



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

1

新媒体大数据分析

New Media Big Data Analysis

第二章 数据分析

黄振亚，朱孟潇，张凯，唐秀秀

课程主页：

<http://staff.ustc.edu.cn/~huangzhy/Course/NM2026.html>

助教：谭施越 朱嘉骏

bigdata_20262026@163.com

9/16/2026



数据分析

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

9/16/2026



数据采集

4

- 数据检索
- 公开数据
- 批量数据获取
 - 网络爬虫
- 数据筛选
- 数据生成



数据采集：数据检索

5

- 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
 - 百度：适合于搜索中文信息
 - Google：更适合搜索英文信息



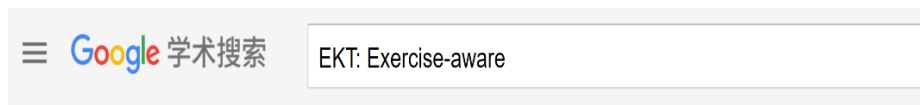
9/16/2026



数据采集：数据检索

6

- 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: 更适合搜索英文信息
 - Google Scholar, DBLP



文章 找到约 82 条结果 (用时0.04秒)

- 时间不限
2021以来
2020以来
2017以来
自定义范围...
- 按相关性排序
按日期排序
- 不限语言
中文网页
简体中文网页
- Ekt: Exercise-aware knowledge tracing for student performance prediction
Q Liu, Z Huang, Y Yin, E Chen, H Xiong... - IEEE Transactions on ..., 2019 - ieeexplore.ieee.org
For offering proactive services (eg, personalized exercise recommendation) to the students in computer supported intelligent education, one of the fundamental tasks is predicting student performance (eg, scores) on future exercises, where it is necessary to track the ...
☆ 99 被引用次数: 77 相关文章 所有 12 个版本
- [HTML] Exercise Hierarchical Feature Enhanced Knowledge Tracing
H Tong, Y Zhou, Z Wang - International Conference on Artificial ..., 2020 - Springer
... Eur. J. Psychol. Assess. 16(1), 3 (2000)CrossRefGoogle Scholar. 5. Huang, Z., et al.: Ekt:Exercise-aware knowledge tracing for student performance prediction. IEEE Trans. Knowl. Data Eng. (2019) Google Scholar. 6. Johnson, SC: Hierarchical clustering schemes ...
☆ 99 被引用次数: 4 相关文章 所有 6 个版本
- mpa Context aware Knowledge Tracing Integrated with The Exercise

Index of /xml

Name	Last modified	Size	Description
Parent Directory		-	
CHANGES.txt	2019-11-22 21:20	3.5K	
README.txt	2019-11-22 21:20	3.5K	
dblp.dtd	2019-11-22 21:20	12K	
dblp.xml.gz	2021-09-14 02:20	624M	
dblp.xml.gz.md5	2021-09-14 02:20	46	
docu/	2018-03-01 16:43	-	
osd.xml	2020-12-18 16:26	1.5K	
release/	2019-08-20 15:57	-	

Apache/2.4.29 (Ubuntu) Server at dblp.uni-trier.de Port 443



数据采集：数据检索

7

- 最简单、最灵活的数据获取方式就是依靠检索

□  × ? !

▼ 上次更新日期 ▼ 下载格式 ▼ 使用权限 ▼ 学科 ▼ 提供方 免费

找到 100 多个数据集



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

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

kaggle 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

8

- 国内常见公开数据渠道
 - 国家相关部门统计信息
 - 中国银行业监督管理委员会
 - 中国国家统计局

数据解读

更多>>

2015年8月社会融资规模增量统计数据报告
2015年8月金融统计数据报告
2015年7月社会融资规模增量统计数据报告
2015年7月金融统计数据报告
2015年上半年地区社会融资规模增量统计数据

统计信息	
• 2015年总资产、总负债 (月度)	2015-08-25
• 2015年总资产、总负债 (季度)	2015-08-10
• 2015年银行业金融机构用于小微企业的贷款情况表 (季度)	2015-08-10
• 2015年商业银行主要指标分机构类情况表 (季度)	2015-08-10
• 2015年商业银行主要监管指标情况表 (季度)	2015-08-10
• 2015年银监会监管统计信息发布日程表	2015-02-13
• 2014年总资产、总负债 (季度)	2015-02-13
• 2014年商业银行主要监管指标情况表 (季度)	2015-02-13
• 2014年商业银行主要指标分机构类情况表 (季度)	2015-02-13

代表性公开数据集

- 1400万的图像数据

■ <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视频的统计与社交网络数据

■ <http://netsg.cs.sfu.ca/youtubedata/>

统计公报

更多

- 年度统计公报
- 经济普查公报
- 人口普查公报
- 农业普查公报
- R&D普查公报
- 其他统计公报
- 基本单位普查公报
- 工业普查公报
- 三产普查公报

