



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

数据科学导论

Introduction to Data Science

第二章 数据分析

陈恩红，黄振亚

Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页:

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2026.html>



数据分析基础

2

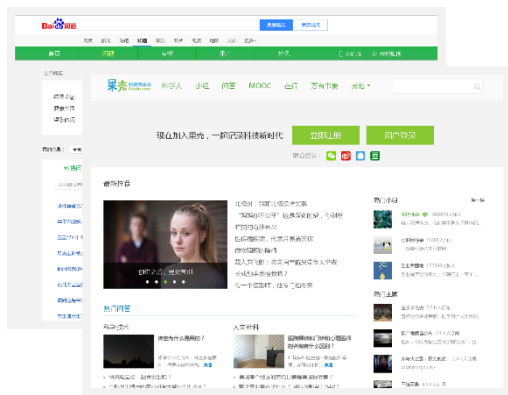
- 数据采集 Data Collection
- 数据预处理 Data Preprocessing
- 特征工程 Feature Engineering



数据采集

3

- 无时无刻产生数据，获得数据的方式多种多样



网页



测量



数据库



监控



传统媒体



数据采集

4

- 数据检索与开源
- 批量数据获取
 - 网络爬虫
- 数据生成



数据采集：数据检索

- 最简单、最灵活的数据获取方式就是依靠检索
- 搜索引擎





数据采集：数据检索

- 最简单、最灵活的数据获取方式就是依靠检索
- 学会使用搜索引擎

□ Google: 更适合:

■ 2018.9, Google

Google 搜索 GPT

找到 100 多个数据集

gpt-dataset
huggingface.co
上次更新日期: Feb 22, 2023

GPT2 Output Dataset
kaggle.com
zip
上次更新日期: Nov 12, 2019

Paraphrased Articles using GPT-3
kaggle.com
zip
上次更新日期: Nov 12, 2019

gpt-dataset
Haziqsayyed/gpt-dataset

探索网页: huggingface.co

73 篇学术文章引用了此数据集 (在 Google 学术搜索中查看)

数据集上次更新日期
Feb 22, 2023

作者
Haazique Sayyed

说明
Haziqsayyed/gpt-dataset dataset hosted on Hugging Face and contributed by the HF Datasets community



数据采集：公开数据

□ 代表性公开数据集

- ImageNet : <http://www.image-net.org/>
- Amazon从2008年开始就为开发者提供几十TB的开发数据
 - <http://aws.amazon.com/datasets>
- YouTube视频与社交网络: <http://netsg.cs.sfu.ca/youtubedata/>
- 用户评分MovieLens: <https://grouplens.org/datasets/movielens/>
- 网络数据-Large scale network: <https://snap.stanford.edu/data/>
- 教育数据:
 - ASSISTmentsData: <https://sites.google.com/site/assistmentsdata/home/>
 - BASEGroup: <https://github.com/bigdata-ustc/EduData>
- 阿里天池数据-数据平台: <https://tianchi.aliyun.com/dataset/>
- 公开大数据竞赛的数据: KDDCup, NeurIPS Challenge



数据采集：批量数据获取

8

- 大量数据的获取难以手动实现，需借助**爬虫程序**
 - 也有可能通过交易（购买）“数据”而得
- 网络爬虫是一个自动在网上抓取数据的程序
 - 爬虫本质上就是**下载**特定网站网页的HTML/JSON/XML数据，并对数据进行**解析**、**提取**与**存储**
 - 通常先定义一组**入口URL**，根据页面中的其他URL，**深度优先**或**广度优先**的遍历访问，逐一抓取数据

天猫怎么设置指纹支付 手机天猫指纹支付功能设置使用教程	支付 手机 购物	0 回答 8 秒钟前
欺凌近义词是什么	学习 教育	0 回答 8 秒钟前
杨志厚的毕业院校	求职就业 升学入学 毕业	0 回答 9 秒钟前
美国人那么喜欢玩车改车，他们的配件都是哪来的啊？原厂不会有多改改装配件给车主玩吧？	汽车 赛车 汽车改装	0 回答 9 秒钟前
PS中怎么建立一个外框比你画的图要灰（图的第一部分还有第3部分）	图像处理 软件 ps3	0 回答 10 秒钟前
王晶的儿子是谁	家庭关系 烦恼	0 回答 10 秒钟前
爱情保卫战上最没礼貌的女嘉宾是哪一期	明星 爱情 恋爱	0 回答 10 秒钟前
水管发热带可以直几个小时	生活常识 理工学科 水管	0 回答 11 秒钟前
草堂祭饭而致近。谓什么动物？	宠物	0 回答 11 秒钟前
梦见别人给我留言说自己的幸福自己把握是	解梦 恋爱	0 回答 11 秒钟前
想看书，没闲钱	学习 美食	0 回答 11 秒钟前
联想A320T手机系统升级失败开不了机怎样重新刷机	联想 刷机 联想手机	0 回答 11 秒钟前

热点要闻

个性推荐

进入推荐版

习近平对智利进行国事访问 发表署名文章 专题

同智利总统会谈 中拉媒体领袖峰会 向印度总理致慰问电 治国理政

李克强：打造改革创新开放的新标杆

“放暑假”座谈会 前正声会见台湾民意代表交流参访团 会见俄国防长

纸币要消失了？央行筹备数字货币

数字货币是否安全？能带来什么好处？纸币会不会作废？

国家卫计委：全国性公共场所控烟条例有望今年出台

韩日正式签署军事情报保护协定 共享朝鲜情报

揭秘：韩十一返回轮开航 曹晋都带回了什么

父子同日被终止人大代表资格 被移送司法机关

北京三审“气象灾害防治条例” 修订“霾”入法

大连一载23吨燃油渔船突燃大火：周围近百余艘渔船

韩媒：韩日签署《军情协定》为深化双边军事合作起步

特朗普：纽约时报对我非常粗鲁 不看出版能多活20多年

e租宝案件8月月底宣判 目前已缴涉案资产逾百亿

湖南浏阳11岁小姑娘一句话“叫停”景观工程

中国首次成为资本净输出国 流量反超日本居全球第二

2020年东京奥运会总经费拟定：约2万亿日元

父亲半道弃女儿遗体于垃圾桶内 当事人接受调查

外交部回应达赖窜访：藏方的错误做法损害中藏关系

北京市食药监局：网传水体污染致淡水鱼下架不可信

习近平出席中拉媒体领袖峰会开幕式

热搜新闻词 HOT WORDS

更多

习近平同智利总统巴切莱特举行会谈

习近平出席中拉媒体领袖峰会开幕式

全国性控烟条例有望今年出台

央行数字货币

韩日签署军事情报保护协定

明年起居民身份证可异地换领

重慶不該稱“不設防城市”

