

人工智能技术 如何在教育行业落地

苗广艺
学霸君技术VP

个人简介

- 毕业于中科院计算机专业，模式识别方向
- 先后就职于央视网、搜狐、YY、奇虎360
- 2014年加入学霸君，目前担任技术VP

目录

1 > 背景介绍

2 > 智慧题库

3 > 自动批改

4 > 自适应学习

5 > 总结

学霸君是一家面向K12的智能化教育公司



学习陪伴工具

- 线上流量入口
- 学习数据采集
- 学习交互社区



滴滴打老师

- 首创滴滴打车模式
- 首创数据工厂
- 首家将数码笔引入教学，实现线上直播互动



2C：线上1对1授课

- 数据驱动，实时测评
- “有序题组”实现教学重构
- 自适应题库替代题海战术



2B：智慧教育平台

- 主观题自动化批改解放老师
- 学校智能化数字化再造实现“Ai学 inside”
- 自适应题库替代题海战术

碎片化学习场景

课外补习场景

课内学习场景



A轮

500万美金

B轮

5000万美金

C轮

1亿美金

人工智能的几个层次

□ 基础层

- 云计算、芯片、TF等框架

□ 中间层

- 语音识别、人脸识别、图像识别

□ 应用层

- AI+行业、行业+AI



学霸君定位

The diagram shows the text 'AI+行业、行业+AI' from the application layer list. The phrase '行业+AI' is circled in blue. A blue line extends from the circle to a dark blue rectangular box on the right containing the text '学霸君定位'.

行业+AI 的关键点

□ 数据

- 大量实际真实场景的数据

□ 行业知识

- 需要多年积累，对具体业务非常熟悉
- 教育行业：教研知识，教学常识，学科知识

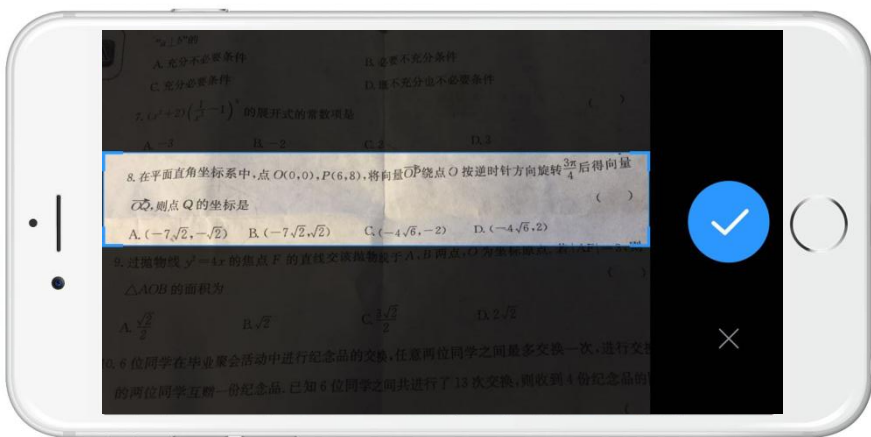
□ 工程与系统

- 最终产品是一个复杂系统
- 不存在“一招制胜”的算法

学霸君App：拍照搜题



学生：遇到难题
对准题目 拍照 框选范围



秒出答案：

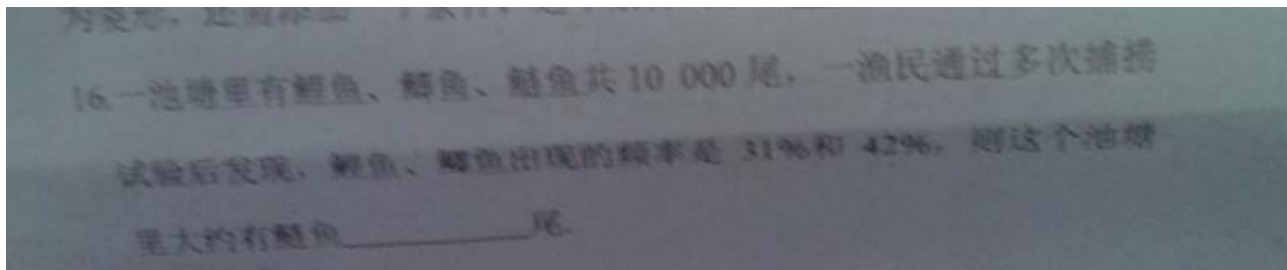
题干
解析
答案详解
点评
考点：

认识考点
考点例题
命题方向

题目识别OCR难点

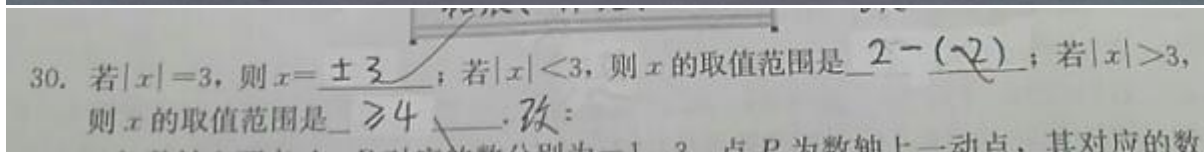
□ 形变

- 褶皱，扭曲
- 纸面透视严重



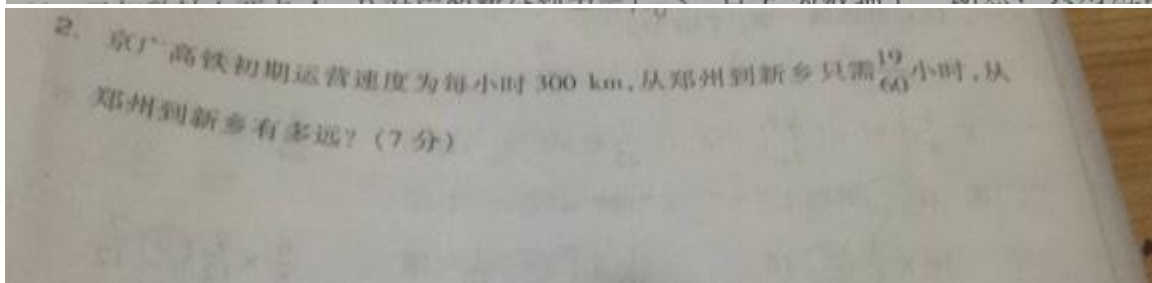
□ 模糊

- 抖动，失焦
- 摄像头差



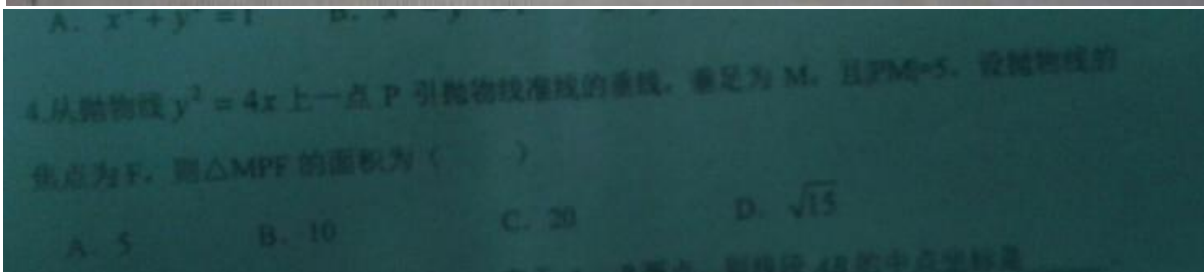
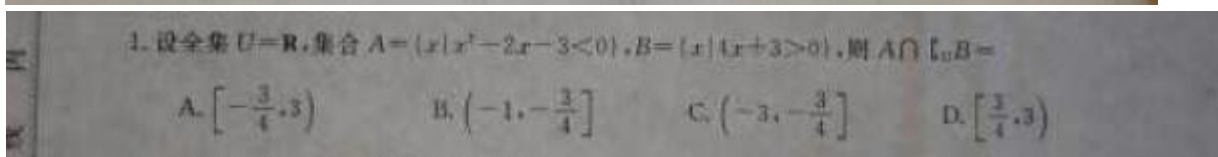
□ 版式复杂

