 2021 - pubsonline.informs.org
... the quasi-experiment, we describe the **empirical** setting and summary statistics of our data ... application programming interface (**API**). Third, we used the trained sentiment **API** to tag the ...
☆ Save 📄 Cite Cited by 6 Related articles All 4 versions [PDF] informs.org
- Using Deep Learning to Overcome Privacy and Scalability Issues in Customer Data Transfer**
P.Arvand, C.Lee - *Marketing Science*, 2022 - pubsonline.informs.org
... **API** calls to the discriminator residing inside the firm's walls. In this situation, the researcher makes **API** ... Thus, our **empirical** evidence suggests that GANs can indeed incorporate multiple ...
☆ Save 📄 Cite [PDF] informs.org

如何利用多方资源获取数据

■ 如何自己爬数据？

- 利用API资源 (*Application Programming Interface*) :

■ API Documents是什么？

- 如何知道我想采集数据的网站有没有API？看看相关文章有没有利用API获取数据！如果他们可以用，那么你大概率也是可以用的。

■ 常用的API有哪些？

- *Twitter*: <https://developer.twitter.com/en/docs/twitter-api>

- *YouTube*: <https://developers.google.com/youtube/v3>

- *IMDb*: <https://developer.imdb.com/>

- *豆瓣*: <https://www.doubanapi.com/>

- *Spotify*: <https://developer.spotify.com/documentation/web-api/>

- *网易云音乐*: <https://github.com/Binaryify/NeteaseCloudMusicApi>

- *Steam*: <https://steamcommunity.com/dev>

如何利用多方资源获取数据

- 以Genius音乐资讯平台为例：

The screenshot displays the Genius website interface. At the top, there is a yellow search bar with the text "Search lyrics & more" and a magnifying glass icon. To the right of the search bar, the word "GENIUS" is prominently displayed. Further right, there are navigation links for "FORUMS", "FEED", "ME", "MESSAGES", "EARN IQ", and a user profile icon. Below this is a black navigation bar with white text for "FEATURED", "CHARTS", "VIDEOS", "PROMOTE", "FORUMS", and "ADD A SONG".