香港黃店被劫案 香港黃店被劫案

246種生猛活蝦 246種生猛活蝦

8亿网友

百度百家 BAIDU JIA

财经 文化 娱乐 体育

加入百家号

百度开放大批语音黑科技

拦截救助：动保恐怖主义

方舟子：木糖醇能预防龋齿吗？

方舟子

9/16/2026

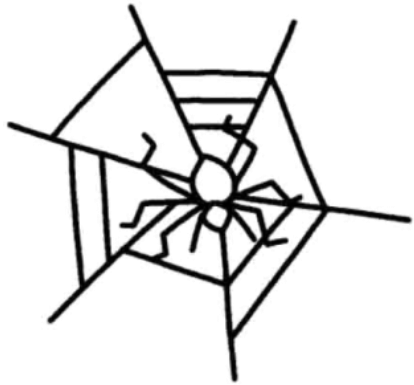


数据采集：网络爬虫

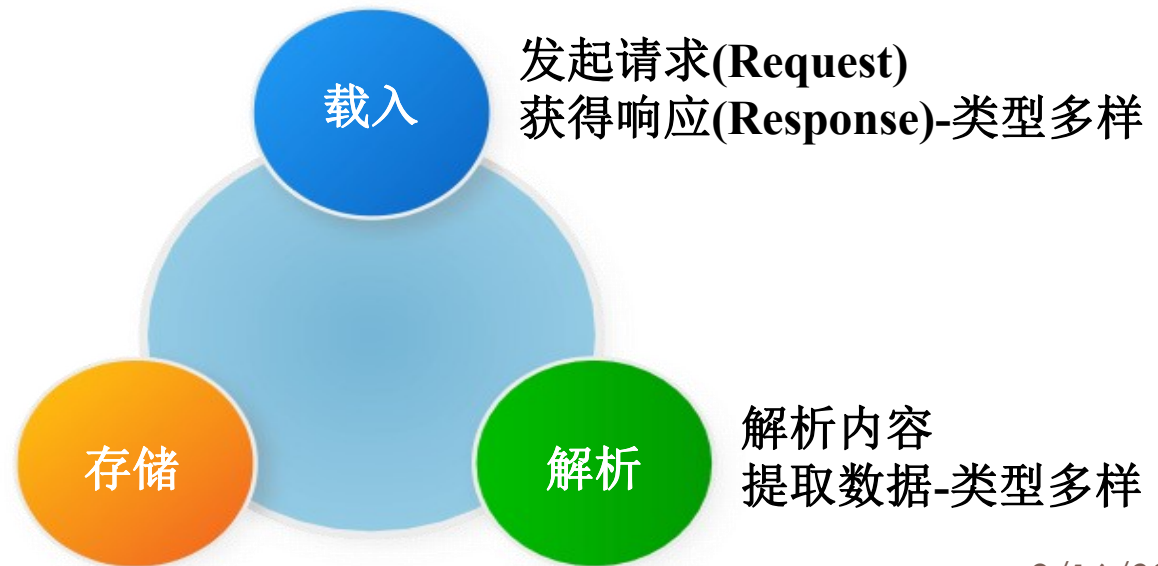
9

□ 网络爬虫是什么？

- 网络爬虫（又被称为网页蜘蛛，网络机器人，网页追逐者），是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。
 - 请求网站并提取数据的自动化程序
- 爬虫的行为可以划分为：载入、解析、存储，
 - 最复杂的部分为载入



获得数据
存储数据-类型多样



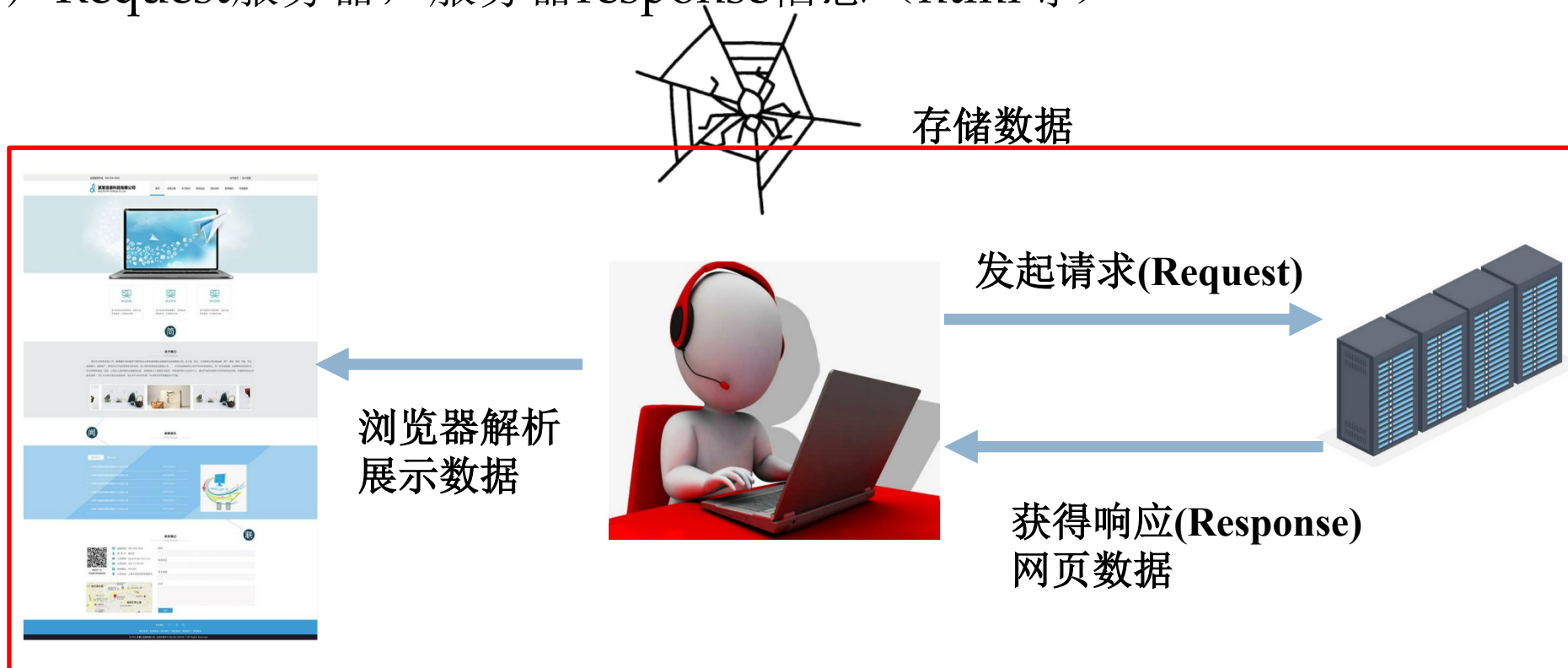
9/16/2026



数据采集：网站数据

10

- 访问网页示例
 - 网站数据主要依托于网页（html, 超文本标记语言）展示
 - 用户Request服务器，服务器response信息（html等）





数据采集：网站数据

11

□ 网页示例

- 网站数据主要依托于网页（html, 超文本标记语言）展示
- 用户Request服务器，服务器response信息（html等）
- 课程主页: <http://staff.ustc.edu.cn/~huangzhy/Course/DS2026.html>
- 查看网页源代码

数据科学导论

Introduction to Data Science

授课教师 [黄振亚](#)、[陈恩红](#) 课程代码 CS4023.01 开课学院 011 计算机科学与技术学院

课程信息

时间 第 1-18 周, 每周二第 8、9 节 (15:55-17:30)

教室 高新校区 GT-B212



数据采集：网站数据

12

□ 网页示例

- 网站数据主要依托于网页（html, 超文本标记语言）展示
- 用户Request服务器，服务器response信息（html等）
- 课程主页<http://staff.ustc.edu.cn/~huangzhy/Course/DS2026.html>
- 右键“检查”查看网页源代码



```

}
</style>
</head>
<body>
<main class="page">
  <header>
    <h1>数据科学导论</h1>
    <p class="english-title">Introduction to Data Science</p>
    <p class="course-meta">
      <span>
        <strong>授课教师</strong>
        <a href="http://staff.ustc.edu.cn/~huangzhy/">黄振亚</a>、
        <a href="http://staff.ustc.edu.cn/~cheneh/">陈恩红</a>
      </span>
      <span><strong>课程代码</strong> CS4023.01</span>
      <span><strong>开课学院</strong> 011 计算机科学与技术学院</span>
    </p>
  </header>

  <section aria-labelledby="course-info">
    <h2 id="course-info">课程信息</h2>
    <div class="section-body">
      <dl>
        <dt>时间</dt>
        <dd>第 1 - 18 周，每周二第 8、9 节（15:55 - 17:30）</dd>
        <dt>教室</dt>
        <dd>高新校区 GT-B212</dd>
      </dl>
    </div>
  </section>

  <section aria-labelledby="teaching-assistants">
    <h2 id="teaching-assistants">助教与作业提交</h2>
    <div class="section-body">
      <dl>
        <dt>助教</dt>
        <dd>罗胤玻、李宁</dd>
        <dt>邮箱</dt>
        <dd><a href="mailto:ds_intro2026@163.com">ds_intro2026@163.com</a></dd>
        <dt>QQ 群</dt>
        <dd><a href="https://qm.qq.com/q/nbiD5WAdvE">1050506077</a></dd>
      </dl>
    </div>
  </section>

```



网络爬虫：载入

14

- 载入：将目标网站数据下载到本地
 - Html, HyperText Markup Language
 - 爬虫程序向服务器发送网络请求 Request，获取相应的网页
 - 网站常用网络协议：http, https
 - 数据常用请求方式：get, post
 - get: 参数常放置在URL中
 - http://www.abc.com?p=1&q=2&r=3,
 - 问号后为参数
 - post: 参数常放置在一个表单中（报文头（header））
 - 在向目标URL发送请求时，将参数放置在一个网络请求的报文头中
 - 更安全