- 插图，复杂排版
- 数学、化学公式



□ 干扰

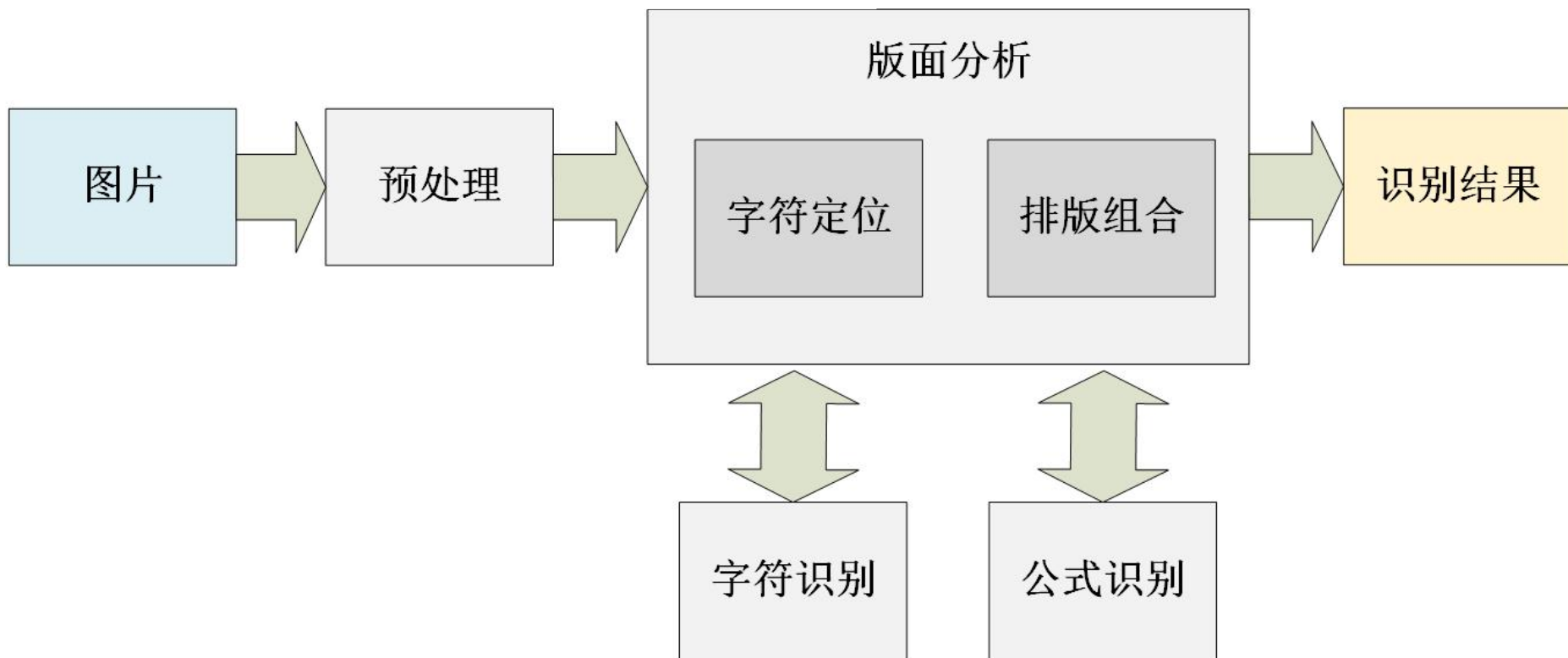
- 手写，划线
- 其他物体



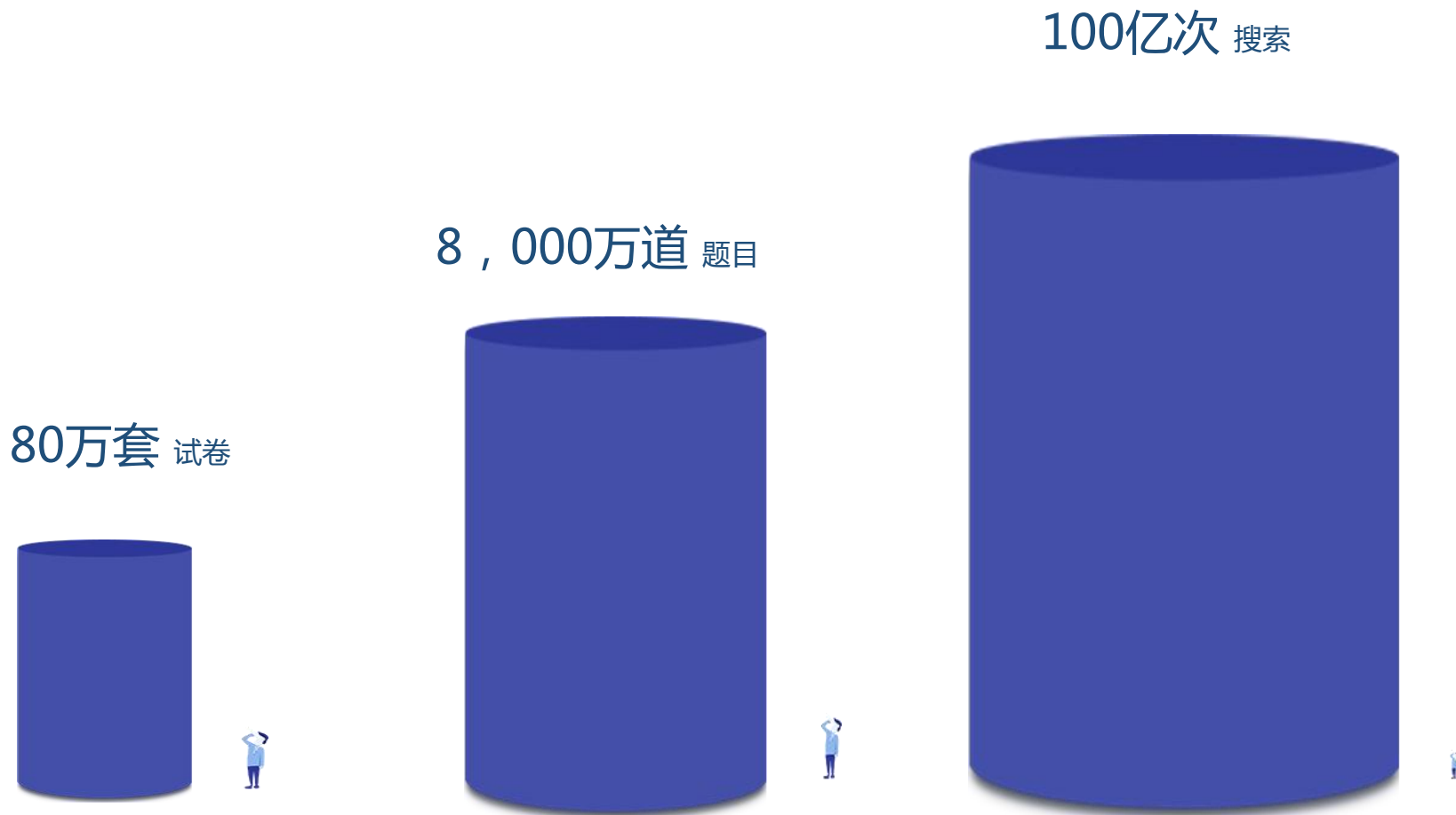
拍照题目OCR识别

□ 识别内核

- 中文：CNN
- 英文：LSTM

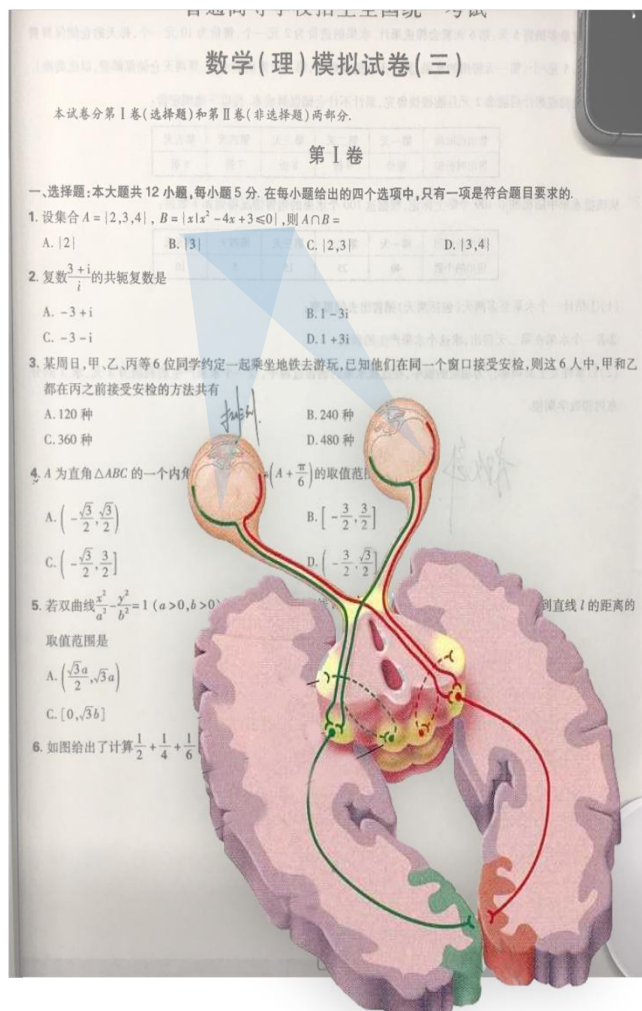


拍照搜题积累了海量题库



仅有数量是远远不够的，我们需要 **智慧题库**

迈向更智能数据认知模拟



通过算法来结构化题目



题目格式结构化

- 自动识别题目属性（如题号、分值、题目类型，选择题选项，填空题空格位置）

题号：10

分数：5

题型：选择题

内容：

已知向量 a , b 满足 $|a| = 2$, $|b| = 3$, $|a + b| = \sqrt{19}$, 向量 c 满足 $(a - 2c) \cdot (b - 3c) = 0$, 记 $|c|$ 的最大值为 m , 最小值为 n , 则 mn 的值为[question-placeholder]

[option] $\frac{1}{2}$

[option] $\frac{\sqrt{2}}{2}$

[option] 1

[option] $\sqrt{2}$

题号：18

分数：12

题型：解答题

内容：

已知 T_n 为数列 $\{a_n\}$ 的前 n 项积, 且满足 $2T_n = 1 - a_n (n \in N^*)$.

(1) 设 $b_n = \frac{1}{T_n}$, 证明数列 $\{b_n\}$ 是等差数列, 并求数列 $\{b_n\}$ 的通项公式;