9/16/2026



数据采集：公开数据

9

□ 代表性公开数据集

- 用户评分MovieLens: <https://grouplens.org/datasets/movielens/>
- 文本数据-头条: <https://github.com/aceimnorstuvwxz/toutiao-text-classfication-dataset>
- 金融数据-股票: <https://github.com/asxinyu/Stock>
- 网络数据-Large scale network: <https://snap.stanford.edu/data/>
- 教育数据:
 - ASSISTmentsData-学业: <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

10

- 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





数据采集：网络爬虫

11

□ 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

存储

载入

发起请求(Request)
获得响应(Response)-类型多样

解析

解析内容
提取数据-类型多样

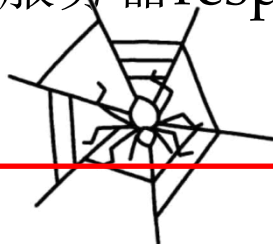


数据采集：网站数据

12

□ 访问网页示例

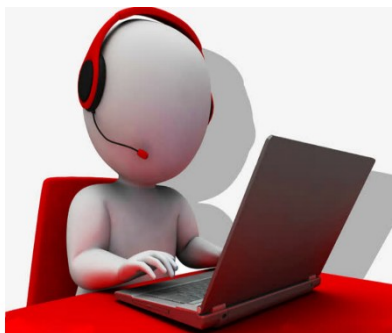
- 网站数据主要依托于网页（html, 超文本标记语言）展示
- 用户Request服务器，服务器response信息（html等）



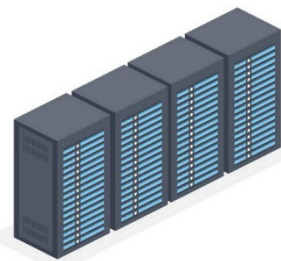
存储数据



浏览器解析
展示数据



发起请求(Request)



获得响应(Response)
网页数据



数据采集： 网站数据

13

□ 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/NM2026.html>
- 右键“检查”查看网页源代码



数据采集：网站数据

14

```
1 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
2 <!-- saved from url=(0039)file:///Users/kevin/Desktop/NM2024.html -->
3 <html><head><meta http-equiv="Content-Type" content="text/html; charset=GBK">
4
5 </head>
6 <body>
7 <font size="8px"><strong> <center>New Media Big Data Analysis </center></strong> <font>
8 <font size="8px"><strong> <center>新媒体大数据分析 </center></strong> <font>
9 <br>
10 <font size="6px"><strong> <center>NNM2012 </center></strong> <font>
11 <font size="6px"><strong> <center>学院：人文学院-025科技传播系 </center></strong> <font>
12 <font size="6px">
13 <strong>
14   <center>课程教师：
15   <a href="https://stc.ustc.edu.cn/2021/0601/c4503a511121/page.htm" target="-parent">朱孟潇</a>,
16   <a href="http://staff.ustc.edu.cn/~huangzhy/" target="-parent">黄振亚</a>,
17   <a href="https://faculty.ustc.edu.cn/zhangkai123/zh_CN/index.htm" target="-parent">张凯</a>,
18   <a href="https://stc.ustc.edu.cn/2026/0513/c4503a738364/page.htm" target="-parent">唐秀秀</a>
19   </center>
20 </strong>
21 <font>
22 <br><font size="6px">
23   <strong>
24   <center>上课时间：(1-18周)每周一第6 、 7节：14:00-15:35，教室：东区2303</center>
25   </strong>
26   <font>
27
28 <hr size="1px" noshade="">
29 助教（含作业提交）：
30 <br><div><font size="5px">
31   <li>谭施越，朱嘉骏</li>
32   <li>课程群-QQ群：1071511308 </li>
33   <li>课程邮箱：bigdata_20262026@163.com </li><br></font>
34 </div>
35
```