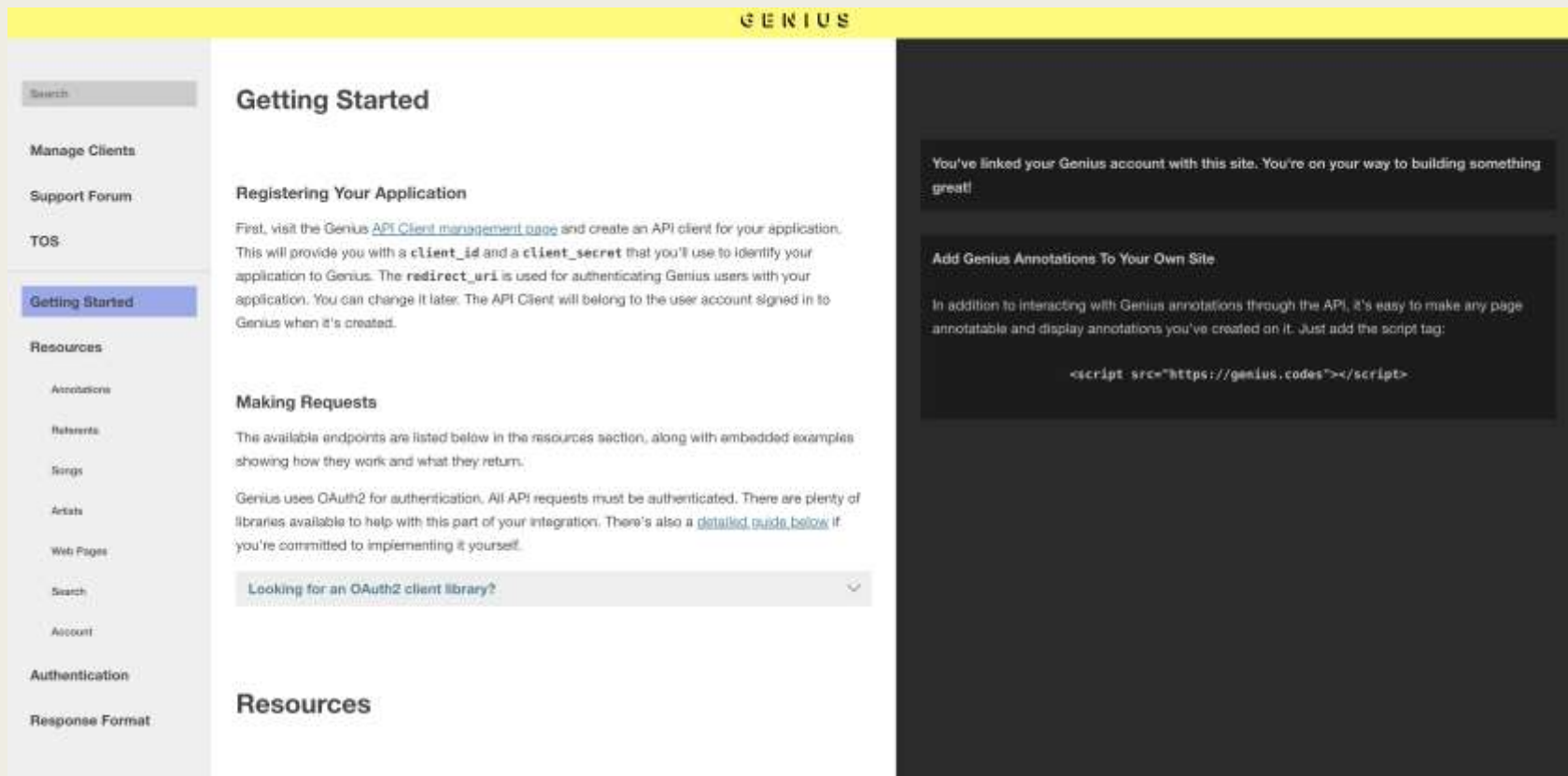
The main content area features a large news article on the left with the headline "NIKI Relives Teenage Love On New Song 'High School In Jakarta'". Below the headline, it states "It's the latest single off her forthcoming album, 'Nicole.'" and is attributed to "by Leah Degrazia / Aug 8, 2022". To the right of the article is a photograph of NIKI in a school hallway, surrounded by other students.

Below the main article are four smaller news snippets, each with a "NEWS" label and a small image:

- Drake Flips '70s Classic On New DJ Khaled And Lil Baby Collab "STAYING ALIVE"** (by Leah Degrazia / Aug 6, 2022)
- Read All The Lyrics To Calvin Harris' New Album 'Funk Wav Bounces, Vol. 2'** (by Leah Degrazia / Aug 5, 2022)
- Benny Blanco Recruits BTS & Snoop Dogg For New Song "Bad Decisions"** (by Leah Degrazia / Aug 5, 2022)
- Read All The Lyrics To YoungBoy Never Broke Again's New Album 'The Last Slimeto'** (by Leah Degrazia / Aug 5, 2022)

