ArchSummit全球架构师峰会北京站2015

如何提供让人耳目一新的在线答疑服务的核心技术？

学霸君App

陈锐锋
(ruifeng.chen@wenba100.com)

Geekbang.
极客邦科技

整合全球最优质学习资源，帮助技术人和企业成长

Growing Technicians, Growing Companies

InfoQ
专注中高端技术人员的
技术媒体

EGO EXTRA GEEKS' ORGANIZATION NETWORKS
高端技术人员
学习型社交网络

StuQ
实践驱动的
IT职业学习和服务平台

GiT GEEKBANG INTERNATIONAL TRAINING 极客邦培训
一线专家驱动的
企业培训服务

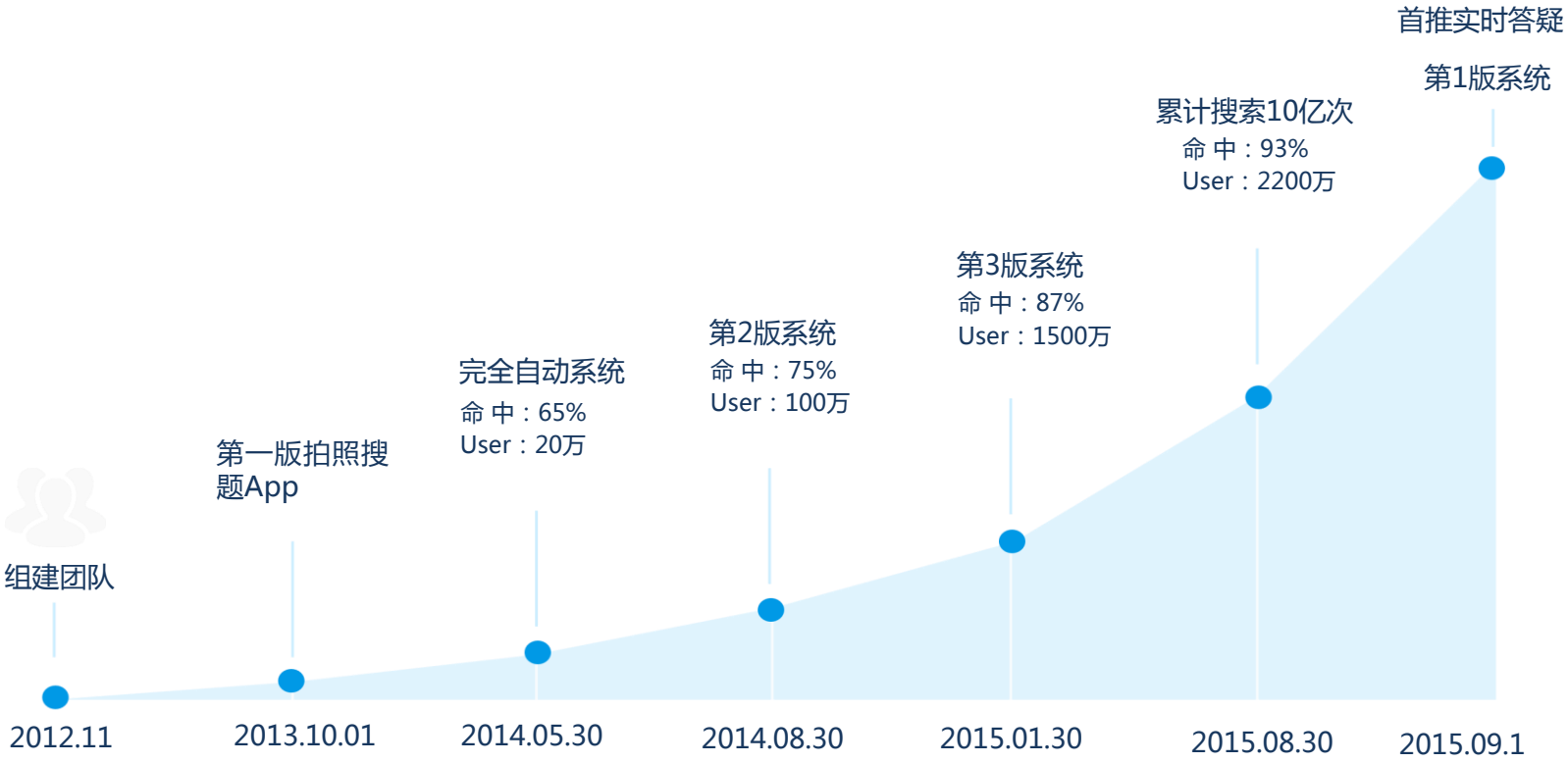
地址：北京市朝阳区洛娃大厦C座6层1607室　　网址：www.geekbang.org