网络爬虫：载入

15

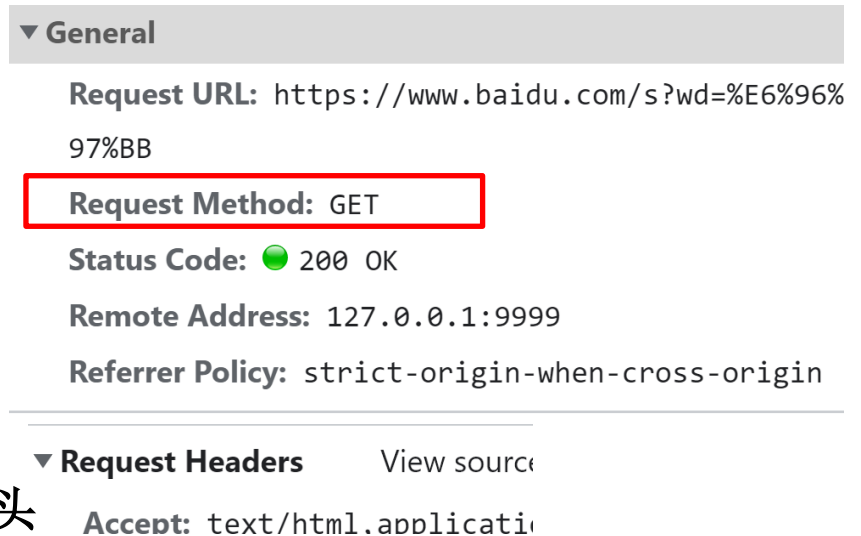
- 载入：将目标网站数据下载到本地
 - Html, HyperText Markup Language
 - 爬虫程序向服务器发送网络请求 Request，获取相应的网页
 - 网站常用网络协议：http, https
 - 数据常用请求方式：get, post
 - get：参数常放置在URL中
 - http://www.abc.com?p=1&q=2&r=3,
 - 问号后为参数
 - post：参数常放置在一个表单中（报文头（header））
 - 在向目标URL发送请求时，将参数放置在一个网络请求的报文头中
 - 更安全
 - https://ustc.edu.cn/



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
 - **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?wd=新闻



请求头



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

■ **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）

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.

n

Request Method: POST

Status Code: 200 OK

Remote Address: 127.0.0.1:9999

Referrer Policy: unsafe-url

请求体

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: huangzhy92

password: Jk5LCPMYHxJr6

JyR31zjQ6WKlg10sW0jf12!

3jz15ofHQtMTeUneEHAbKWk

9/16/2026



网络爬虫：载入

18

- 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

响应体，即网页源代码

▼ 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

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<!-- saved from url=(0039)file:///Users/kevin/Desktop/NM2024.html -->
<html><head><meta http-equiv="Content-Type" content="text/html; charset=GBK">
</head>
<body>
<font size="8px"><strong> <center>New Media Big Data Analysis </center></strong> <font>
<font size="8px"><strong> <center>新媒体大数据分析 </center></strong> <font>
<br>
<font size="6px"><strong> <center>NM2012 </center></strong> <font>
<font size="6px"><strong> <center>学院：人文学院-025科技传播系 </center></strong> <font>
<font size="6px">
<strong>
<center>课程教师：
<a href="https://stc.ustc.edu.cn/2021/0601/c4503a511121/page.htm" target="-parent">朱孟潇</a>,
<a href="http://staff.ustc.edu.cn/~huangzhy/" target="-parent">黄振亚</a>,
<a href="https://faculty.ustc.edu.cn/zhangkai123/zh_CN/index.htm" target="-parent">张凯</a>,
<a href="https://stc.ustc.edu.cn/2026/0513/c4503a738364/page.htm" target="-parent">唐秀秀</a>
</center>
</strong>
<font>
<br><font size="6px">
<strong>
<center>上课时间：(1-18周)每周一第6、7节：14:00-15:35，教室：东区2303</center>
</strong>
<font>
<br>
<font size="1px" noshade="">
助教（含作业提交）：
<br><dir><font size="5px">
<li>谭施越，朱嘉骏</li>
<li>课程群-QQ群：1071511308 </li>
<li>课程邮箱：bigdata_20262026@163.com </li><br></font>
</dir>
```



网络爬虫：载入

19

- 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

20



□ 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

21

- **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

请输入六位数字验证码

重新发送(57)

确定



网络爬虫：解析

22

□ 解析：在载入的结果中**抽取特定的数据**，载入的结果主要分成三类html、json、xml

□ html

- Java工具包：jsoup等
- Python工具包：beautifulSoup等

□ json

- Java工具包：json-lib、org-json、jackson等
- Python工具包：json、demjson等

□ Xml

- Java工具包：dom4j等
- Java工具包：xml、libxml2等



解析内容
提取数据-
类型多样



网络爬虫：解析(对比JSON与XML)

23

```
{  
  "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

24

**邓超**
8-17 20:49 来自 iPhone 7 Plus

跑男最新名单.....

344900

303031

转发 344900

评论 303031

**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zI**
08-17
迪丽热巴💕💕

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

9/16/2026



```

{
  "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": @Object {...},
      "text": "回复
```