(2) 求数列 $\{\frac{b_n}{2^n} + 2^n\}$ 的前 n 项和 s_n

数学公式LaTeX化

若函数 $f(x) = \frac{3a-1}{\sqrt{1-ax}}$ 在区间 $[0,1]$ 上单调递增

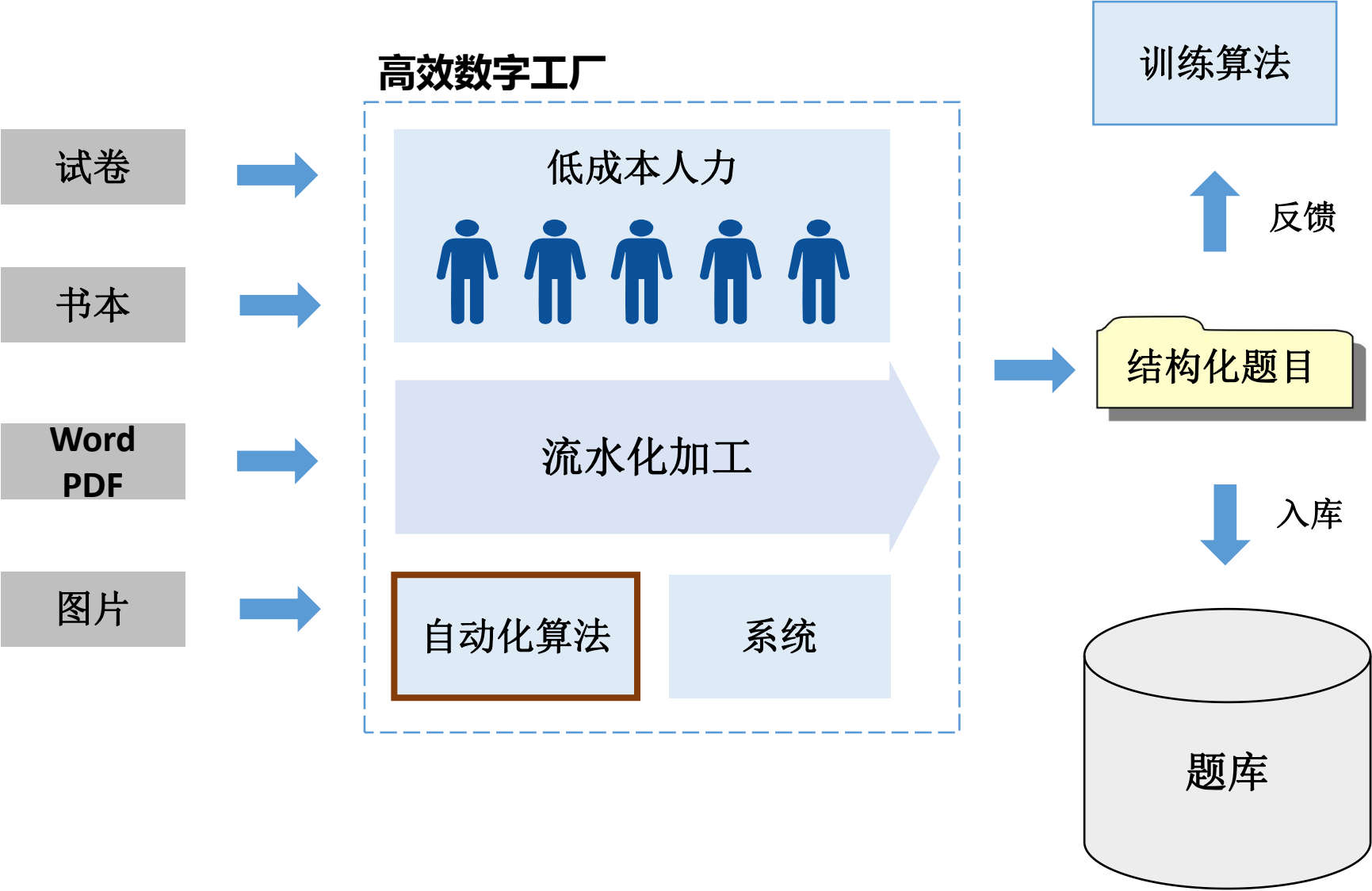
识别算法

若函数 $f(x) = \frac{3a-1}{\sqrt{1-ax}}$ 在区间
 $[0,1]$ 上单调递增

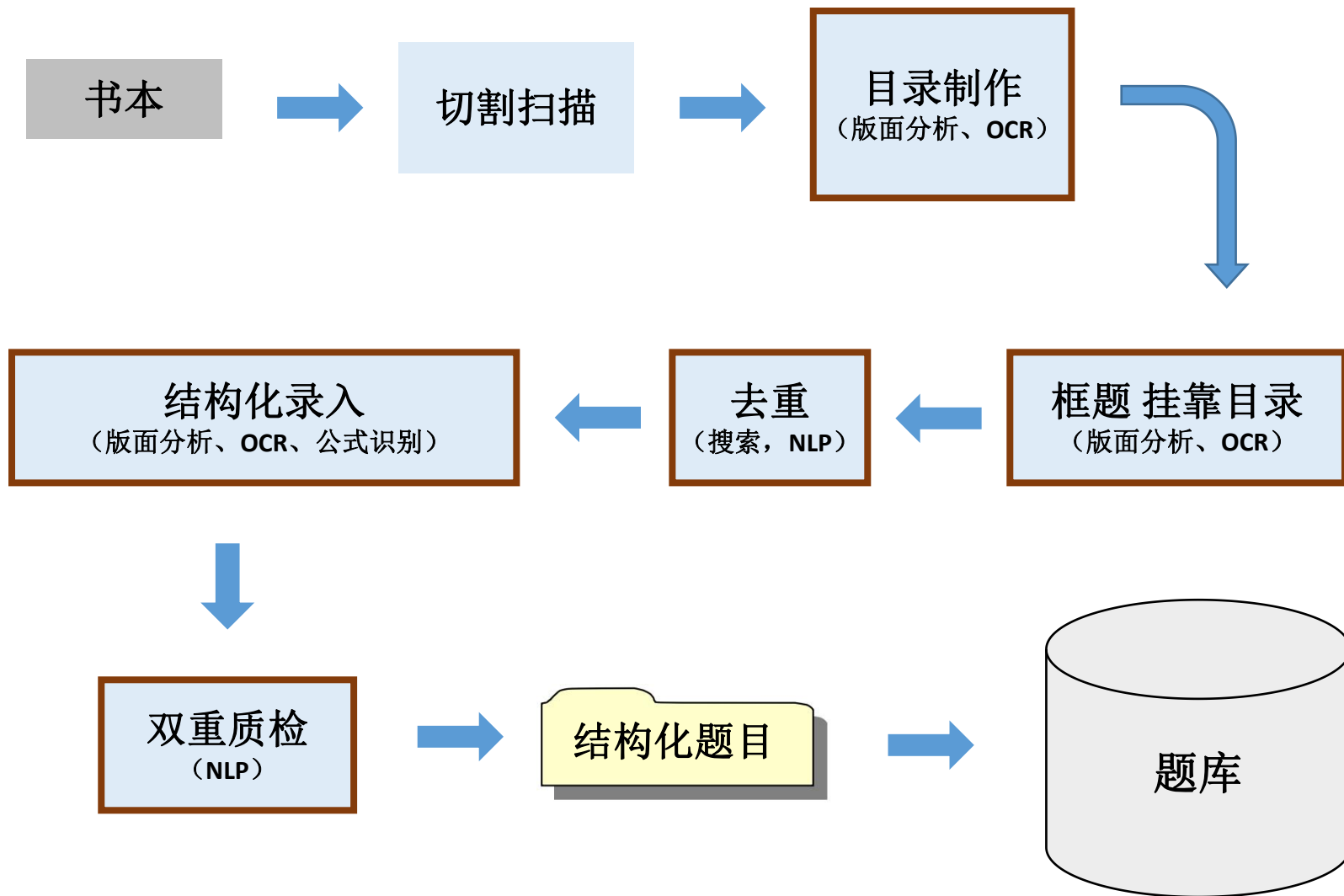
渲染算法

若函数 $f(x) = \frac{3a-1}{\sqrt{1-ax}}$ 在区间 $[0,1]$ 上单调递增

流水化生产题目



通过算法提高生产效率



题目知识点分类

难点：短文本、多层次、需要语义/公式层面信息

已知函数 $f(x)$ 是定义在 $[-2, 2]$ 上的增函数, 且 $f(1-m) < f(m)$, 求实数 m 的取值范围_____.

已知函数 $f(x) = 8 + 2x - x^2$, $g(x) = f(2 - x^2)$, 试求 $g(x)$ 的单调区间.

已知函数 $f(x) = \frac{x^2 + 2x + a}{x}$, $x \in [1, +\infty)$.

(1) 当 $a = \frac{1}{2}$ 时, 判断并证明 $f(x)$ 的单调性;

(2) 当 $a = -1$ 时, 求函数 $f(x)$ 的最小值.

▼ 函数

▼ 函数概念与性质

► 函数的概念与表示

▼ 函数的性质

函数单调性的判定与证明

单调性与函数的最值

分段函数的单调性

复合函数的单调性

抽象函数的单调性

利用函数单调性比较大小

利用函数单调性解方程/不等式

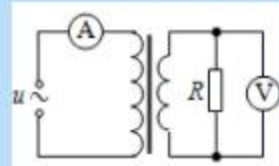
函数奇偶性的判定与证明

奇偶函数的性质

利用奇偶性求解析式

结构化知识点

如图所示, 理想变压器的原线圈接在 $u = 220\sqrt{2} \sin \pi t(\text{V})$ 的交流电源上, 副线圈接有 $R = 55 \Omega$ 的负载电阻, 原、副线圈匝数之比为 $2:1$, 电流表、电压表均为理想电表. 下列说法正确的是 ()