假如记忆可以移植…

目录

1.1.学霸君的创业历程——创新、再创新

- 学霸君简史



- 现状 -

[学生]
不敢问
不想问
不会问

[老师]
薪酬低
空余时间多
压力大

[家长]
望子成龙
不计成本
无力辅导

- 我们的解决方案 -

实时答疑

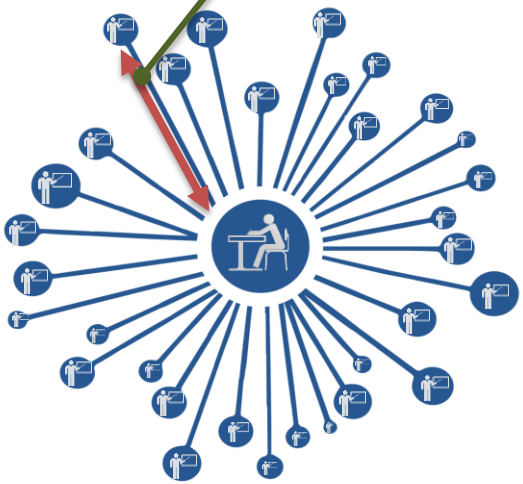
请求答疑

精确解答

掌握情况

距离=千山万水

距离=10秒



学霸君老师答疑——难题即刻得到讲解

老师：拿题就讲

学生：不懂就问，问懂为止

真正做到今日难题今日毕

目录

2.1. 发动一场大规模的垂直领域数据采集

- 2012~2013年我们思考的核心问题，如何获得最学生个体学习信息？
  - 举个例子，随便给一个学生，如何知道他在学什么？

- 数十次头脑风暴后，学霸君决定把注意力落于他日常接触的书、试卷。

- 技术孵化早期典型心理——在盲目乐观与盲目悲观中游荡

文字提取是个坎

pulled the boat into open water. The pair panicked and tried to row back to shore. But they were no match for it and the boat was out of control.

Tim knew it would soon be swallowed by the waves.

"Everything went quiet in my head," Tim recalls (回忆). "I was trying to figure out how to swim to the boys in a straight line."

Tim took off his clothes and jumped into the water. Every 500 yards or so, he raised his head to judge his progress. "At one point, I considered turning back," he says. "I wondered if I was putting my life at risk." After 30 minutes of struggling, he was close enough to yell to the boys, "Take down the umbrella!"

Christian made much effort to take down the umbrella. Then Tim was able to catch up and climb aboard the boat. He took over rowing, but the waves were almost too strong for him.

"Let's aim for the pier (码头)," Jack said. Tim turned the boat toward it. Soon afterward, waves crashed over the boat, and it began to sink. "Can you guys swim?" he cried. "A little bit," the boys said.

Once they were in the water, Tim decided it would be safer and faster for him to pull the boys toward the pier. Christian and Jack were wearing life jackets and floated on their backs. Tim swam toward land as water washed over the boys' faces.

"Are we almost there?" they asked again and again. "Yes," Tim told them each time.

After 30 minutes, they touched the pier.

35. Why did the two boys go to the sea?
A. To go boat rowing.
B. To get back their football.
C. To swim in the open water.
D. To test the umbrella as a sail.

36. What does "it" in Paragraph 2 refer to?
A. The boat. B. The water.
C. The boat. D. The wind.

37. Why did Tim raise his head regularly?
A. To take in enough fresh air.
B. To consider turning back or not.
C. To check his distance from the boys.
D. To ask the boys to take down the umbrella.

38. How did the two boys finally reach the pier?
A. They were dragged to the pier by Tim.
B. They swam to the pier all by themselves.
C. They were washed to the pier by the waves.
D. They were carried to the pier by Tim on his back.

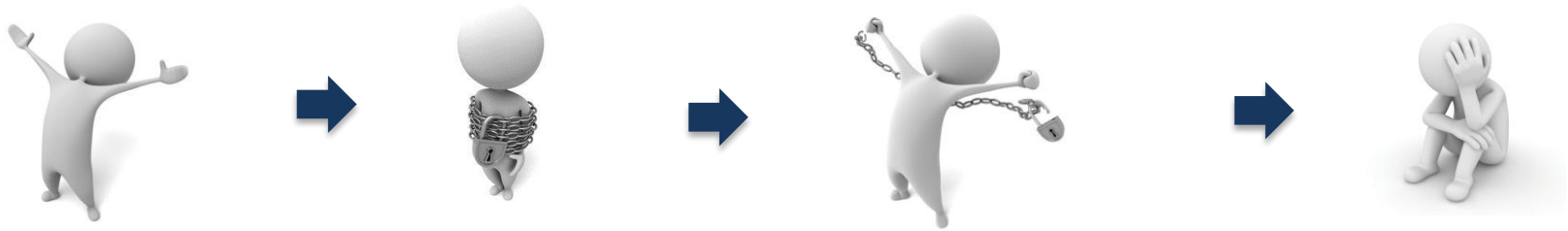
C

LONDON — A British judge on Thursday sentenced a businessman who sold fake (假冒的) bomb detectors (探测器) to 10 years in prison, saying the man's lies had caused almost potentially deadly consequences.

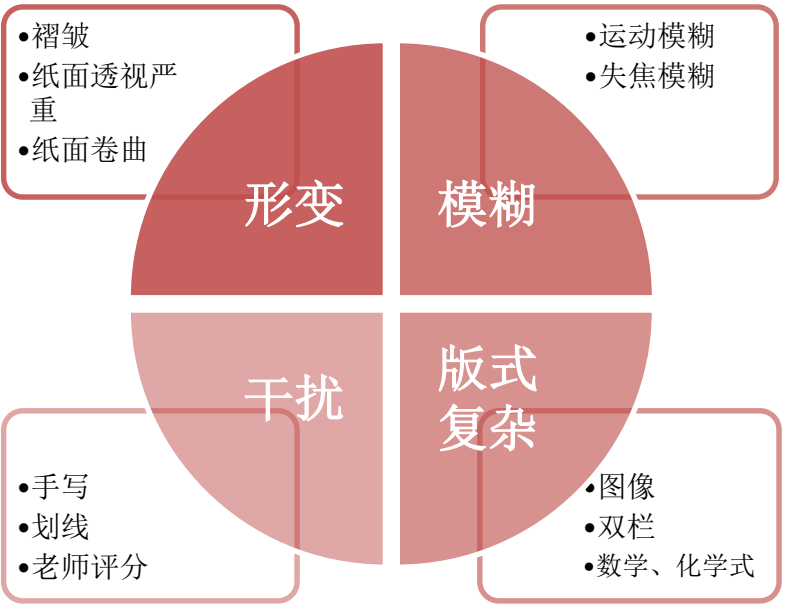
It is believed that James McCormick got about £77.8 million from the sales of his detectors — which were based on a kind of golf ball finder — to countries including Iraq, Belgium and Su-