解析出需要的 的字段

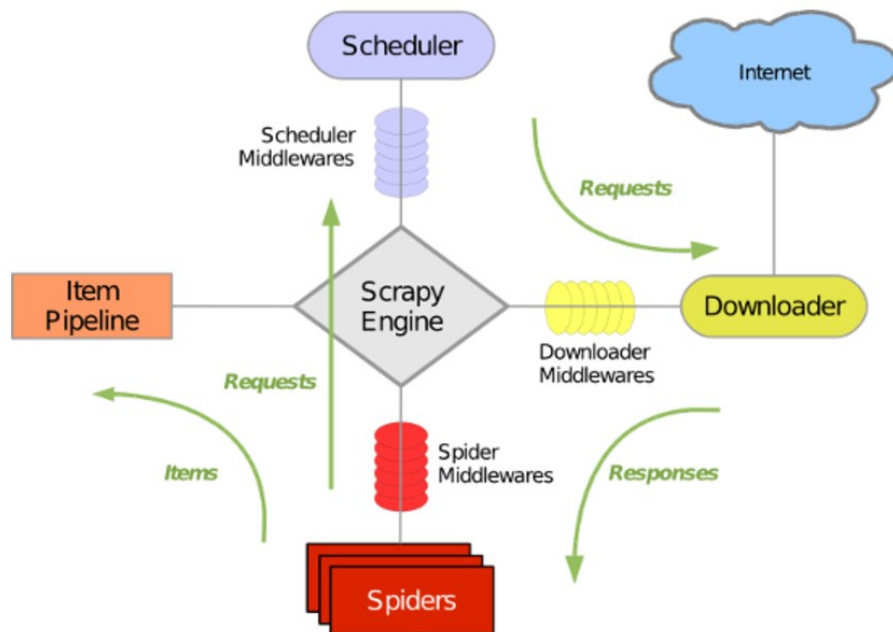
9/16/2026



网络爬虫：现有技术

26

- 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



数据采集

27

- 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。



数据采集：课后练习（不考核）

28

- 写一个爬虫，爬取一个感兴趣的网站

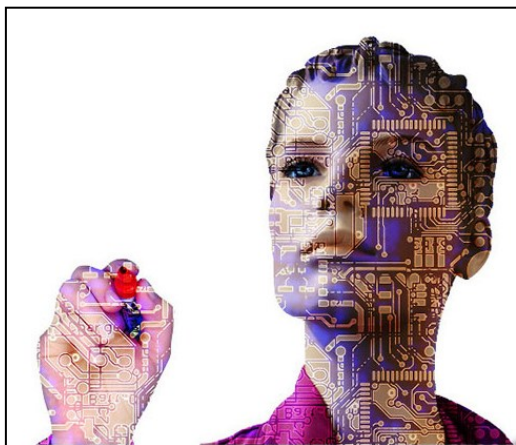


数据采集：自动生成数据