原副线圈中电压、电流与线圈匝数关系
电磁学 > 交变电流 > 变压器

A. 原线圈的输入功率为 $u = 220\sqrt{2} \text{ W}$

A

交变电流的有效值

电磁学 > 交变电流 > 交变电流的四值

原副线圈中电压、电流与线圈匝数关系

电磁学 > 交变电流 > 变压器

B. 电流表的读数为 1 A

B

交变电流的有效值

电磁学 > 交变电流 > 交变电流的四值

原副线圈中电压、电流与线圈匝数关系

电磁学 > 交变电流 > 变压器

C. 电压表的读数为 $u = 110\sqrt{2} \text{ V}$

C

交变电流的有效值

电磁学 > 交变电流 > 交变电流的四值

原副线圈中电压、电流与线圈匝数关系

电磁学 > 交变电流 > 变压器

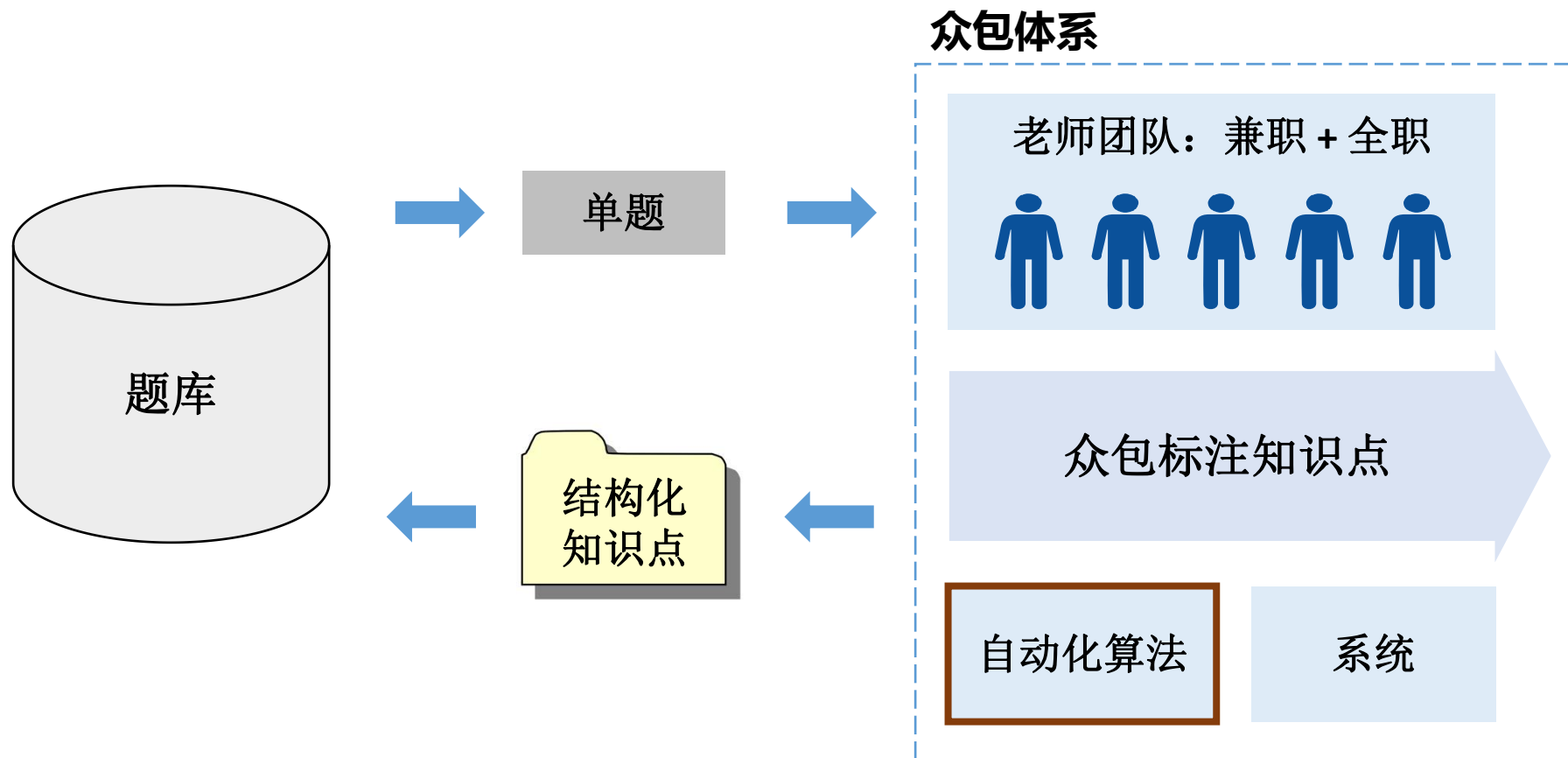
D. 副线圈输出交流电的周期为 50 s

D

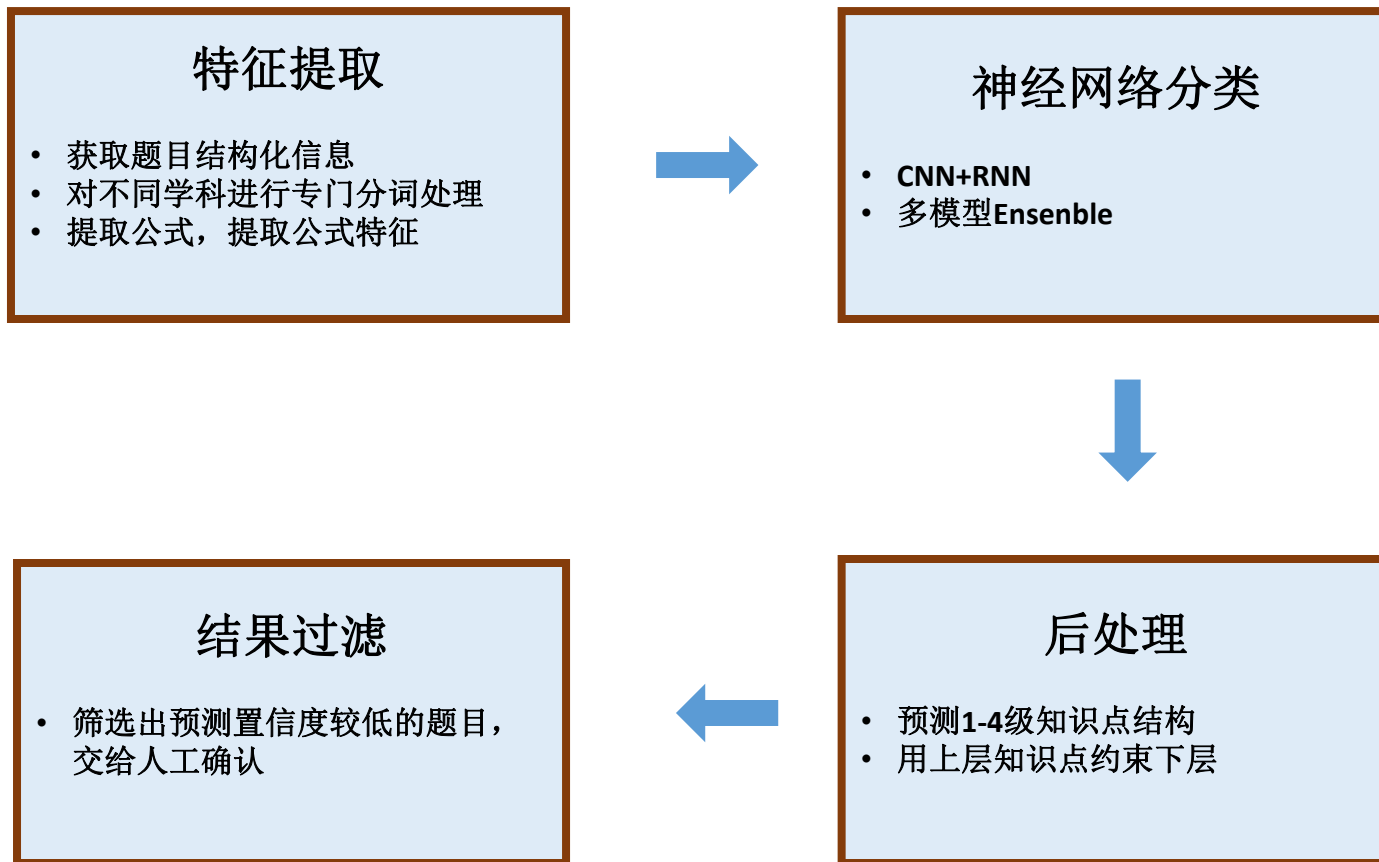
正弦式交变电流的表达式及推导

电磁学 > 交变电流 > 交变电流的产生以及描述

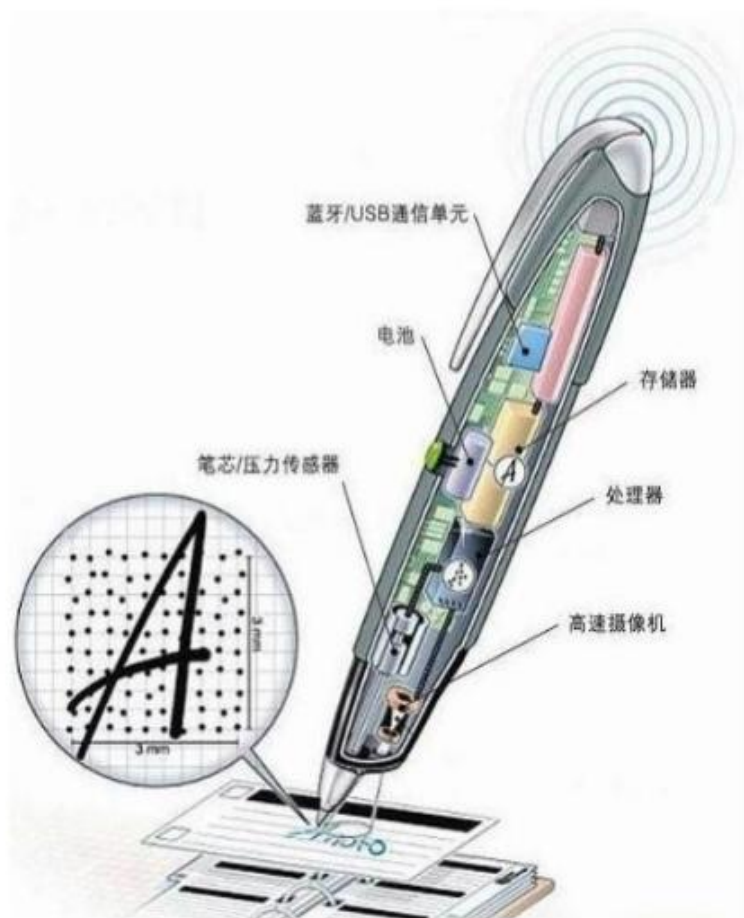
通过算法提高生产效率



知识点分类算法



手写笔记同传技术



使用场景



红外线

笔芯

CMOS摄像机

学校里常规使用



自动批改

学校现状



老师每天至少花费**2个小时**批改作业

未来情况

学生做题数据全量电子化

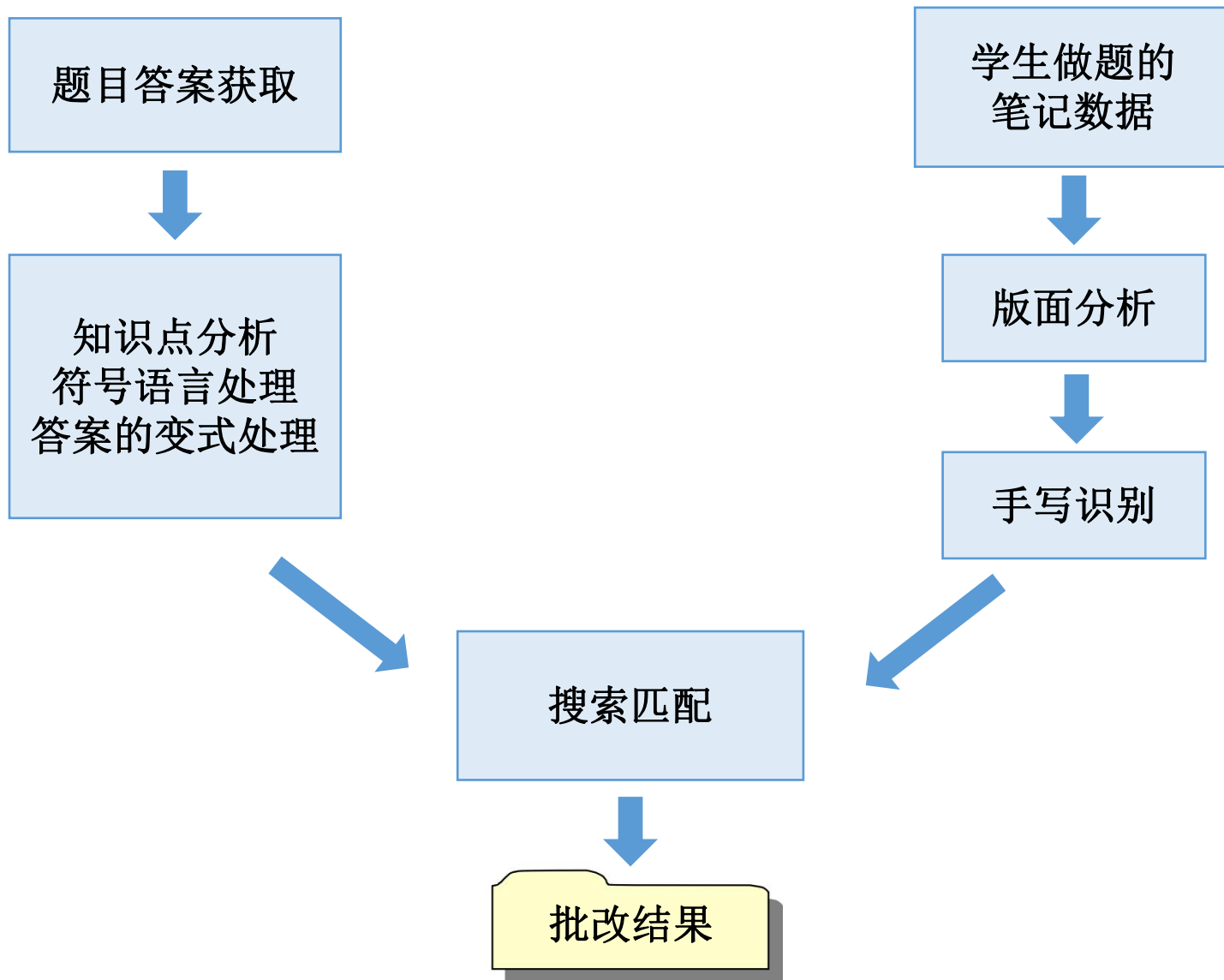


系统自动批改作业



老师随时查看作业报告

自动批改算法架构



题目与学生笔迹

题干：

函数 $y = \sqrt{x(x-1)} + \sqrt{x}$ 的定义域为_____。