11

最后我们发现，2013年拍照搜题没有特别现成技术，识别效果是个未知数

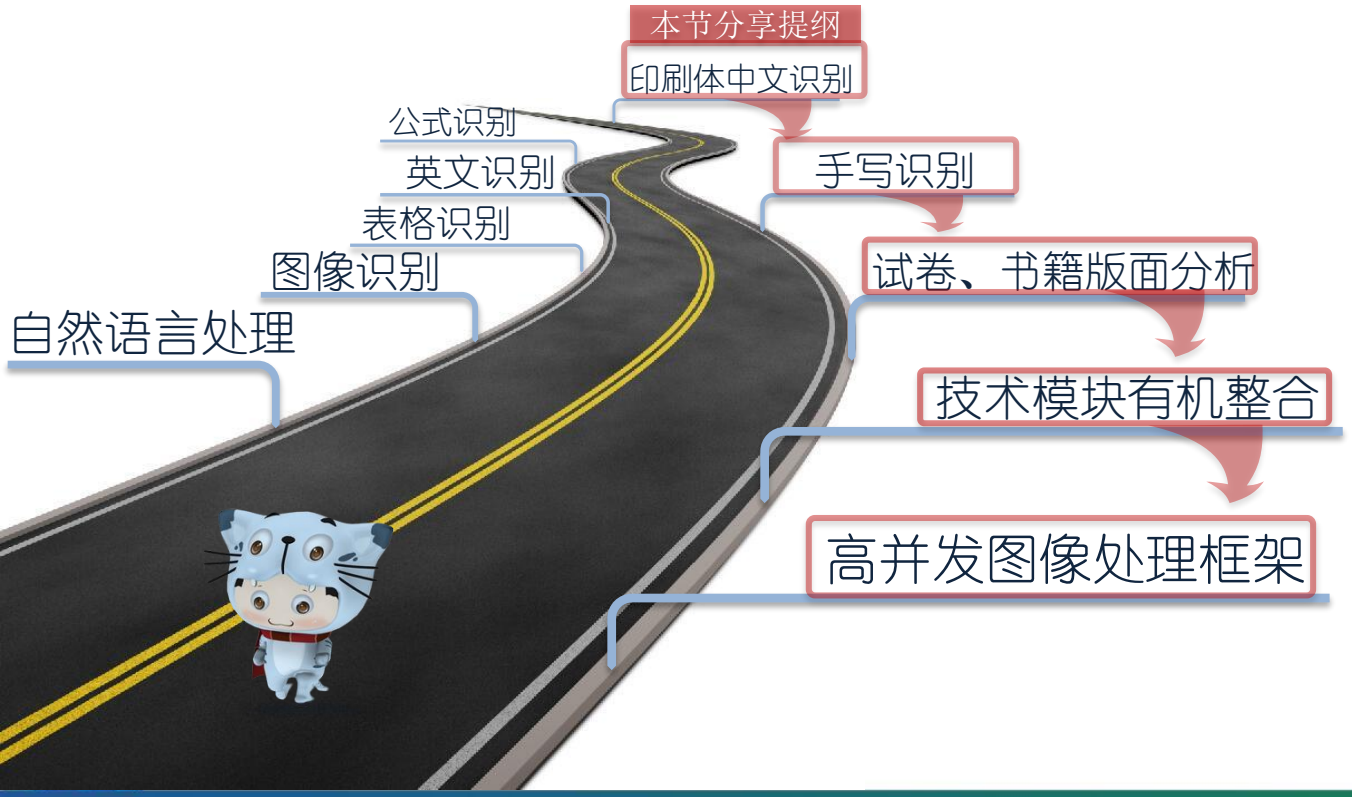


- 为什么拍照搜题的识别技术充满挑战：



- 这些难点也在逐步被攻克

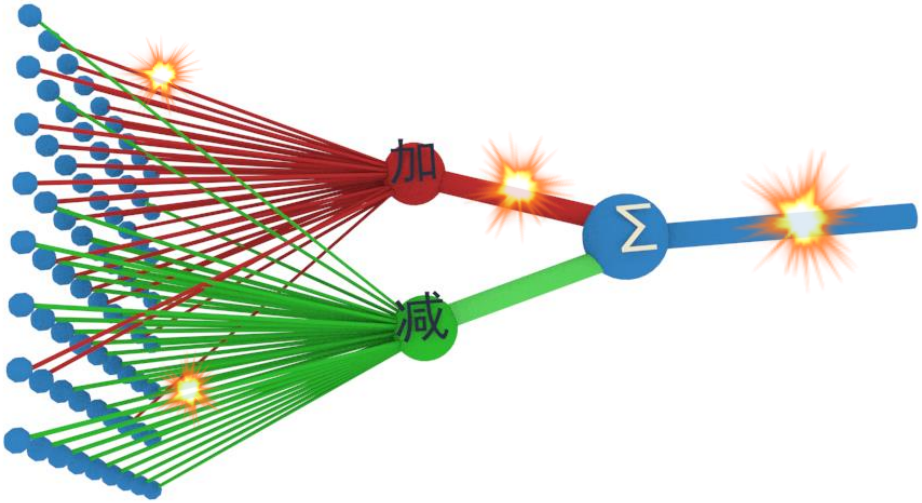
- 2013年初，学霸君开始开拓拍照搜题的核心识别技术道路



2.2. 学霸君文字识别

- 深度学习技术CNN+RNN