29

➤ 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

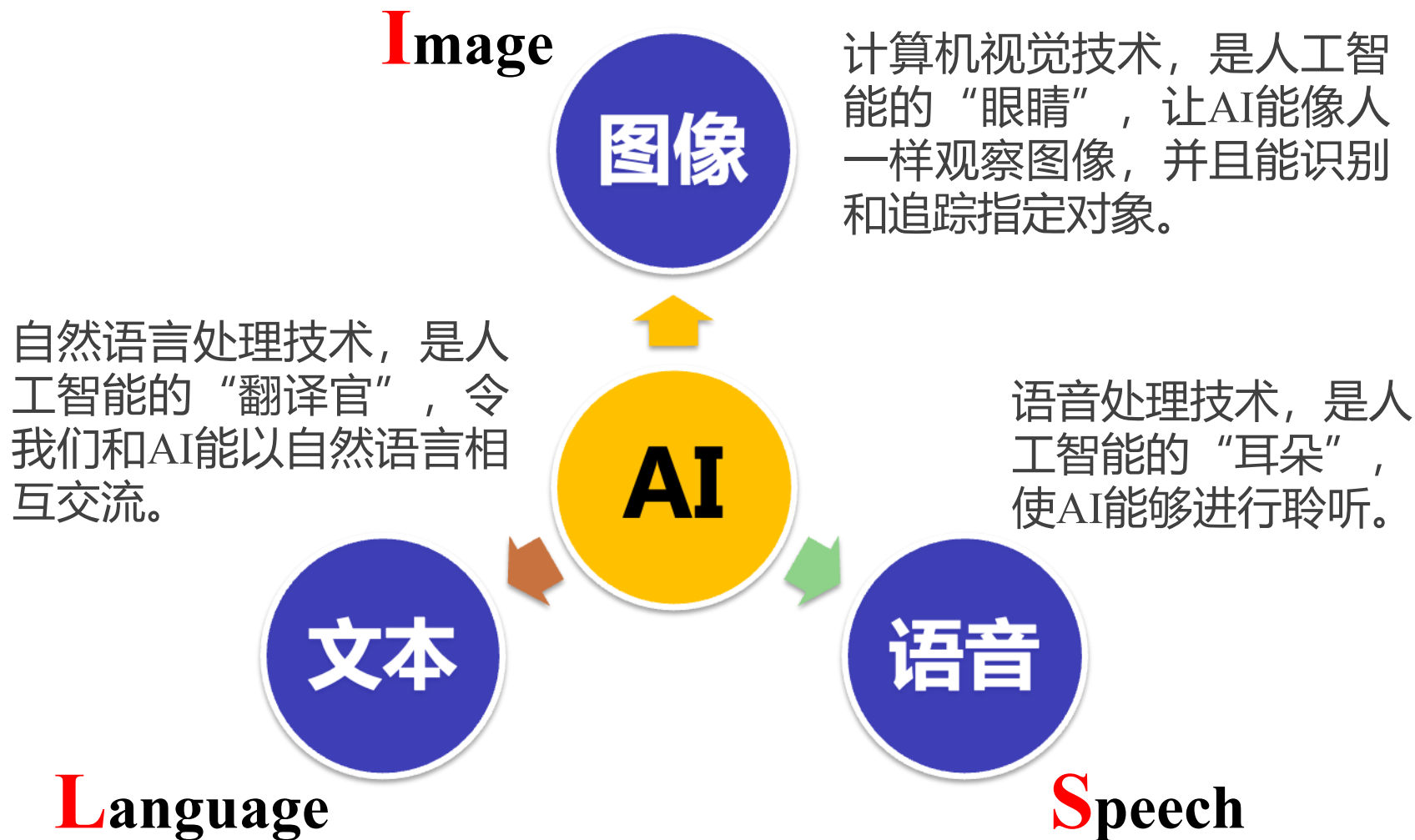


9/16/2026



数据采集： 自动生成数据

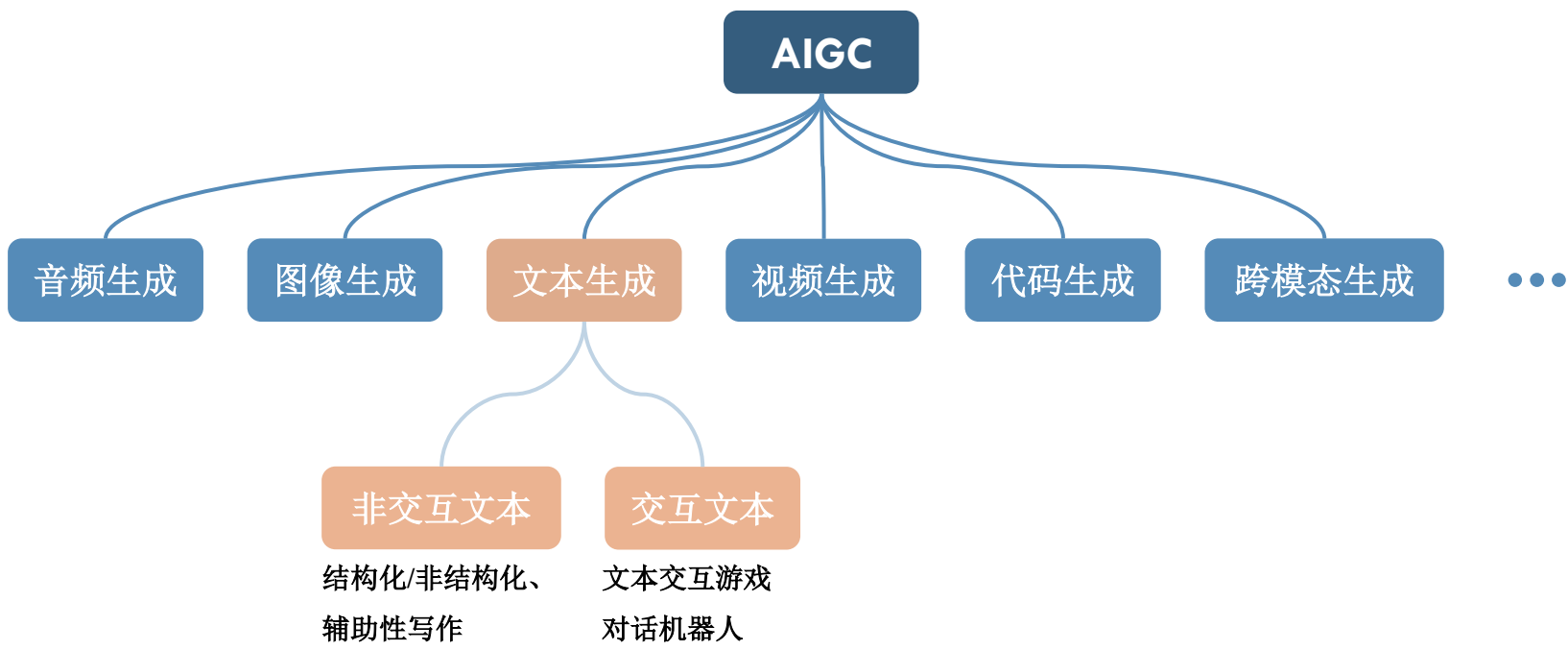
30





数据采集： 自动生成数据

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，可以在创意、表现力、迭代、传播、个性化等方面，充分发挥技术优势，打造新的数字内容生成与交互形态。