如何利用多方资源获取数据

- 以Genius音乐资讯平台为例：



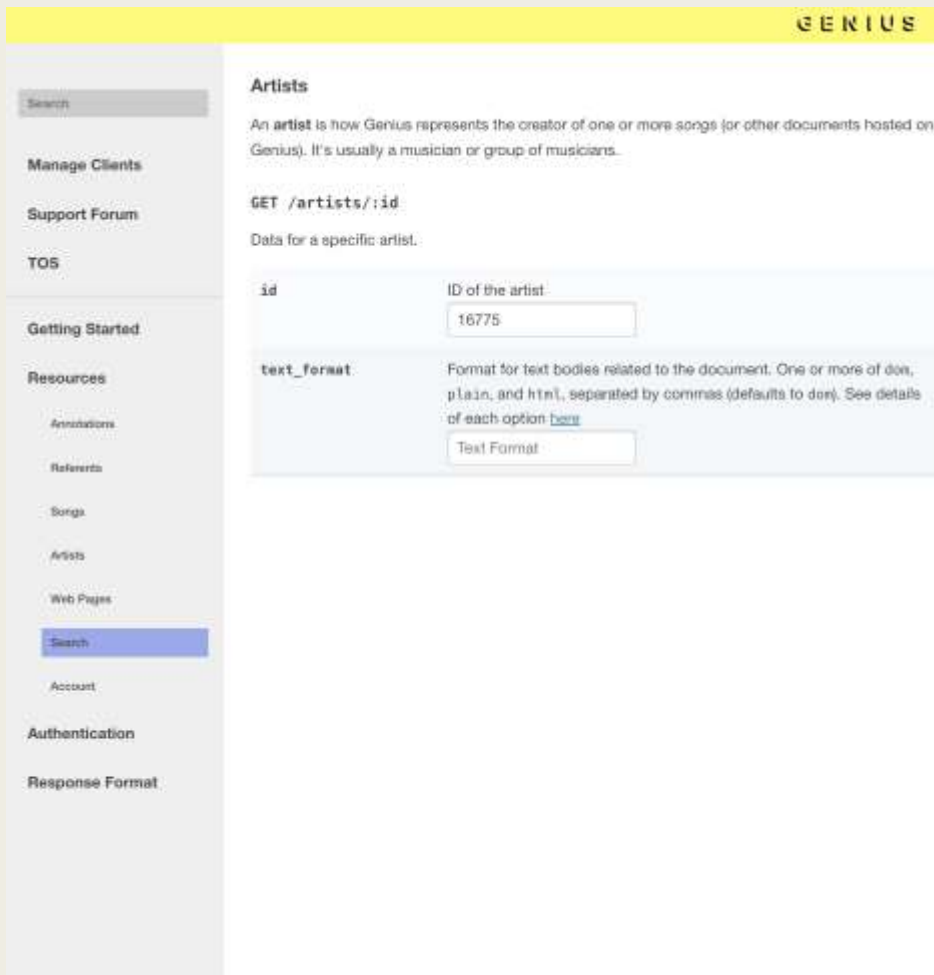
The screenshot displays the Genius API documentation page. The top navigation bar is yellow with the word "GENIUS" in white. On the left, there is a sidebar menu with a search bar and various navigation links: "Manage Clients", "Support Forum", "TOS", "Getting Started" (highlighted in blue), "Resources" (with sub-links for Annotations, Albums, Songs, Artists, Web Pages, Search, and Account), "Authentication", and "Response Format". The main content area is titled "Getting Started" and contains three sections: "Registering Your Application", "Making Requests", and "Resources". The "Registering Your Application" section explains the process of creating an API client, mentioning `client_id`, `client_secret`, and `redirect_uri`. The "Making Requests" section discusses authentication using OAuth2 and provides a search bar for "Looking for an OAuth2 client library?". The "Resources" section is partially visible. On the right side of the page, there are two dark grey boxes with white text. The top box says "You've linked your Genius account with this site. You're on your way to building something great!". The bottom box is titled "Add Genius Annotations To Your Own Site" and contains the code snippet: `<script src='\"https://genius.codes\"'></script>`.