1. 利用CNN进行中文字符识别（训练字符库：20亿，单字符识别率：99.5%）
2. 利用RNN进行英文单词识别 （训练单词库：10亿，单词识别率： 98.3%）

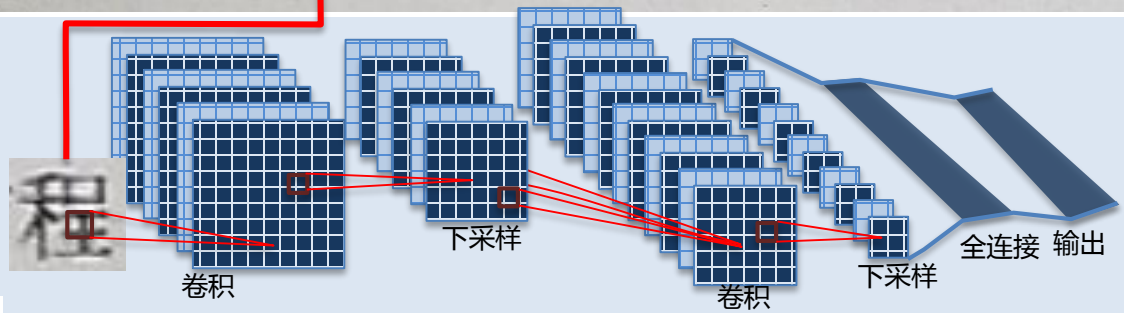


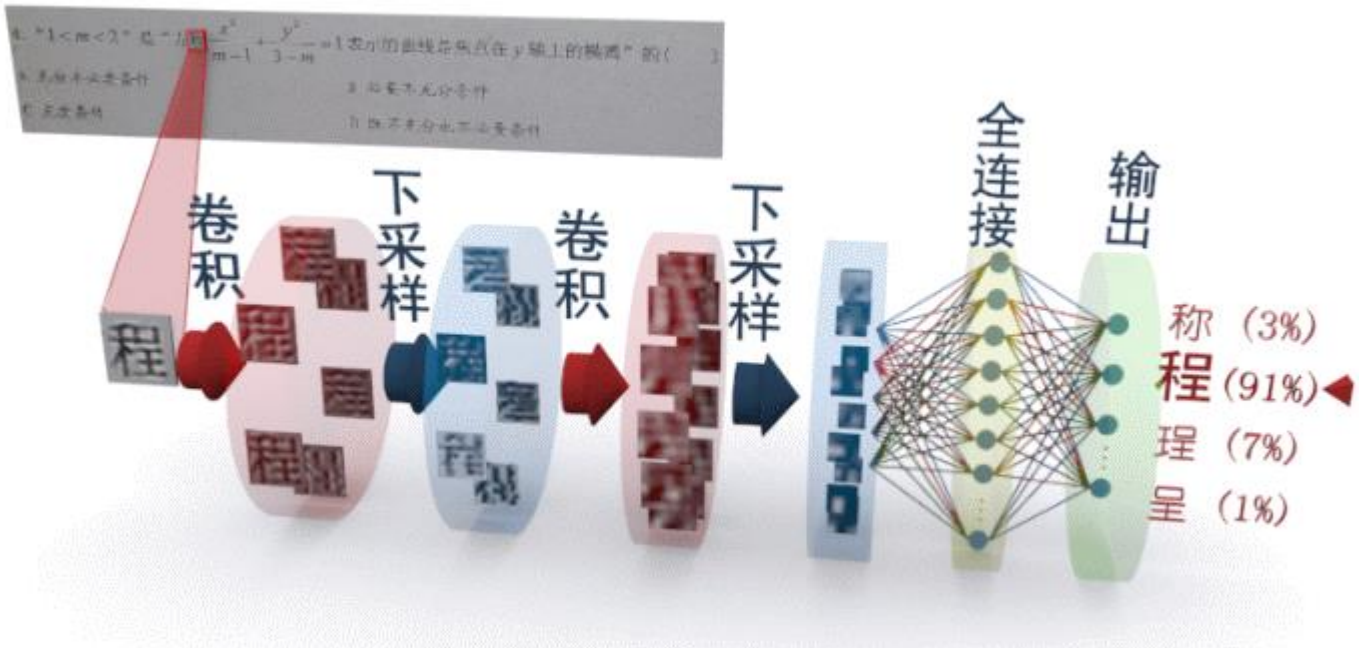
4. “$1<m<2$”是“方程$\frac{x^2}{m-1}+\frac{y^2}{3-m}=1$表示的曲线是焦点在$y$轴上的椭圆”的(　　)