□ 数据常用请求方式: get, post

- **get:** 参数常放置在URL中

- `http://www.abc.com?p=1&q=2&r=3`, 问号后为参数

- 例如，[https://www.baidu.com/s?ie=\[中国科学技术大学\]](https://www.baidu.com/s?ie=[中国科学技术大学])





网络爬虫：载入

16

- 载入：将目标网站数据下载到本地
 - 数据常用请求方式：get, post
 - **post**：参数常放置在一个表单中
 - 在向目标URL发送请求时，将参数放置在一个网络请求的报文头中
 - 相比于Get，多了Form Data部分（请求体）
 - 更安全：登录操作常用（不会放在URL后面）

请求体



Baidu 用户名密码登录

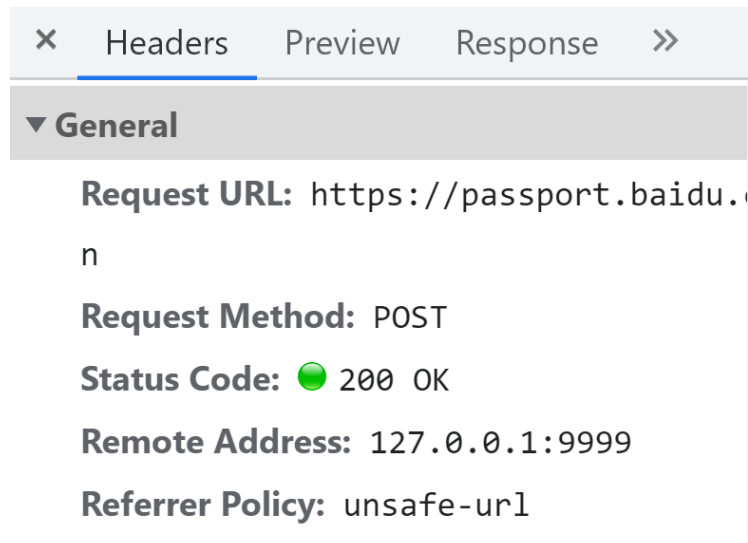
手机号/用户名/邮箱

密码

登录

[忘记密码?](#)

扫码登录 | 立即注册



Headers Preview Response >>

▼ General

Request URL: https://passport.baidu.com/signout/login?targetURL=https%3A%2F%2Fwww.baidu.com%2F%3C%2Fhtml%2Fv3Jump.html%3E&tpl=mn

Request Method: POST

Status Code: 200 OK

Remote Address: 127.0.0.1:9999

Referrer Policy: unsafe-url



▼ Form Data view source view URL-encoded

staticpage: https://www.baidu.com/cache/user/html/v3Jump.html

charset: UTF-8

token: ddfac7e17ce70dc6187ce33dffee73ed

tpl: mn

username:

password:



网络爬虫：载入

17

- 载入：将目标网站数据下载到本地
 - 数据常用请求方式：get, post
 - 获得服务器的响应：Response，即获取网页源代码

响应头

▼ Response Headers View source

Bdpagetype: 2

Bdqid: 0xdabb3fd4000073b3

×	Headers	Preview	Response	Initiator	Timing
▼ General					
Request URL:	http://staff.ustc.edu.cn/~huangzhy/				
Request Method:	GET				
Status Code:	200 OK (from disk cache)				
Remote Address:	127.0.0.1:10001				
Referrer Policy:	strict-origin-when-cross-origin				

响应状态：200， 404等

HTTP状态码

响应体，即网页源代码

```
</style>
</head>
<body>
  <main class="page">
    <header>
      <h1>数据科学导论</h1>
      <p class="english-title">Introduction to Data Science</p>
      <p class="course-meta">
        <span>
          <strong>授课教师</strong>
          <a href="http://staff.ustc.edu.cn/~huangzhy/">黄振亚</a>、
          <a href="http://staff.ustc.edu.cn/~chench/">陈恩红</a>
        </span>
        <span><strong>课程代码</strong> CS4023. 01</span>
        <span><strong>开课学院</strong> 011 计算机科学与技术学院</span>
      </p>
    </header>
    <section aria-labelledby="course-info">
      <h2 id="course-info">课程信息</h2>
      <div class="section-body">
        <dl>
          <dt>时间</dt>
          <dd>第 1 - 18 周，每周二第 8、9 节 (15:55 - 17:30) </dd>
          <dt>教室</dt>
          <dd>高新校区 GT-B212</dd>
        </dl>
      </div>
    </section>
    <section aria-labelledby="teaching-assistants">
      <h2 id="teaching-assistants">助教与作业提交</h2>
      <div class="section-body">
        <dl>
          <dt>助教</dt>
          <dd>罗胤璇、李宇</dd>
          <dt>邮箱</dt>
          <dd><a href="mailto:ds_intro2026@163.com">ds_intro2026@163.com</a></dd>
          <dt>QQ 群</dt>
          <dd><a href="https://qm.qq.com/q/nb1D5WAdvE">1050506077</a></dd>
        </dl>
      </div>
    </section>
  </main>
</body>
</html>
```

9/16/2026



网络爬虫：载入

18

- 实际操作：抓取一个静态网页步骤
 - 首先确定URL，例如：<http://www.baidu.com>
 - 其次确定请求的方式以及相关参数：
 - 直接用浏览器实现：chrome, firefox浏览器抓包工具，详见
 - <http://jingyan.baidu.com/article/3c343ff703fee20d377963e7.html>
 - 或者抓包工具：charles等，详见
 - <http://blog.csdn.net/jiangwei0910410003/article/details/41620363/>
 - 最后在代码中按照特定的请求方式（get, post）向URL发送参数，即可收到网页的结果



网络爬虫：载入

19

□ 但部分页面的数据是动态加载的

□ Ajax异步请求

- 网页中的部分数据需要浏览器渲染 (JavaScript调用接口获取数据)
- 用户的某些点击、下拉的操作触发才能获得

□ 解决方案：

- 借助抓包工具，分析Ajax某次操作所触发的请求，通过代码实现相应的请求
 - 有技术难度，但抓取速度快。
- 利用智能化的工具：selenium webdriver
 - 用程序控制驱动浏览器，模拟浏览器
 - 可以模拟实现人的所有操作
 - 操作简单，但是速度慢
 - 因为爬虫需要启动浏览器，浏览器需要渲染页面，所以速度比较慢
- 其他：Splash, Pyv8等





网络爬虫：载入

20

- **反爬虫：**随着网络爬虫对目标网站**访问频率的加大**，网站禁止爬虫程序继续访问
- **常见反爬手段：**
 - 出现用户登录界面，需要验证码
 - 禁止某个固定帐号或ip一段时间内访问网站
 - 更有甚者，直接返回错误的无用数据
- **应对措施：**
 - 优化爬虫程序，尽量减少访问次数，尽量不抓取重复内容
 - 使用多个cookie（网站用来识别用户的手段，每个用户登录会生成一个cookie）
 - 使用多个ip（可以用代理实现）

安全验证 ×

您的帐号可能存在安全风险，为了确保为您本人操作，请先进行安全验证。

发送成功

验证方式

186*****23手机 ▼

请输入六位验证码

重新发送(57)

确定



网络爬虫：解析(对比JSON与XML)

21

```
{
  "name": "中国",
  "province": [{
    "name": "黑龙江",
    "cities": {
      "city": ["哈尔滨", "大庆"]
    }
  },
  {
    "name": "广东",
    "cities": {
      "city": ["广州", "深圳", "珠海"]
    }
  },
  ....
}]
```

对象，成员：键值对

```
<?xml version="1.0" encoding="utf-8"?>
<country>
  <name>中国</name>
  <province>
    <name>黑龙江</name>
    <cities>
      <city>哈尔滨</city>
      <city>大庆</city>
    </cities>
  </province>
  <province>
    <name>广东</name>
    <cities>
      <city>广州</city>
      <city>深圳</city>
      <city>珠海</city>
    </cities>
  </province>
  .....
</country>
```