数据采集：自动生成数据

32

2022年11月30日，OpenAI 推出全新的对话式通用人工智能工具——**ChatGPT**。短短几天，其注册用户超过100万，引爆全网热议，成为历史上增长最快的消费者应用程序，掀起了人工智能领域大模型（参数量巨大的模型）技术巨浪。

ChatGPT 是一个聚焦于对话生成的大语言模型，其能够根据用户的文本描述，产生相应的智能回复。GPT通过学习大量网络已有文本数据，获得了像人类一样流畅对话的能力。





数据采集：自动生成数据

33

国内外大模型产业竞争目前呈现“激烈迅猛”之势



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-智谱清言



百川-百川大模型

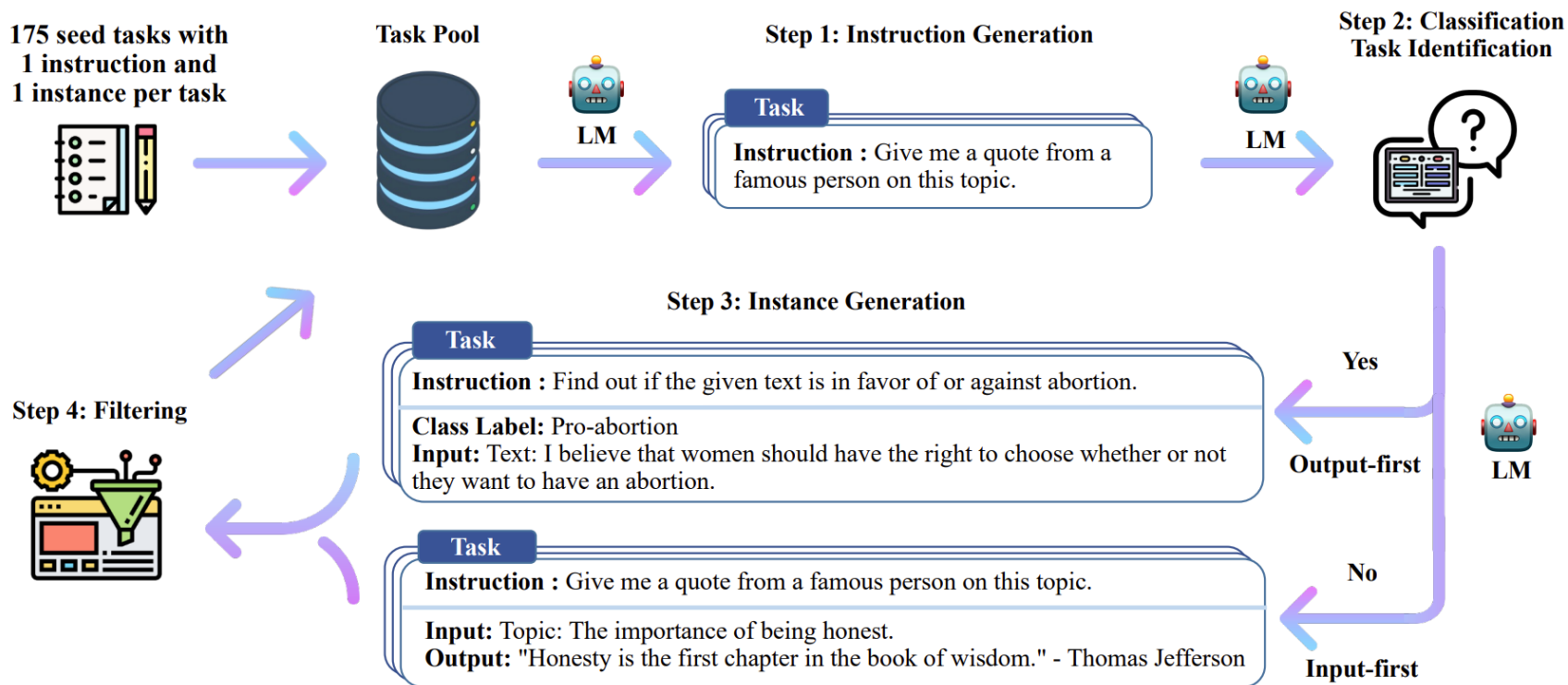
国内大模型从大模型多重能力构建，多模态语义理解等角度快速发展，在多个领域超过ChatGPT。