A 充分不必要条件　　B 必要不充分条件

C 充要条件　　D 既不充分也不必要条件

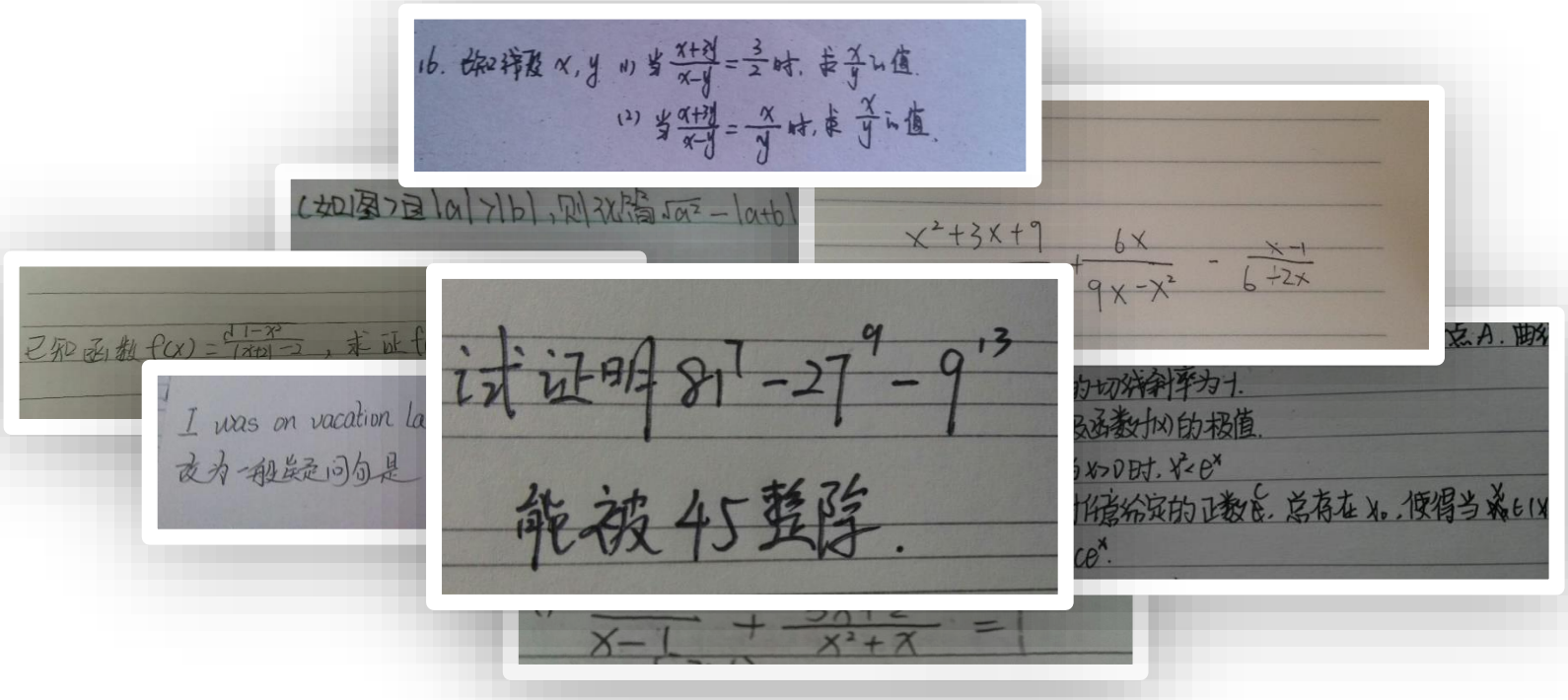
深度神经网络





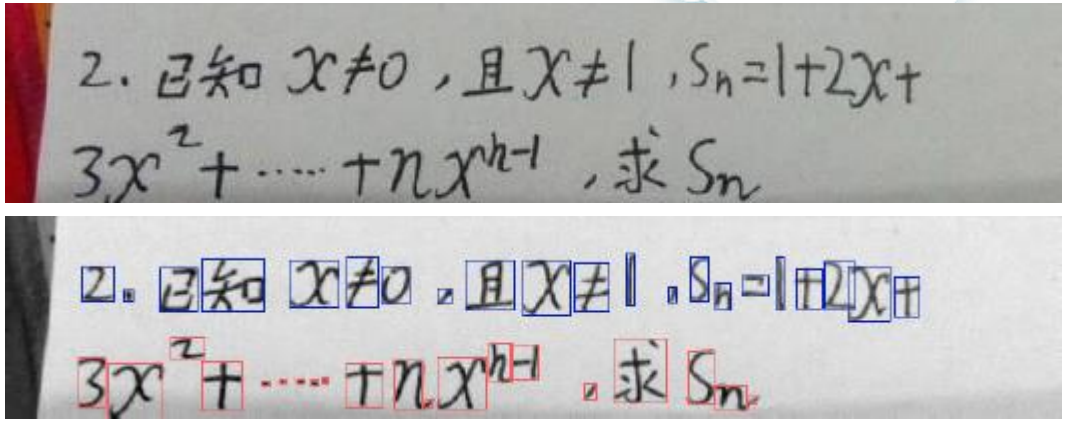
2.3. 手写识别

- 手写题目直接拍照答疑，拓宽了应用场景识别率将能进一步提升！



底纹滤除 → 版面分析 → 识别 → 自然语言处理