网络爬虫：抓取微博评论

22



抓包工具
获取请求

▼ General

Request URL: https://m.weibo.cn/api/comments/show?
Request Method: GET
Status Code: 200 OK
Remote Address: 123.125.106.67:443
Referrer Policy: no-referrer-when-downgrade

► Response Headers (14)

▼ Request Headers

[view source](#)

Accept: application/json, text/plain, */*
Accept-Encoding: gzip, deflate, br
Accept-Language: zh-CN,zh;q=0.8,en;q=0.6
Connection: keep-alive
Cookie: _T_WM=d9a7dba4dd130f79eaecac13c8906050; ALbktAKLUXNkW1un7fu00CXjkppVYn1wGjJ3knF4g.; SUBP=0p5NHD95Q0So5Re0.cS020Ws4Dqcjn-fHBxHzLxK-LB.eLBK5L505136002; M_WEIBOCN_PARAMS=featurecode%3D200003236170084375%26uicode%3D20000061%26fid%3D414183617
Host: m.weibo.cn
Referer: https://m.weibo.cn/status/4141836170084375
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 12.113 Safari/537.36
X-Requested-With: XMLHttpRequest



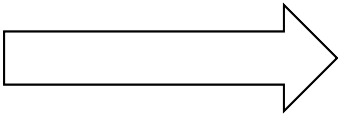
网络爬虫： 抓取微博评论

23

获得评论的json格式

```
京ICP备15025187号-1 邮箱: service@json.cn
{
  "mod_type": "mod/pagelist",
  "previous_cursor": "",
  "next_cursor": "",
  "card_group": [
    {
      "id": 4142016554789113,
      "created_at": "08-18 08:46",
      "source": "柔光自拍vivo X7",
      "user": {
        "id": 4142015488402958,
        "text": "回复
      "reply\_id": 4142015488402958,
      "reply\_text": "<a href='/u/'5187664653>@邓超</a>: 我也不知道<i class=
      "like\_counts": 10811,
      "liked": false,
      "mod\_type": "mod/single/infobox"
```

解析出需要的字段



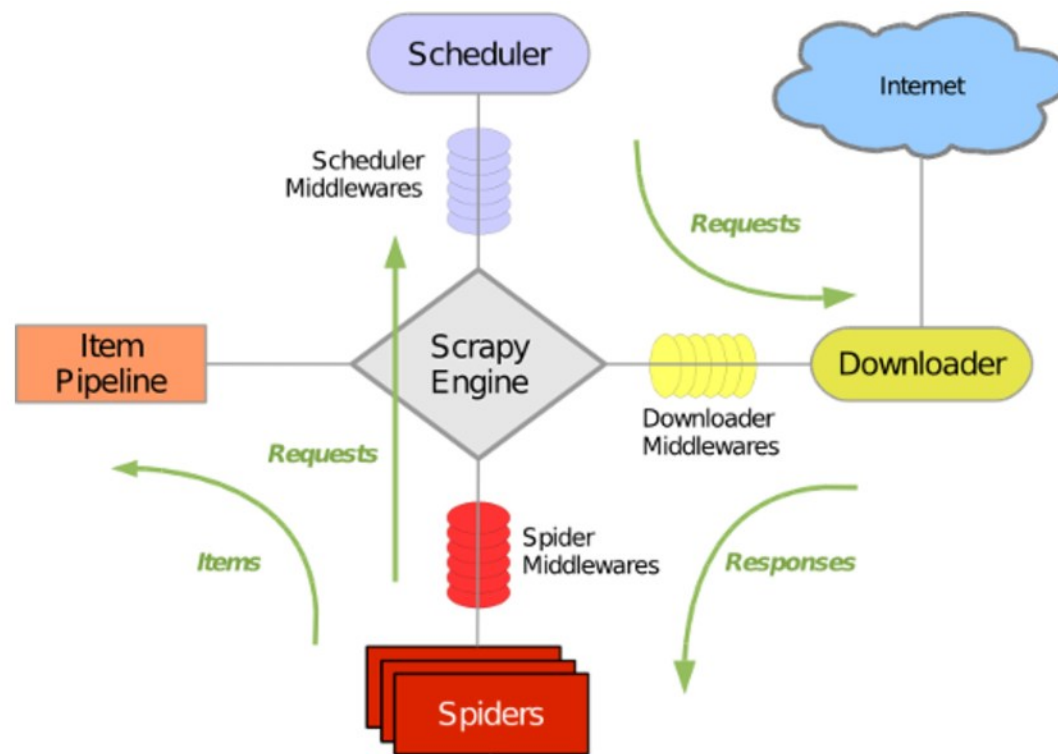
用户id	时间	内容
陈赫	08-18	天霸
邓超	08-18	我们都很好，谢谢大家
邓超	08-18	我也不知道
贼亮z1	08-17	迪丽热巴
.....		



网络爬虫：现有技术

24

- 基于Java的工具
 - HttpClient
 - Jsoup
- 基于Python的工具
 - **Scrapy**
 - BeautifulSoup



现有的爬虫框架很成熟，能够合理的控制爬取的过程，并有效的处理爬取过程中出现的各种异常，推荐使用Scrapy



- 注意网站规定
- 注意法律规定
 - 2021年6月1日, 《中华人民共和国数据安全法》
- 注意数据使用规范
- etc

课达编程 (CODIA) · 隐私政策

引言

课达编程重视用户的隐私。您在使用我们的服务时, 我们可能会收集和使用您的相关信息。我们希望通过本《隐私政策》向您说明, 在使用我们的服务时, 我们如何收集、使用、储存和分享这些信息, 以及我们为您提供的访问、更新、控制和保护这些信息的方式。本《隐私政策》与您所使用的课达编程服务息息相关, 希望您仔细阅读, 在需要时, 按照本《隐私政策》的指引, 作出您认为适当的选择。本《隐私政策》中涉及的相关技术词汇, 我们尽量以简明扼要的表述, 并提供进一步说明的链接, 以便您的理解。

您使用或继续使用我们的服务, 即意味着同意我们按照本《隐私政策》收集、使用、储存和分享您的相关信息。

如对本《隐私政策》或相关事宜有任何问题, 请通过 codia-support@bdaa.pro 与我们联系。



扩展阅读：PageRank算法

26

1. 提出者与时间



提出者：

Larry Page、Sergey Brin



提出时间：

1998年



背景：

斯坦福大学，作为Google早期搜索排序核心算法之一



2. 核心逻辑

- 1 将互联网看作由网页和超链接构成的有向图
- 2 一个网页被越多高质量网页引用，其重要性通常越高
- 3 网页的重要性会通过链接传递，并通过迭代计算不断更新