数据采集：自动生成数据

34

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

图片数据生成



Input Canny edge



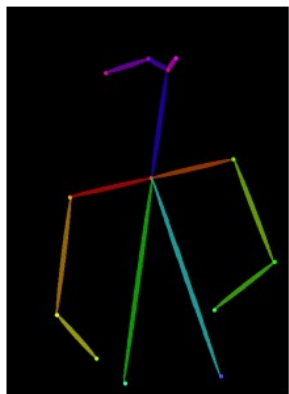
Default



“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



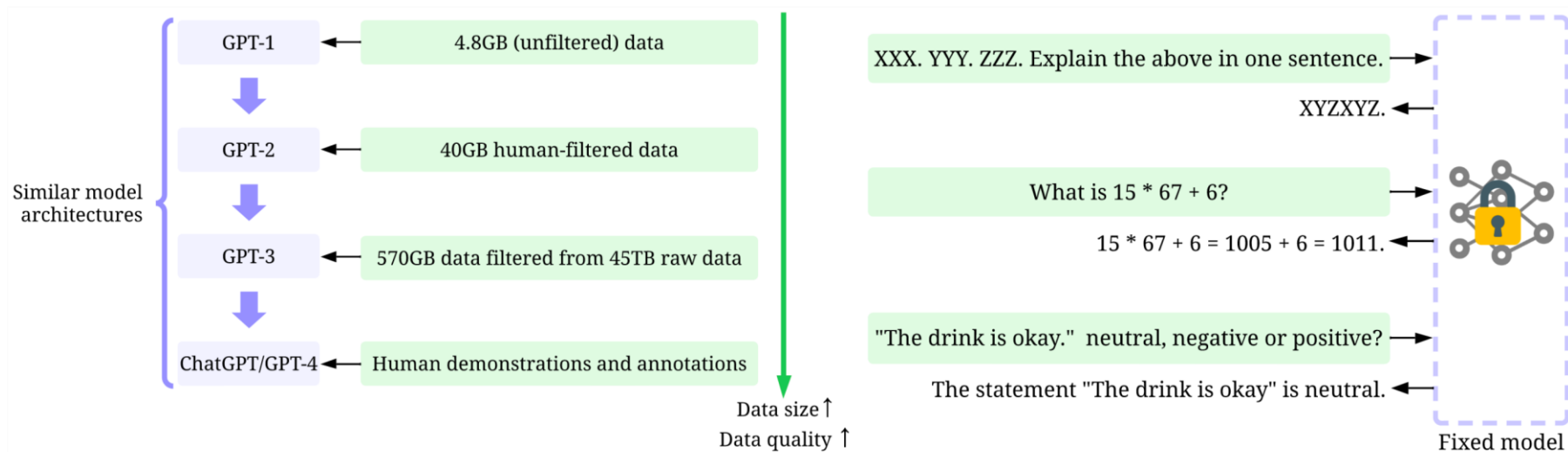
【提示词】逼真的特写视频展现了两艘海盗船在一杯咖啡中激烈交战，它们在咖啡液中航行。



数据科学基础

37

- 人工智能逐渐从以模型为中心过渡到以数据为中心
 - GPT成功依赖数据基石：** 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
 - 数据导向的模型应用：** 当模型足够强大，仅仅需要修改推理数据（提示工程）便可完成目标任务