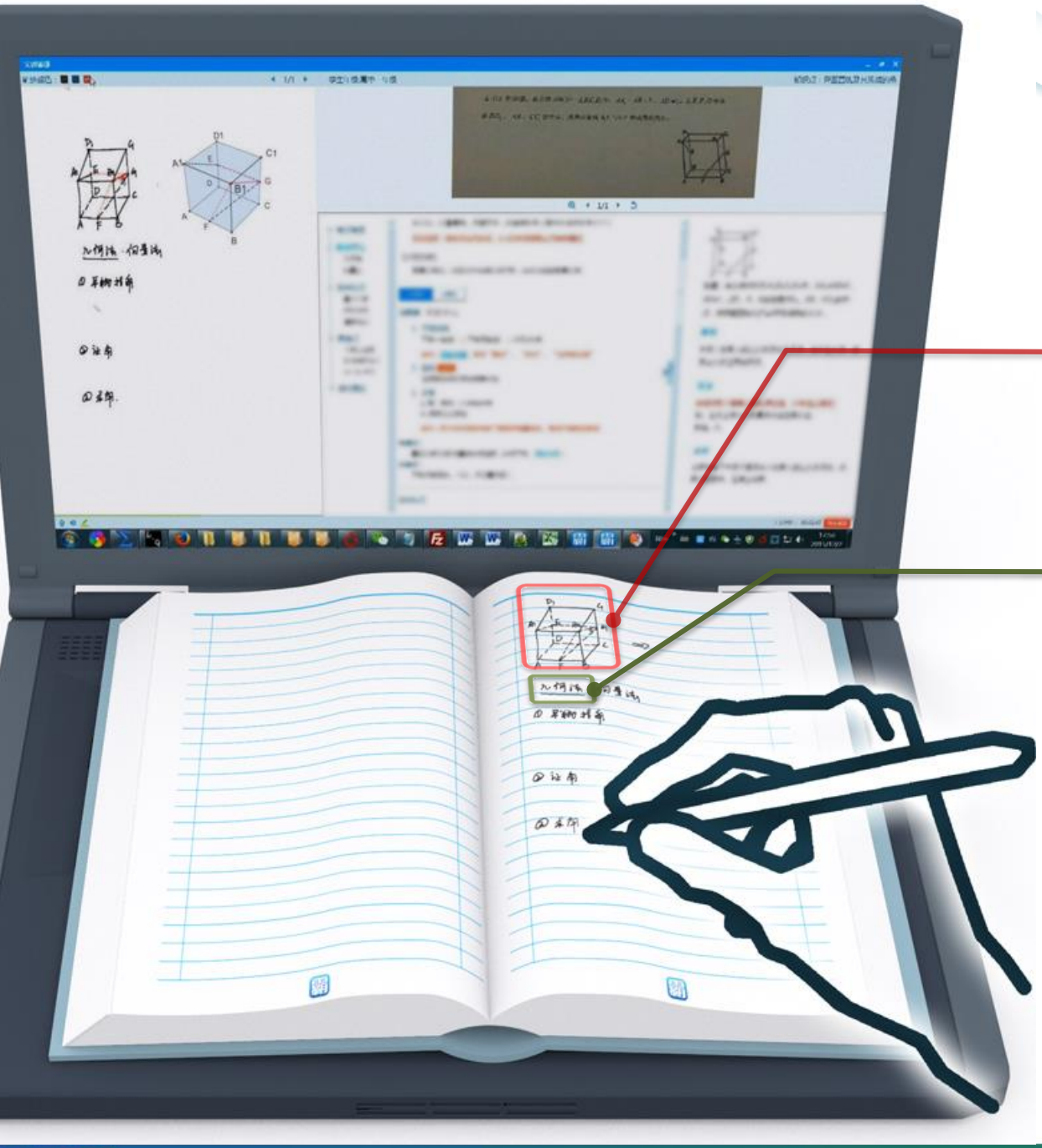
- 识别结果



2	已	知	x	≠	0	且	x	≠	1	S	n	=	1	+	2	x	+
2	已	知	x	≠	0	且	x	≠	1	s	n	=	1	+	2	x	+

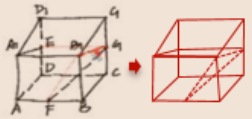
3	x	2	+	.	.	.	+	n	x	n	H	求	S	n
3	x	2	+	.	.	.	+	n	x	n	H	求	s	n