$$PR(A) = \frac{(1-d)}{N} + d \times \sum \left(\frac{PR(T_i)}{C(T_i)} \right)$$

d: 阻尼系数，通常取 0.85
 T_i : 指向A的页面
 $C(T_i)$: 页面 T_i 的出链数



3. 算法的价值与对内容检索的重要性



显著提升搜索结果排序质量，帮助用户更快找到权威内容



不仅看关键词是否出现，还利用链接关系评估内容的重要性



降低低质量页面和简单堆砌关键词页面对排序的干扰



为后续搜索引擎、学术检索、推荐系统和信息检索研究提供重要基础



搜索



内容



排序提升



更广泛的应用



PageRank通过“链接投票 + 重要性传递”重塑了网页排序思想，是现代内容检索发展的关键里程碑。

BETTER SEARCH
A BRIGHTER WORLD



数据采集：课后练习（不考核）

27

- 写一个爬虫，爬取一个感兴趣的网站

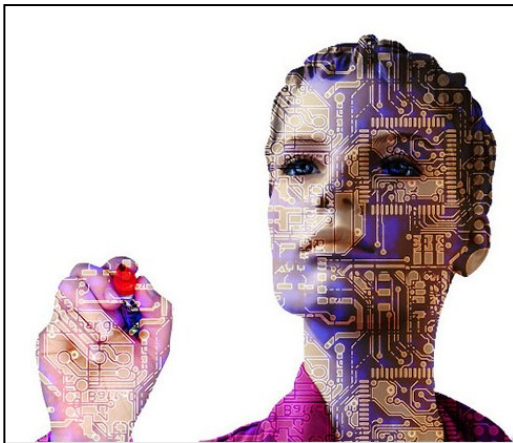


数据采集：自动生成数据

28

➤ AI: 1993 – 现在

- ❑ 信息检索、互联网搜索
- ❑ 机器学习、自然语言处理
- ❑ 数据挖掘、大数据分析
- ❑ “深蓝” 战胜人类国际象棋冠军卡斯帕罗夫



9/16/2026



数据采集：自动生成数据

29



OpenAI-ChatGPT

OpenAI于2022年11月30日发布的文本模态聊天机器人，被社会广泛关注。



讯飞-星火大模型



百度-文心一言



阿里-通义千问



腾讯-混元大模型



智谱-GLM



月之暗面-KIMI

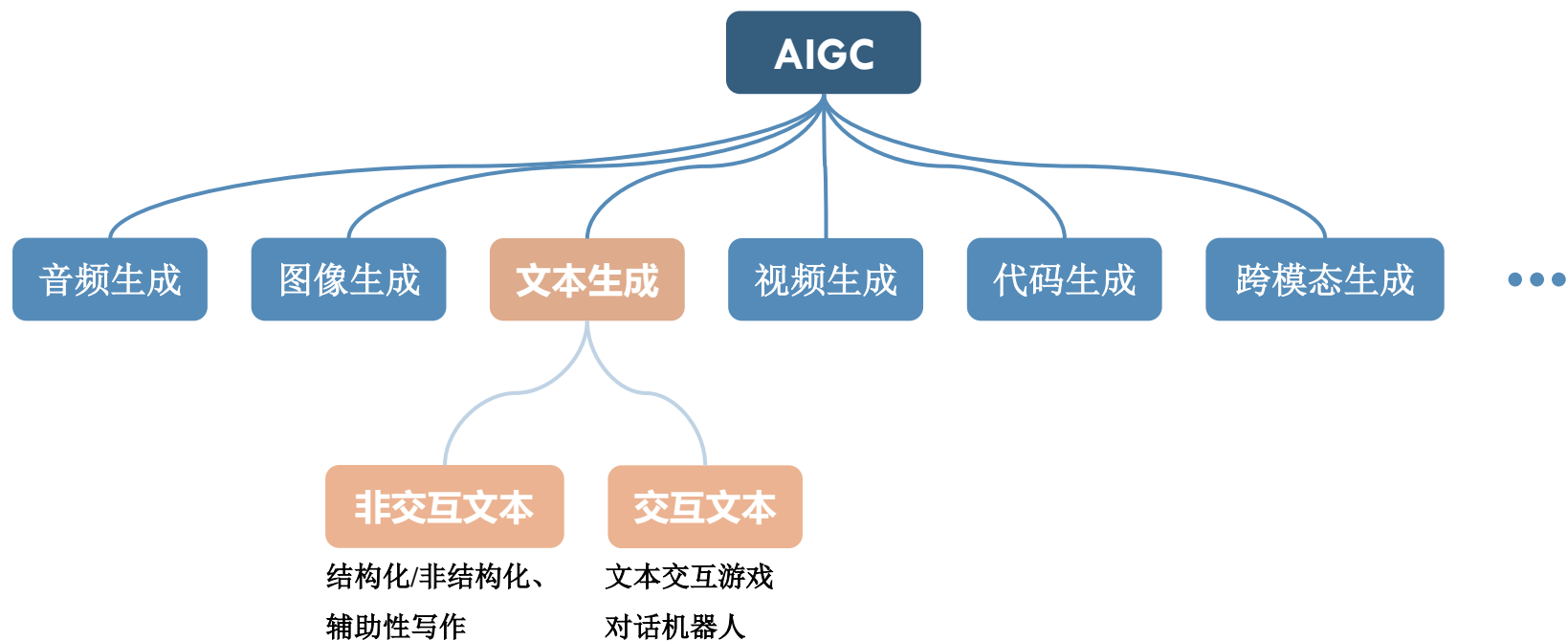
国内大模型从大模型多重能力构建，多模态语义理解等角度快速发展，在多个领域超过ChatGPT。



数据采集： 自动生成数据

30

AI Generated Content (AIGC, 人工智能自动生成内容) 技术，是一种新型内容创作方式，可以定制化生成需要的数据。

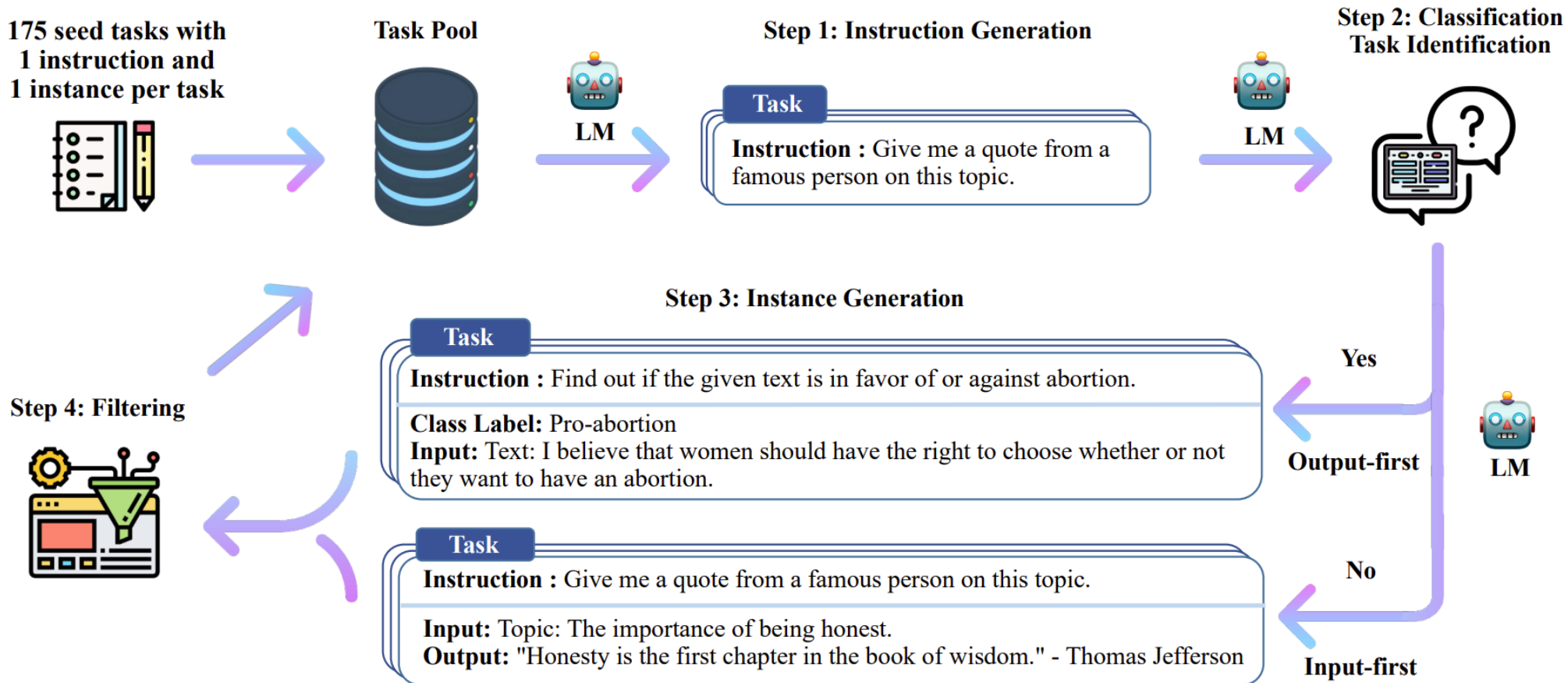




数据采集：自动生成数据

31

文本数据生成



示例：基于少量种子任务，驱动大模型生成新指令、识别任务类型、合成输入输出实例、过滤清洗，自动化迭代扩充数据



数据采集：自动生成数据

32

图片数据生成



Input Canny edge



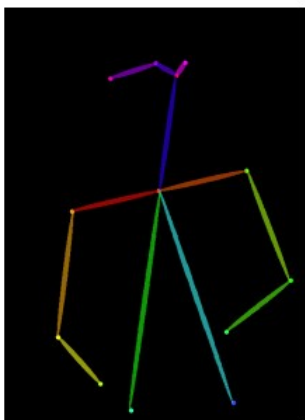
Default



"masterpiece of fairy tale, giant deer, golden antlers"