参考答案： $\{x|x \geq 1\} \cup \{0\}$

学生手写笔迹：

解 要使函数有意义

$$\begin{cases} x(x-1) \geq 0 \\ x \geq 0 \end{cases}$$

得

$$\begin{cases} x \geq 1, x \leq 0 \\ x \geq 0 \end{cases}$$

\therefore 定义域为 $x \in [1, +\infty) \cup \{0\}$

学生笔迹版面分析

公式定位



文本行提取

解 要使函数有意义

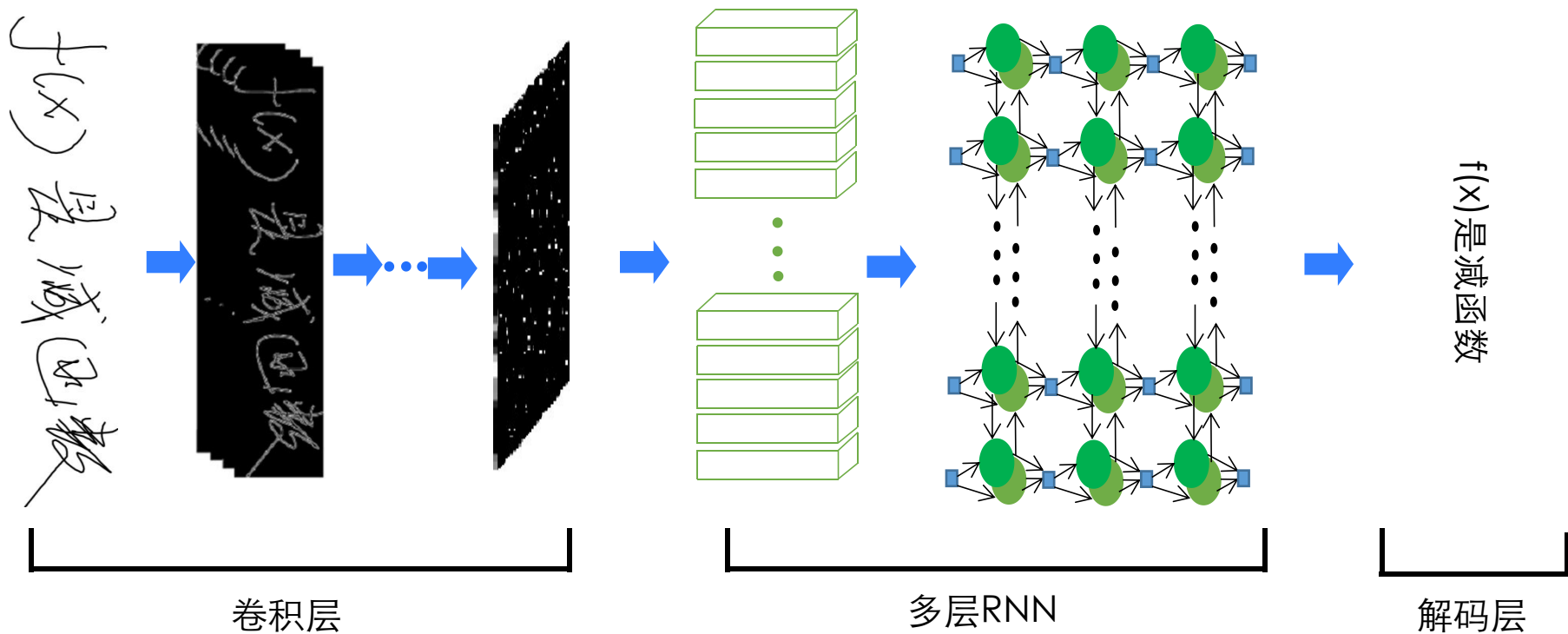
$$\begin{cases} x(x-1) \geq 0 \\ x \geq 0 \end{cases}$$

得

$$\begin{cases} x \geq 1, x \leq 0 \\ x \geq 0 \end{cases}$$

∴ 定义域为 $x \in [1, +\infty) \cup \{0\}$

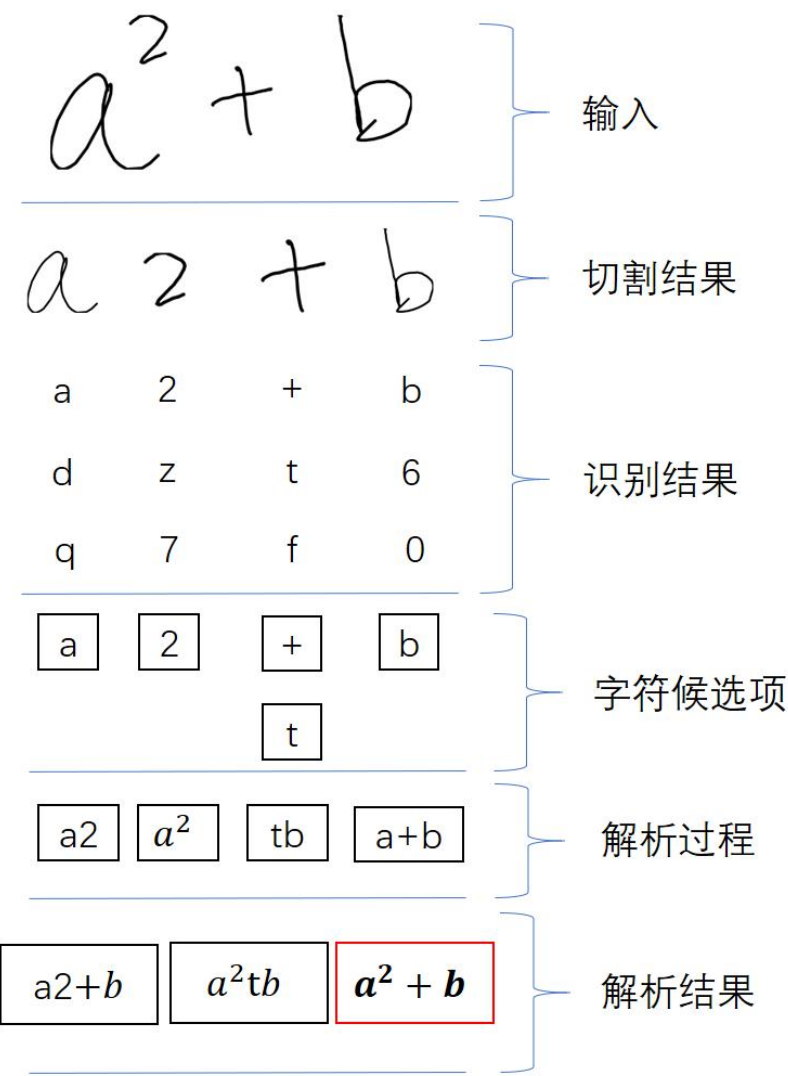
端到端识别 (CNN + LSTM + CTC)



基于2D空间结构识别

算法主要步骤:

- 1. 字符切分
- 2. 字符识别
- 3. 公式结构解析
- 4. 后处理



基于搜索匹配的批改

参考答案:

$$\{x \mid x \geq 1\} \cup \{0\}$$

数学符号
语言处理

$$x \in \{0\} \cup [1, +\infty)$$

$$\{x \mid 1 \leq x\} \cup \{0\}$$

$$x \in [1, +\infty) \cup \{0\}$$

$$\{x \mid x \geq 1\} \cup \{0\}$$

...