错误

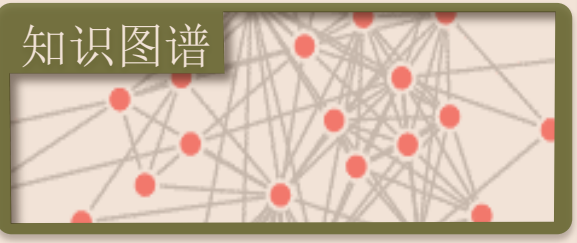


- 实时分析+知识联动算法架构

图像解读

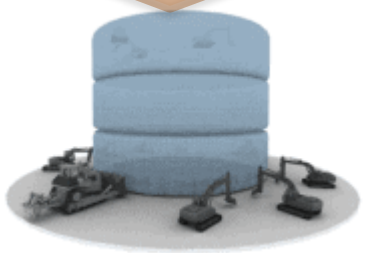


知识图谱

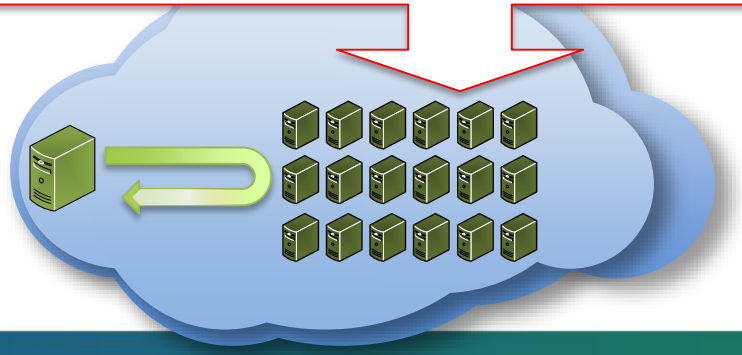


数学引擎

$$\sqrt{x}$$

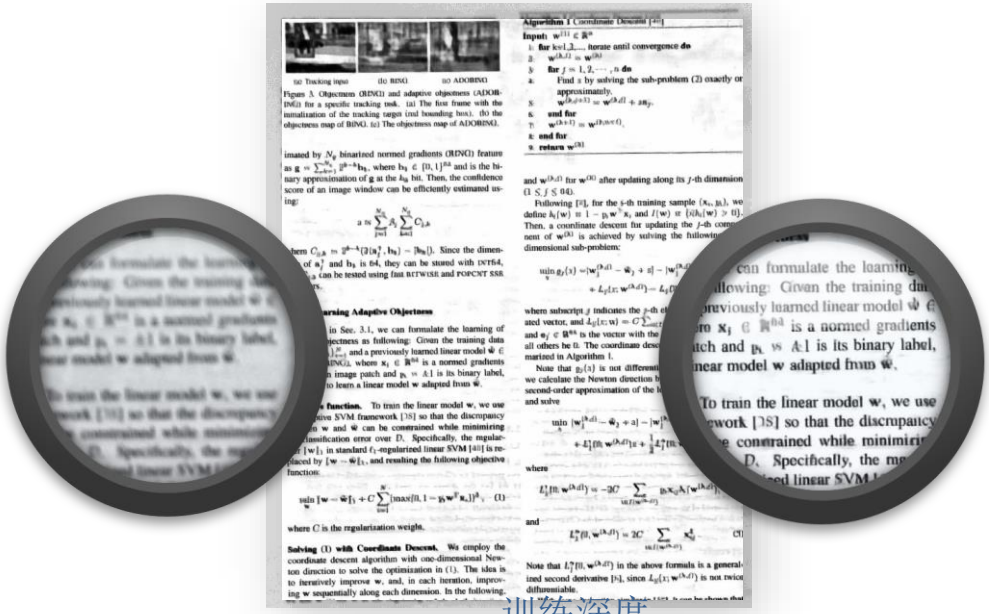