"..., quaint city Galic"



Input human pose



Default



"chef in kitchen"



"Lincoln statue"

ControlNet通过注入边缘图、人体骨骼姿态等方式，定向合成符合特定几何与物理约束的高质量训练图像



数据采集： 自动生成数据

33

视频生成



【提示词】逼真的特写视频展现了两艘海盗船在一杯咖啡中激烈交战，它们在咖啡液中航行。

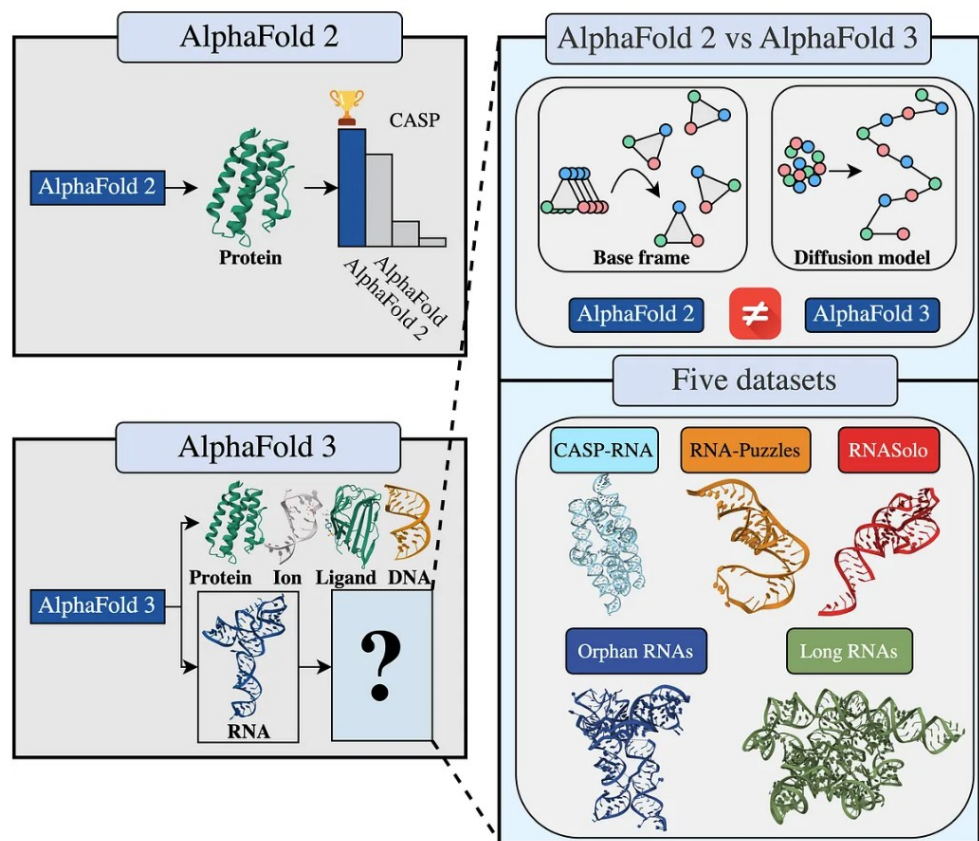


数据采集：自动生成数据

34

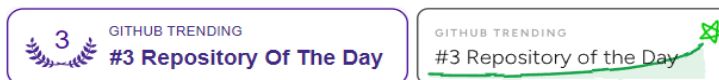
非结构化数据生成

- 通过图片、模板等资源生成PPT
- 生成蛋白质数据



PPT Master — AI generates native PowerPoint from any document

version v6.3.0 License MIT Stars 53k G-Star 445 Stars agentic leaderboard #3





大模型的扩展定律

Scaling law

- 损失值随着模型规模、训练数据量、训练计算量的增加，呈现出幂律下降的趋势

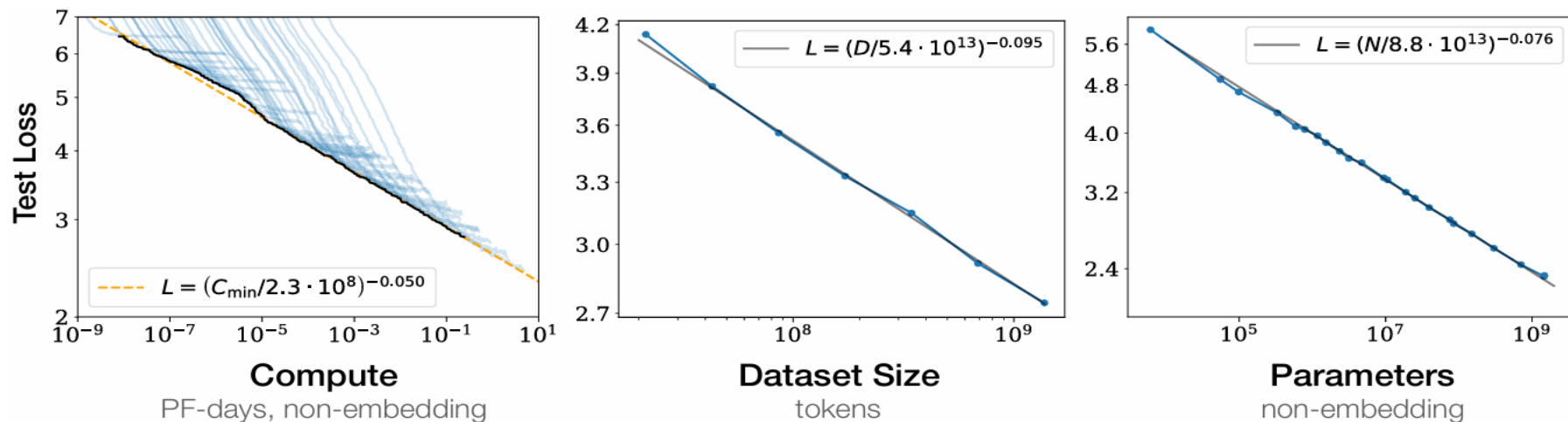


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

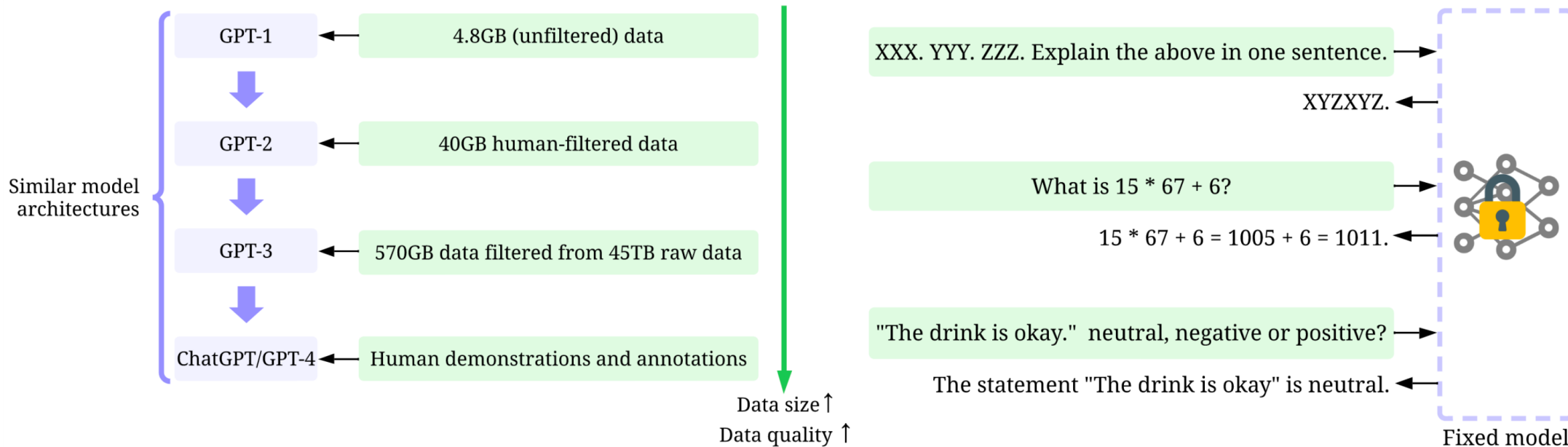
$$L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$$