匹配

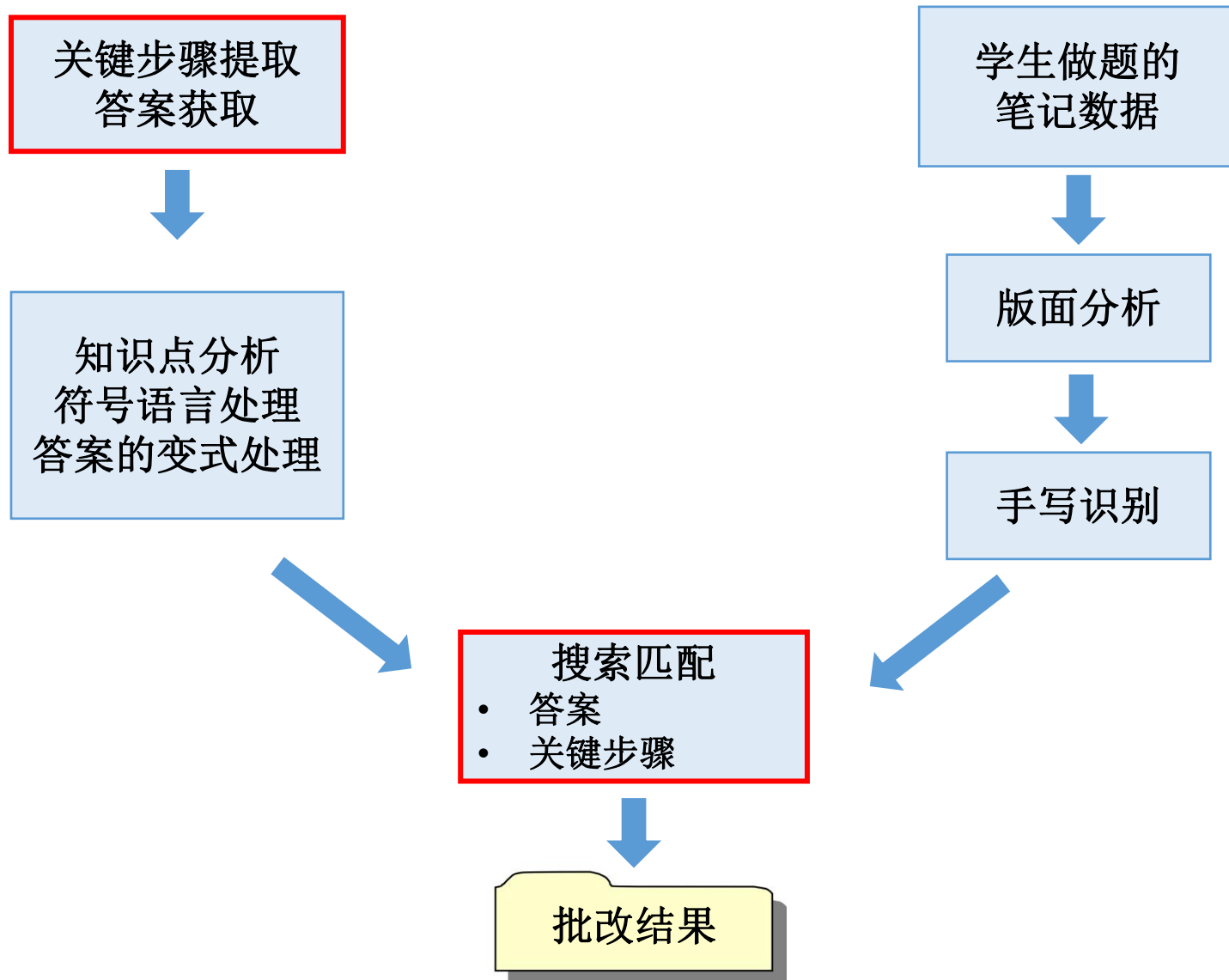
$$', \text{定义域为 } x \in [1, +\infty) \cup \{0\}$$

$$', \text{定义域为 } x \in [1, +\infty) \cup \{0\}$$

批改结果:




解答题的自动批改：给步骤分



多个关键步骤的批改

N行手写数据:

解: 由题意可得数轴为



则 $a=1$

$\therefore A \cap B = \emptyset$

$\therefore a-1 \neq 0 \therefore a-1 \geq 1$

或 $a \geq 2$

$2a+1 \leq 0$

$a \leq -\frac{1}{2}$

若 $a+1 > 2a+1$

则 $a < -2$

综上所述 $\{a \mid a \leq -\frac{1}{2} \text{ 或 } a \geq 2\}$

M个关键步骤:

关键步骤1

$$a \geq 2$$

关键步骤2

$$a \leq \frac{1}{2}$$

最终得分 =

答案分数 + 步骤1分数 + 步骤2分数

作业自动批改



缩小学习闭环时间



由48小时缩短至8小时

学情分析

作业概况

习题数量：24 题

单选题：18 多选题：2 填空题：2 实验题：0 推断题：2 计算题：0

未提交学生名单(1/41)

章梓贞

未批改学生名单(0/41)