<https://docs.genius.com/>

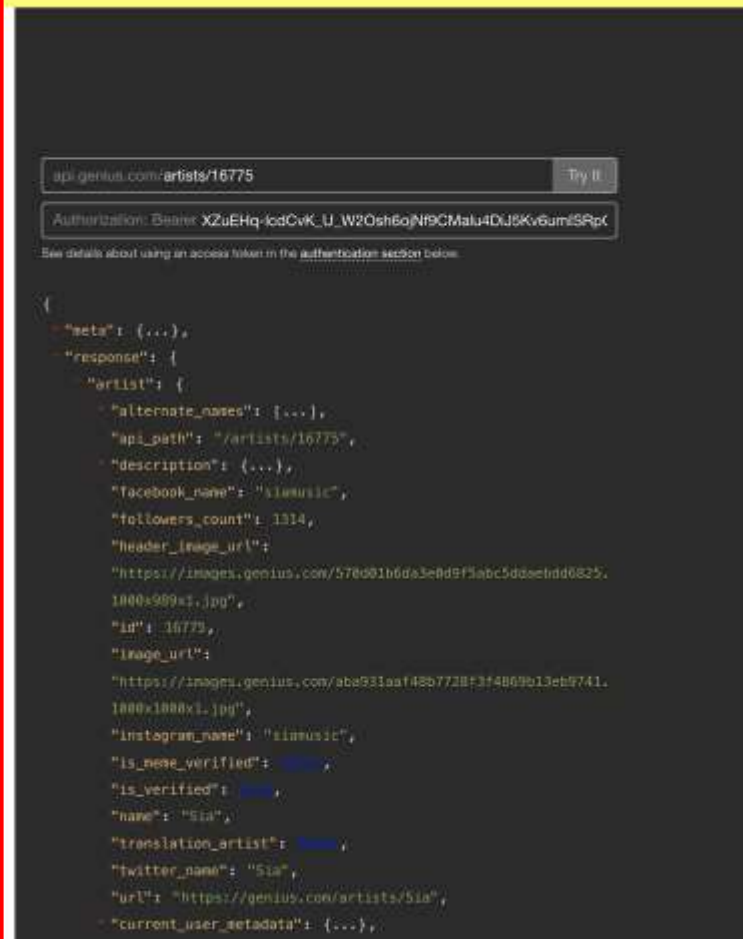
如何利用多方资源获取数据

- 以Genius音乐资讯平台为例：

检索到了歌手编号为16775的Sia的所有作品



The screenshot shows the Genius API documentation for the `GET /artists/:id` endpoint. The page includes a search bar, navigation links, and a description of an artist. The endpoint is used to retrieve data for a specific artist, with the `id` parameter being the ID of the artist (e.g., 16775). The `text_format` parameter allows for different output formats: `dox`, `plain`, and `html`, separated by commas. The `text_format` dropdown is currently set to `Text Format`.



The screenshot shows a REST client interface with the following details:

- URL: `api.genius.com/artists/16775`
- Authorization: `Bearer XZuEHq-4cdCvK_U_W9Osh6jNf9CMalu4DU5Kv6umISRpC`
- Response (JSON):

```
{
  "sets": (...),
  "response": {
    "artist": {
      "alternate_names": [...],
      "api_path": "/artists/16775",
      "description": (...),
      "facebook_name": "siamusic",
      "followers_count": 1314,
      "header_image_url": "https://images.genius.com/578d01b6da3e8d9f3abc5d0aebdd0825.1800x909x1.jpg",
      "id": 16775,
      "image_url": "https://images.genius.com/aba931aaf48b7728f3f4869b13eb9741.1800x1800x1.jpg",
      "instagram_name": "siamusic",
      "is_name_verified": true,
      "is_verified": true,
      "name": "Sia",
      "translation_artist": true,
      "twitter_name": "Sia",
      "url": "https://genius.com/artists/Sia",
      "current_user_metadata": {...}
    }
  }
}
```

如何利用多方资源获取数据

■ 利用集成式采集器爬取数据？

- 相较于API而言，更加快速便捷；通常不需要掌握编程技术，而是可以通过按步骤点击按钮，实现数据的采集与存储
- 常用渠道：
 - 八爪鱼：<https://www.bazhuayu.com/>（网站内有非常丰富的学习资源）
 - Stevesie：<https://stevesie.com/>
 - 小红书、抖音、快手、淘宝：<https://www.postman.com/solar-flare-375895/workspace/spider/documentation/3194348-da2b8207-b341-4679-845e-57ff2b9a2ec5>
 - Kickstarter：<https://apify.com/jaroslavhejlek/kickstarter-search#input>

如何利用多方资源获取数据

- 如果以上途径都无法获得的话，应该怎么办？

- 自己编写爬虫程序，可以以Python作为主要语言

- 网络上有非常丰富的视频、文字的学习资源：

- <https://www.runoob.com/w3cnote/python-spider-intro.html>

- 找淘宝店铺或其他渠道，一般数据集价格通常在几十~几百元不等。

优点：可以按照需求指定自己需要的字段、速度非常快（大约3~5天可以拿到数据）、避免自己爬数据不断返工以及不断遇到突发状况的情形（如随着采集数量增加，被平台封IP的可能性也会增加，需要不断更换IP；此外还有大量认证、调转等问题）、可以将大量宝贵的时间精力投入在数据分析、理论推导等更为重要的任务上。

- 在条件允许的情况下，专业的事情可以交给更为专业的人来做！