以数据为中心的AI

36

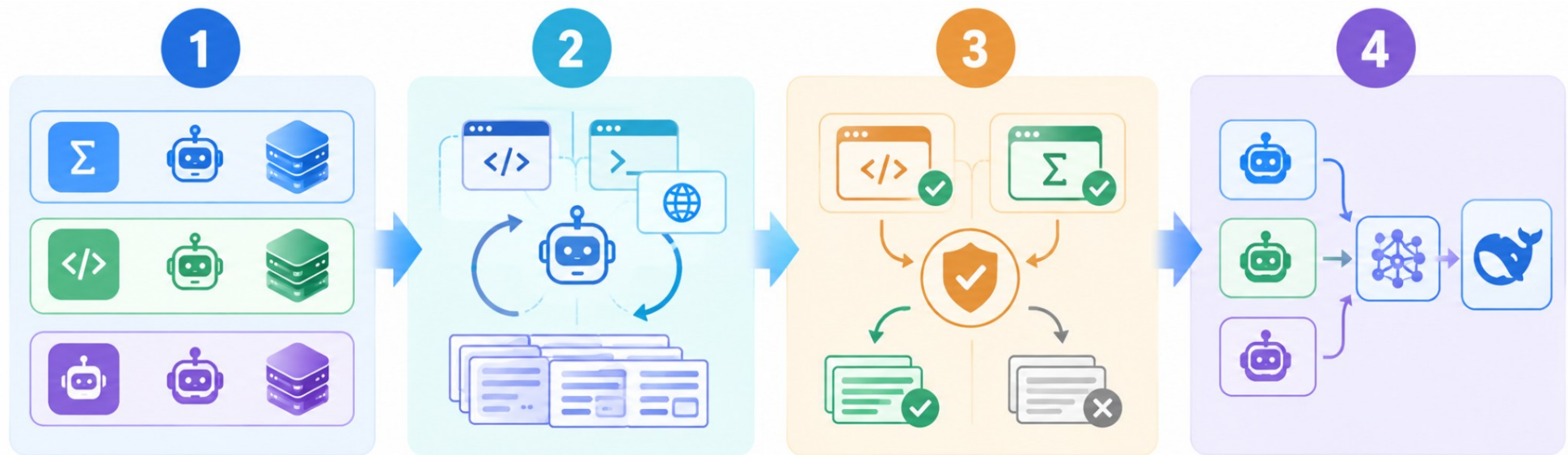
- 人工智能逐渐从以模型为中心过渡到以数据为中心
 - GPT成功的数据基石：** GPT进化中，模型结构保持相似，训练数据的规模、质量得到极大提升
 - 模型对数据的需求急速上升





以数据为中心的AI

- DeepSeek-V4: 多专家训练与在线策略蒸馏
 - **专家分立**: 针对数学、代码与智能体任务独立训练垂直专家模型
 - **动态生成**: 依托代码沙箱与交互环境, 自主探索并合成数万步执行轨迹
 - **真值验证**: 引入真实编译器与求解器, 闭环校验并过滤合成数据的幻觉
 - **在线蒸馏**: 利用 Reverse-KL 散度将多专家策略无损聚合回统一大模型





以数据为中心的AI

□ FAIR数据原则：

- 可发现（Findable）
- 可访问（Accessible）
- 可互操作（Interoperable）
- 可复用（Reusable）

传统科学数据原则

□ FAIR-R数据原则：

新时代科学数据原则

- **AI就绪（Ready for AI）**：审计数据偏见、强化机器可读性与版权合法性，专为模型训练赋能



以数据为中心的AI

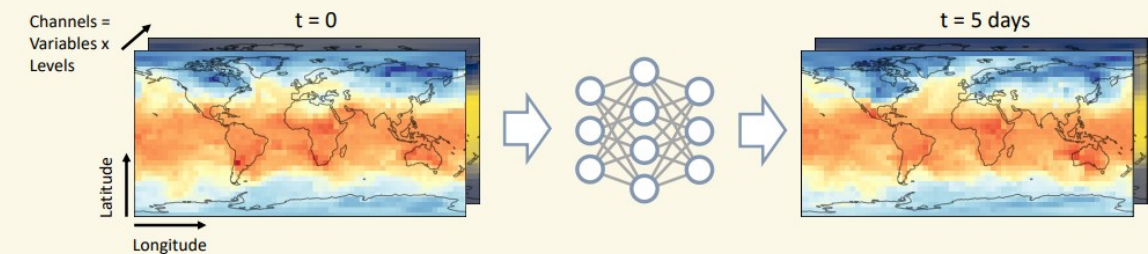
□ AI-Ready 数据集

- **数据困境：** 原始数据格式不一、质量参差且存在大量噪声；数据清洗消耗算力与人效
- **AI-Ready：** 经过格式统一、清洗除错和精准标注的数据，让数据可以即取即用
 - **【气象】** WeatherBench
 - **【医学】** MIMIC
 - **【智驾】** nuScenes
 - **【材料】** Materials Project
 - **【天文】** Galaxy Zoo
 - **【金融】** LOBSTER
 -