没有未批改学生了

86.5%

平均正确率

平均作答时长:60分钟24秒

知识点掌握程度

二、元素的性质与原子结构

83%

元素非金属性的判断
1,10 题

80%

元素性质与原子结构和周期
表位置的关系
6,21,22 题

63%

碱金属元素单质与氧气、水
的反应
14,24 题

三、核素

92%

质量数、质子数、中子数和
相对原子质量
2,3 题

88%

元素、核素、同位素和同素
异形体
5 题

三、元素周期表和元素周期律的应用

96%

主族元素化合价与周期表中
的位置关系
8,16 题

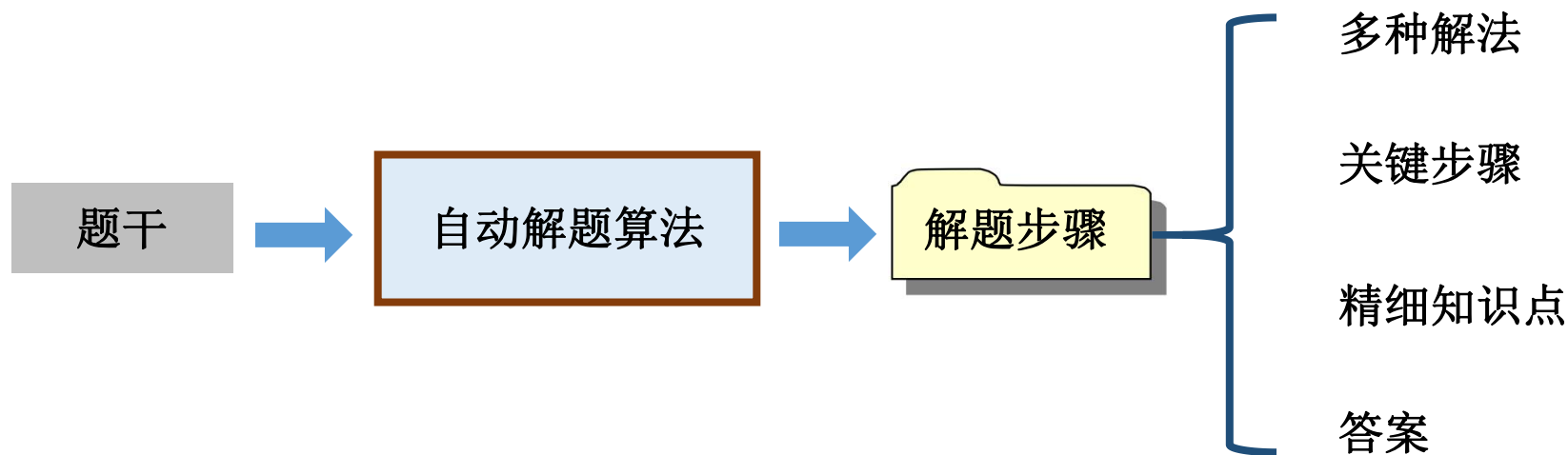
82%

周期表中的规律三角
23 题

解答题批改引来的问题

几个问题：

- 关键步骤如何提取？
- 多种解法如何处理？

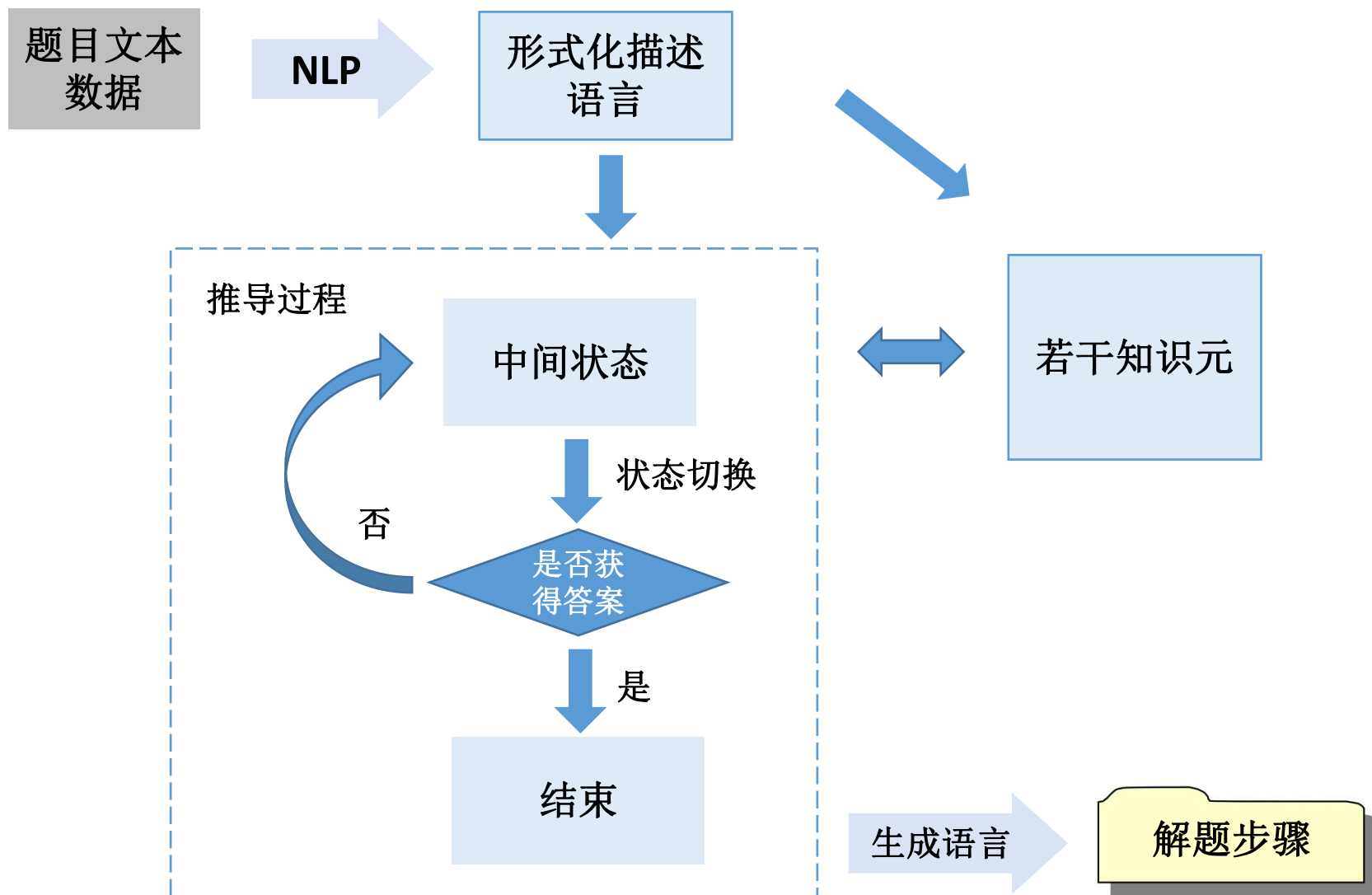




2017年6月7号，学霸君高考机器人Aidam与历年状元同场竞技，挑战2017年数学高考试卷。Aidam仅用10分钟就完成了所有题目，并取得了134分的成绩，现实中挑战高考状元不落下风。目前Aidam已全面接入AI学智慧教育平台，帮助老师批改日常作业。

高考机器人Demo

解题算法架构





知识图谱

新函数的单调性利用函数的单调性解不等式：4

函数 $f(x) = x^3 + ax^2 + 3x - 9$, 已知 $f(x)$ 在 $x =$ 处取得极值, 则 $a = ()$;

函数 $f(x) = x^3 + ax^2 + 3x - 9$, 已知 $f(x)$ 在 $x =$ 处取得极值, 则 $a = ()$;

函数 $f(x) = x^3 + ax^2 + 3x - 9$, 已知 $f(x)$ 在 $x =$ 处取得极值, 则 a 等于();

已知 $f(x) = x^3 + 3ax^2 + bx + a^2$, 在 $x = -1$ 时有值0, 求常数 a, b 的值.;

已知关于 x 的函数 $f(x) = -\frac{1}{3}x^3 + bx^2 + cx + b$

若函数 $f(x)$ 在 $x = 1$ 处取得极值 $-\frac{4}{3}$, 则 $b = ()$,
 $= ()$;

已知函数 $f(x) = \frac{1}{3}a^2x^3 + \frac{1}{2}ax^2 - 2x + 1$ 在 $x =$
2处取得极值, 则实数 a 的值为();

已知函数 $f(x) = x^3 + ax^2 + 3x - 9$ 在 $x = -3$ 处
得极值, 则 $a = ()$;

已知函数 $y = ax^3 - 15x^2 + 36x - 24$ 在 $x = 3$ 处

个性化学习

IRT理论 : Item Response Theory

广泛应用于心理和教育测量领域

原理:

- 通过建模学生做题数据，量化学生能力特征和题目特征
 - 学生能力值 θ_m
 - 题目难度 β
 - 题目区分度 α
 - 题目猜测度 χ
- 预测学生答对概率

$$P(x = 1|\theta, \beta, \alpha, \chi) = \chi + (1 - \chi) \frac{e^{\alpha(\theta - \beta)}}{1 + e^{\alpha(\theta - \beta)}}$$

应用：学生能力评估

答题数据

学生ID	题目ID	学科ID	知识点	答题情况
1000	2000	数学	集合关系	✓
1000	2001	数学	集合关系	✓
1001	2002	英语	定语从句	✗
1001	2003	英语	感叹句	✓
...



学生能力数据

学生ID	学科	知识点	能力值
1000	数学	集合关系	1.12
1001	英语	定语从句	-0.23
1001	英语	感叹句	0.87
...

用以精准评估

题目属性数据

题目ID	难度	区分度	答对概率
2000	0.67	0.4	23%
2001	0.54	0.56	5%
2002	-1.2	1.2	4.2%
2003	-0.12	0.76	26%
...

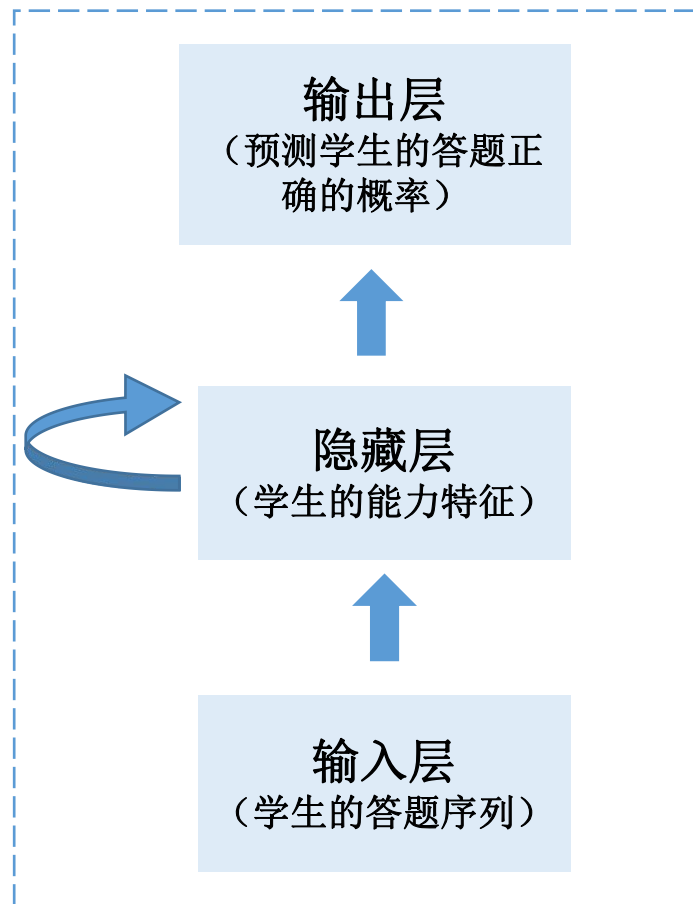
用以完善教学数据

利用RNN训练学生能力模型

Deep Knowledge Tracing (DKT)

核心：

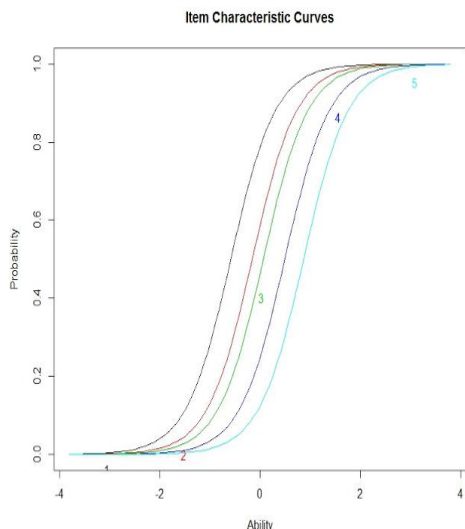
- 以RNN为基本架构
- 以学生的能力作为隐藏的特征
- 预测学生答对概率，或者结合IRT模型，将训练得到的能力特征作为输入



自适应学习框架

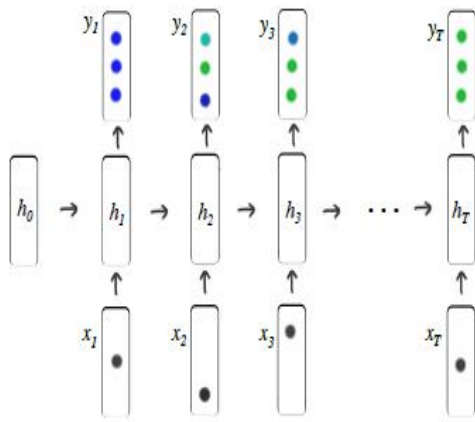
传统Item Response Theory (IRT)