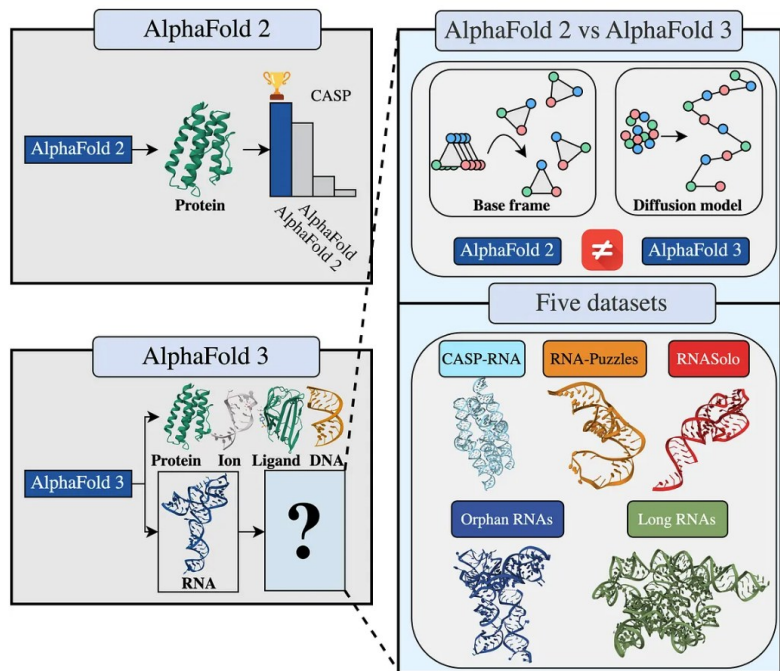


数据科学基础

38

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

3 GITHUB TRENDING #3 Repository Of The Day GITHUB TRENDING #3 Repository of the Day





数据科学基础

39

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

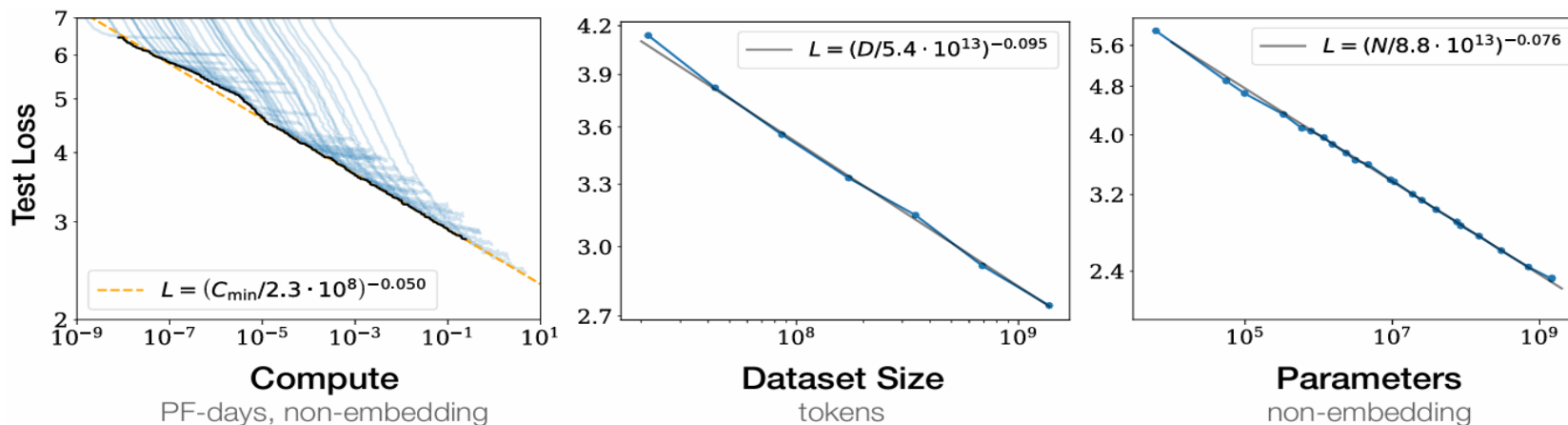


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not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

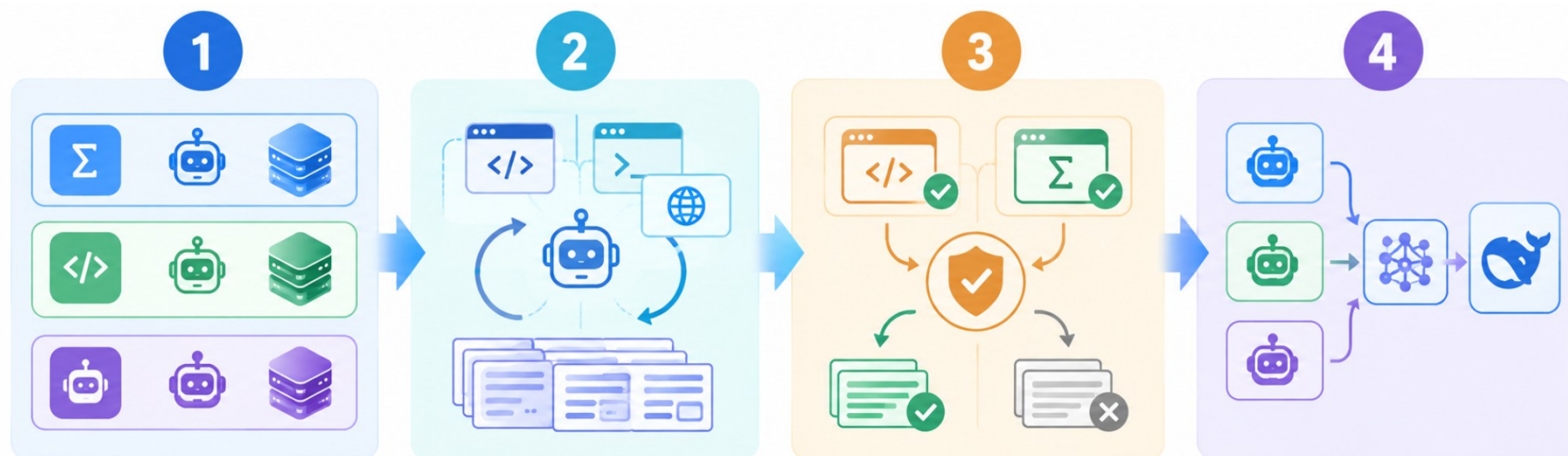
$$L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$$



数据科学基础

40

- 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





数据科学基础

41

□ 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



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