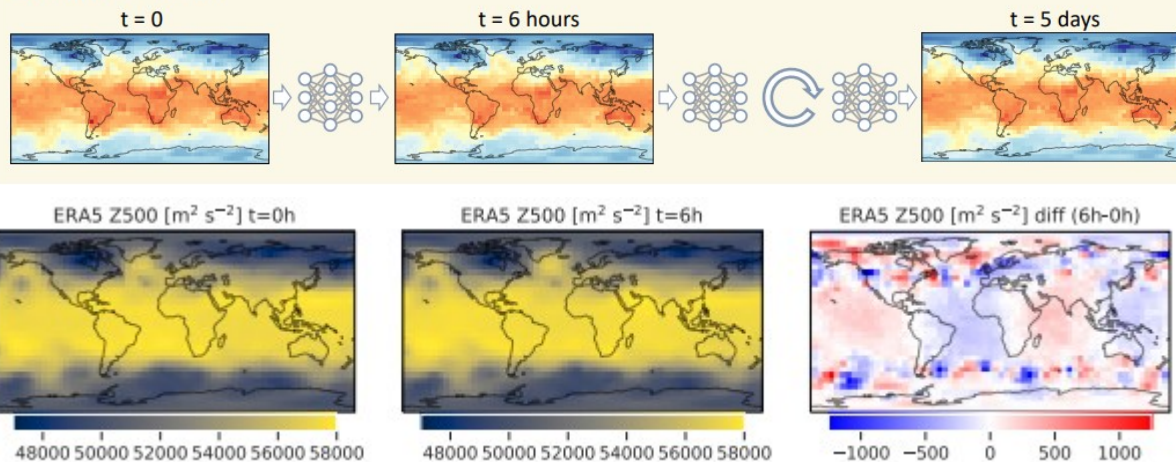


以数据为中心的AI

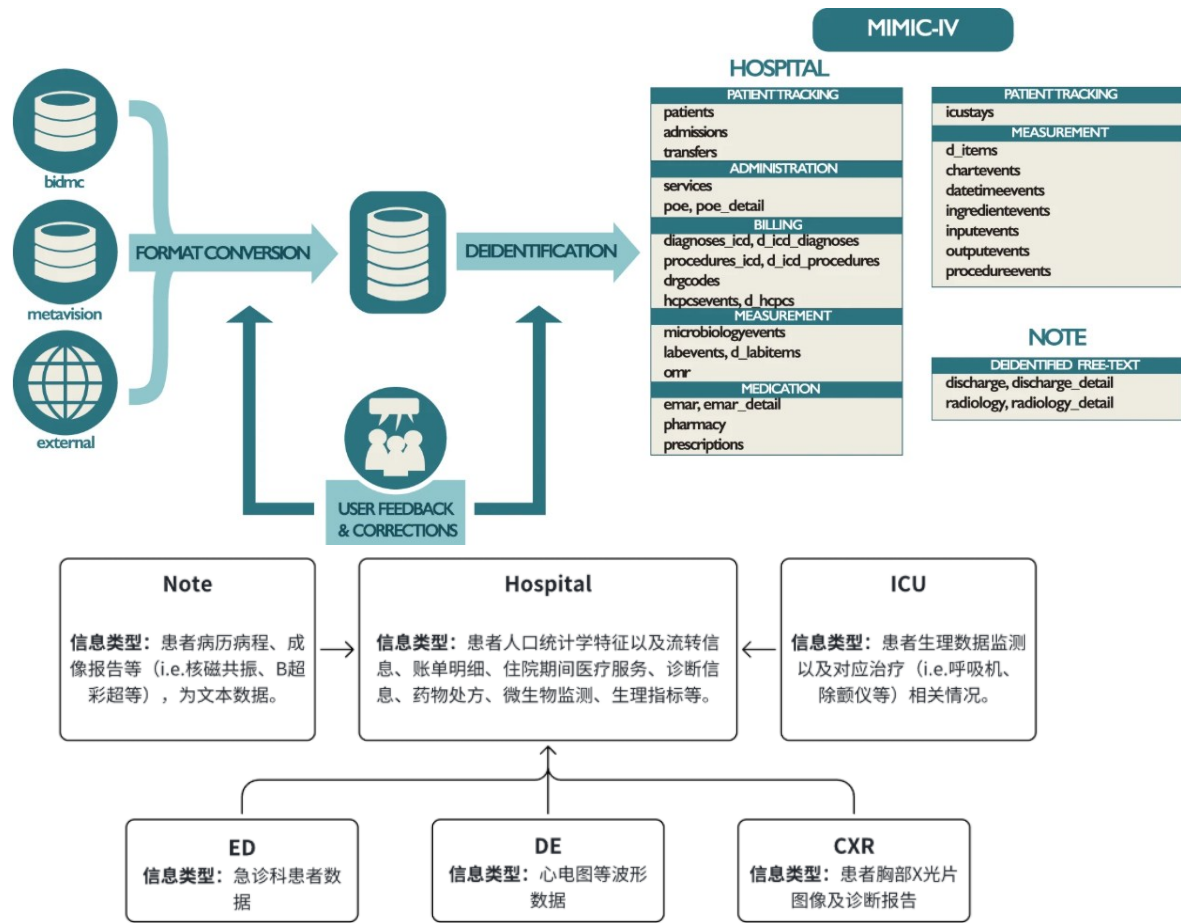
a) Direct prediction



b) Iterative prediction



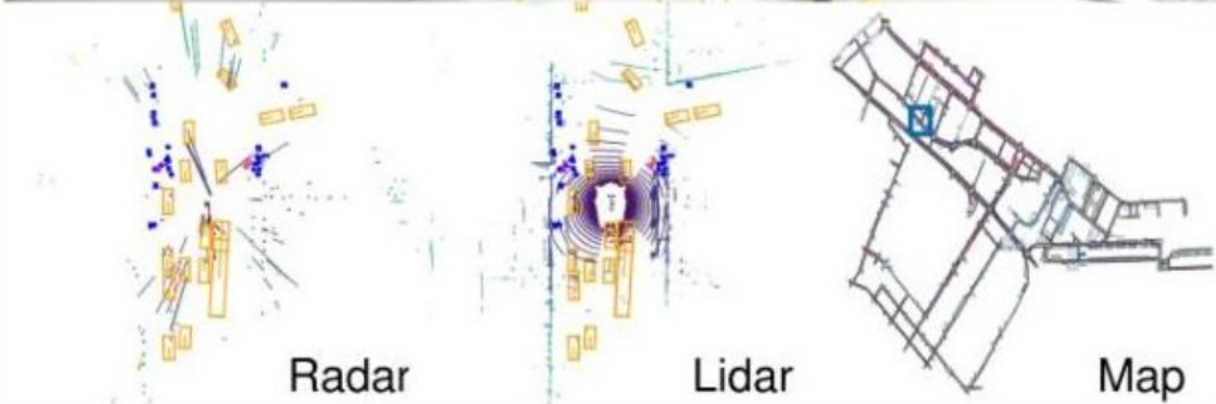
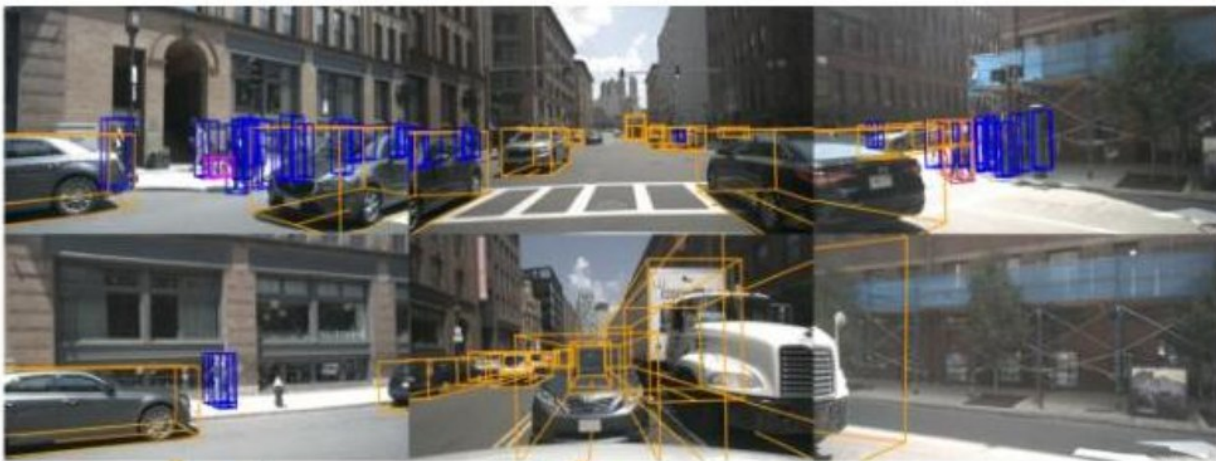
WeatherBench 气象数据集：欧洲气象中心的原始文件被谷歌等团队重构为规则的空间三维网格（标准经纬度、统一气压高度层），并以 Zarr 等高效格式存储



MIMIC 医学数据集：几十万重症患者杂乱的监护仪波形、用药记录、化验单等异构数据进行了统一对齐，构建为一个关系数据库



以数据为中心的AI



nuScenes 智能驾驶数据集：将车身传感器数据做了毫秒级的时间硬同步与空间位置标定。人工统一打标了精确的 3D 边界框，给行人和车辆标定了唯一的追踪 ID。

PrNi₃
mp-1102445

TABLE OF CONTENTS

- Summary
- Crystal Structure
- Properties
- Property Origins
- Contributed Data
- Literature References
- External Links
- More
- Related Materials

doi 10.17188/1743196

Energy Above Hull	0.000 eV/atom
Space Group	R3m
Band Gap	0.00 eV
Predicted Formation Energy	-0.284 eV/atom
Magnetic Ordering	Unknown
Total Magnetization	0.52 μB/f.u.
Experimentally Observed	Yes

Description (Auto-generated)

PrNi₃ crystallizes in the trigonal R3m space group. There are two inequivalent Pr sites. In the first Pr site, Pr is bonded in a 12-coordinate geometry to twelve Ni atoms. There are a spread of Pr-Ni bond distances ranging from 2.90-3.18 Å. In the second Pr site, Pr is bonded in a 18-

Symmetry	
Crystal System	Tetragonal
Lattice System	Tetragonal
Hall Number	-1 4 2
International Number	139
Symbol	I4/mmm
Point Group	4/mmm

Number of Atoms	10
Density	5.16 g·cm ⁻³
Dimensionality	3D
Possible Oxidation States	Co ^{1, 50+} , K ⁺ , Se ²⁻

Materials Project 材料数据集：晶体结构全部用CIF格式重构。统一用密度泛函理论 DFT 计算出了数百万种无机化合物的真实物理性质（如能带间隙、形成能），直接作为算法训练的真值标签



数据分析基础

42

- 数据采集 Data Collection
- 数据预处理 Data Preprocessing
- 特征工程 Feature engineering