拟合学生能力及题目难度等属性



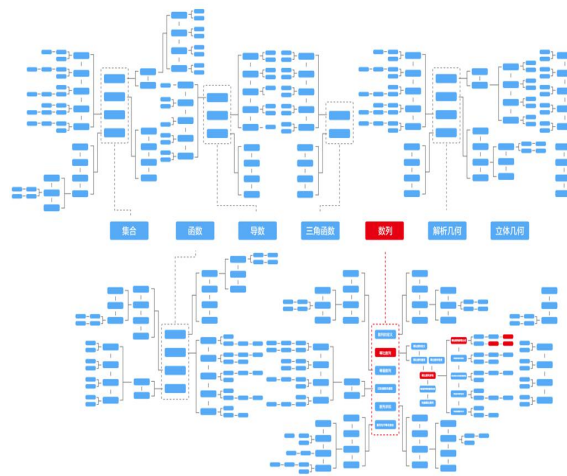
深度学习RNN

与IRT模型组合，提升预测学生答题对错的精度



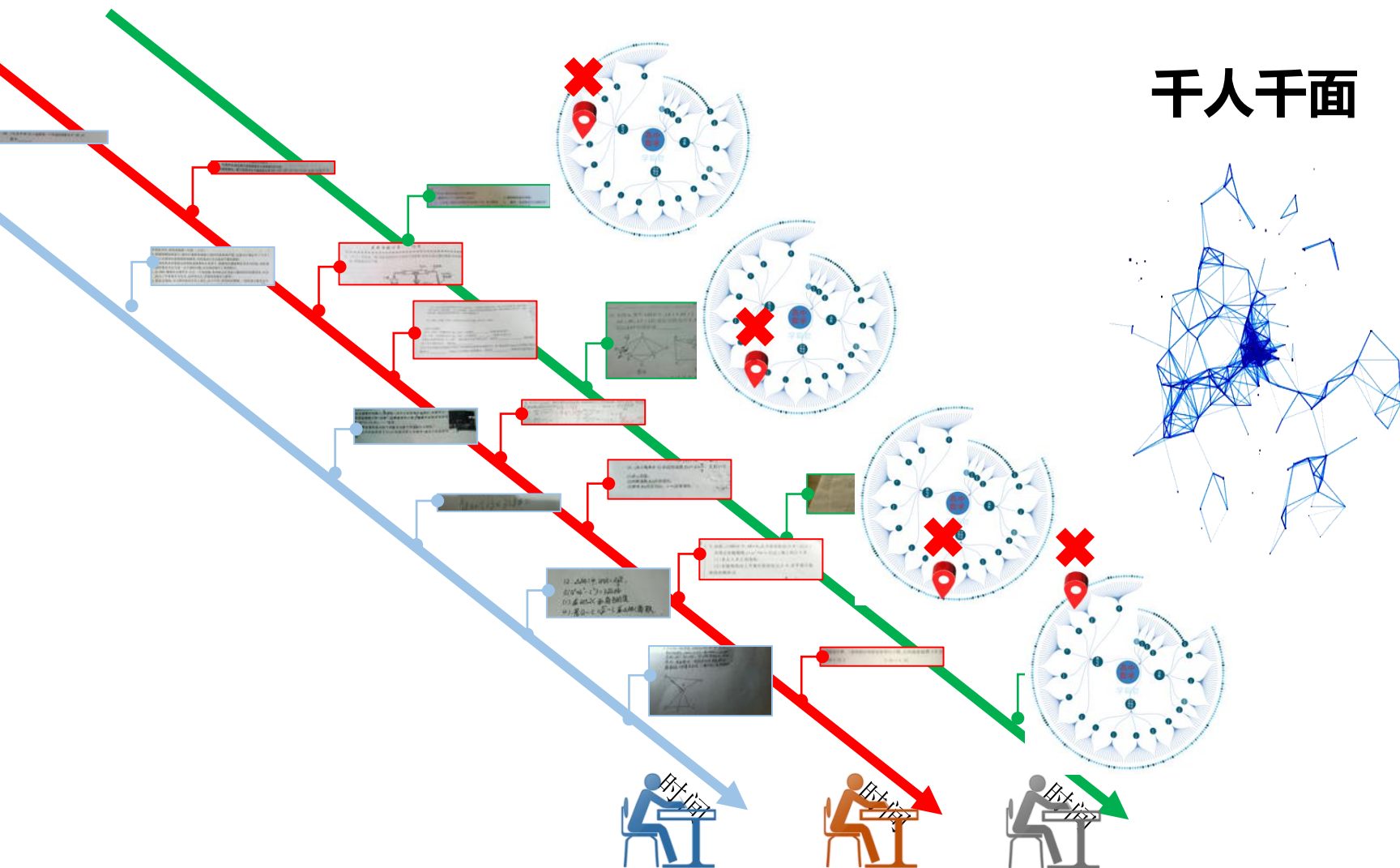
知识图谱

跨知识点推题、规划学习路径

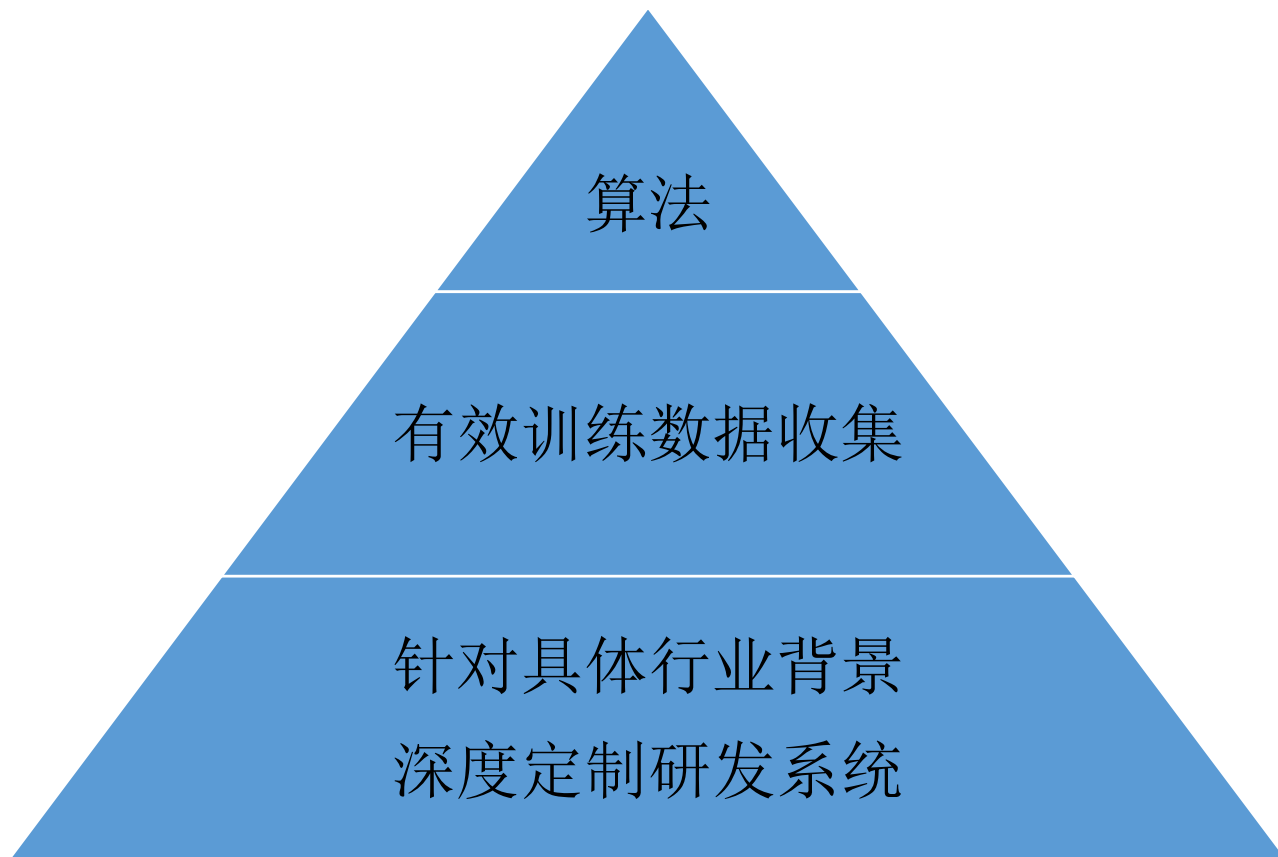


自适应学习模型

千人千面

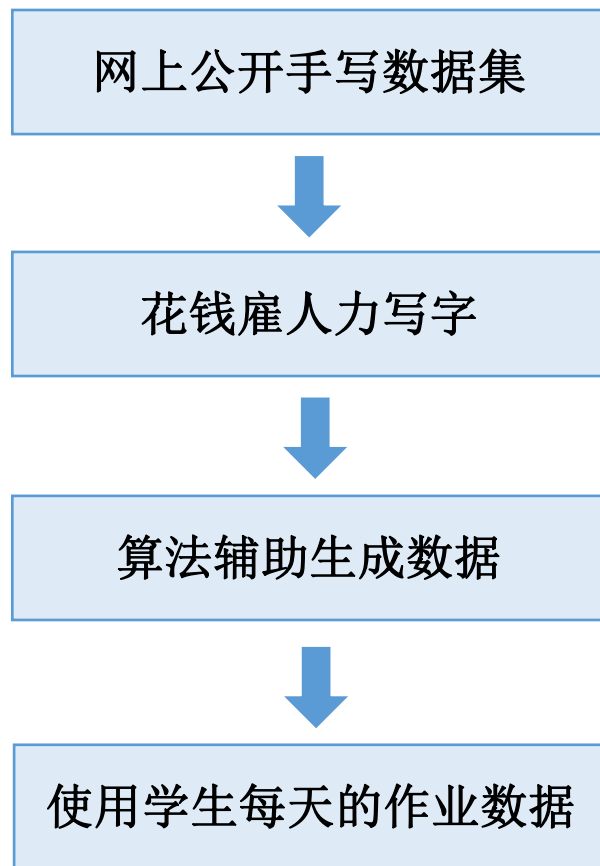


对于精力的消耗



数据收集

手写数据为例：



深度定制的算法与系统

自动批改

自适应学习

自动解题
机器人

手写公式
识别

自然语言
处理

逻辑推理

基于中考高考
的知识图谱

结构化
知识点

四级教研知
识点体系

结构化
题目格式

LaTeX
公式格式

基于点阵笔
的笔记数据

带来的困扰

□ 算法不通用

- 各种场景都需要定制，开发量很大
- 业务变动，不可复用，基本要重新开发

□ 对个人依赖度较大

- 只有实际开发者最懂，别人接手时间很长

□ 对人才复合要求较高

- 懂算法，懂业务，懂教研
- 有较好的系统架构和编码能力

谢谢！

苗广艺 学霸君技术VP
微信号：miaoguangyi