2.4. 智能化的版面分析和题目提取



2.5.图像恢复技术，解决移动端采集图片质量低的问题

可“读”即可“识”，非可“读”也可能“识”，挑战人眼文字辨识极限



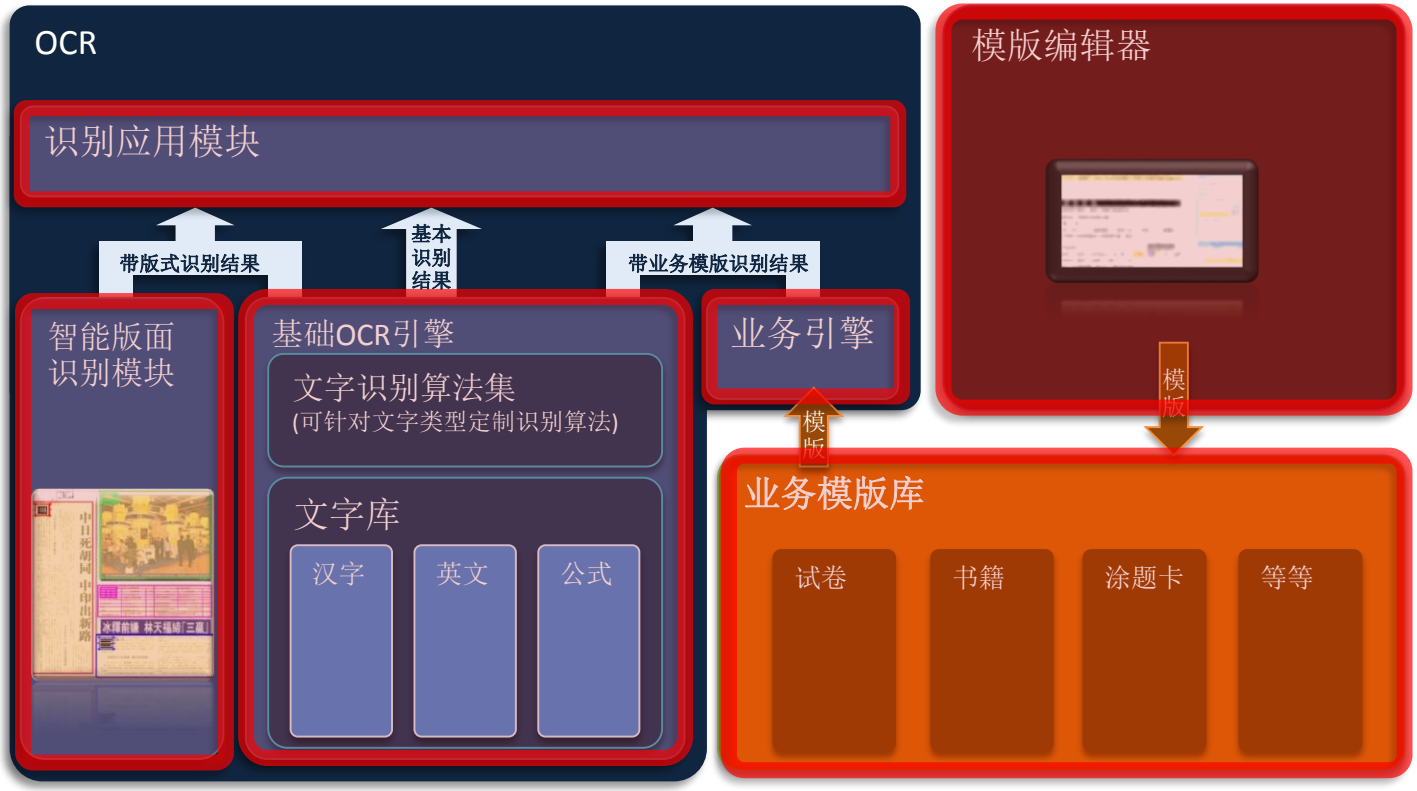
收集、标注
原始图片

训练深度
神经网络
（GPU集
群）

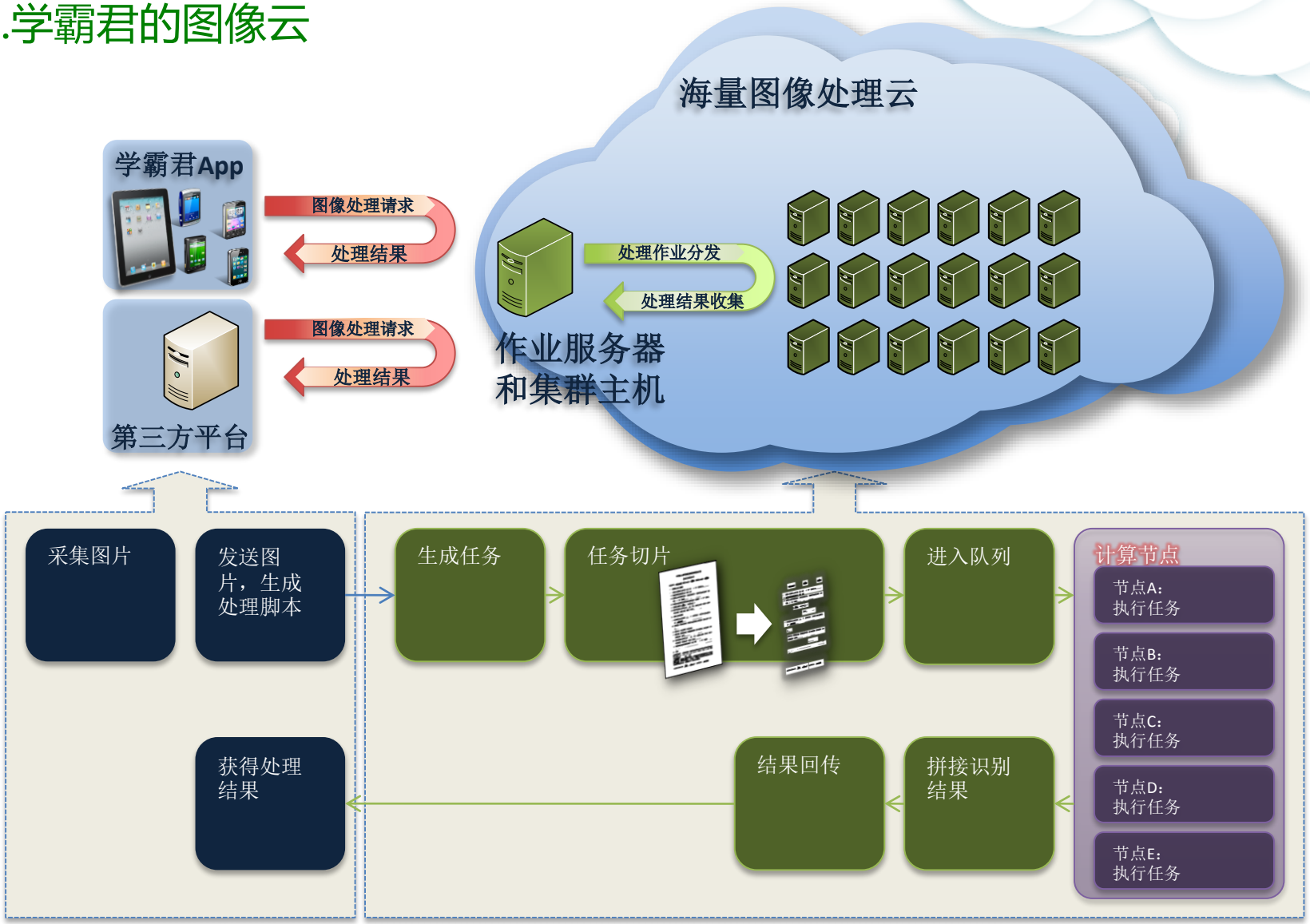
生成模型
应用模型

2.6. 学霸君文字识别技术框架

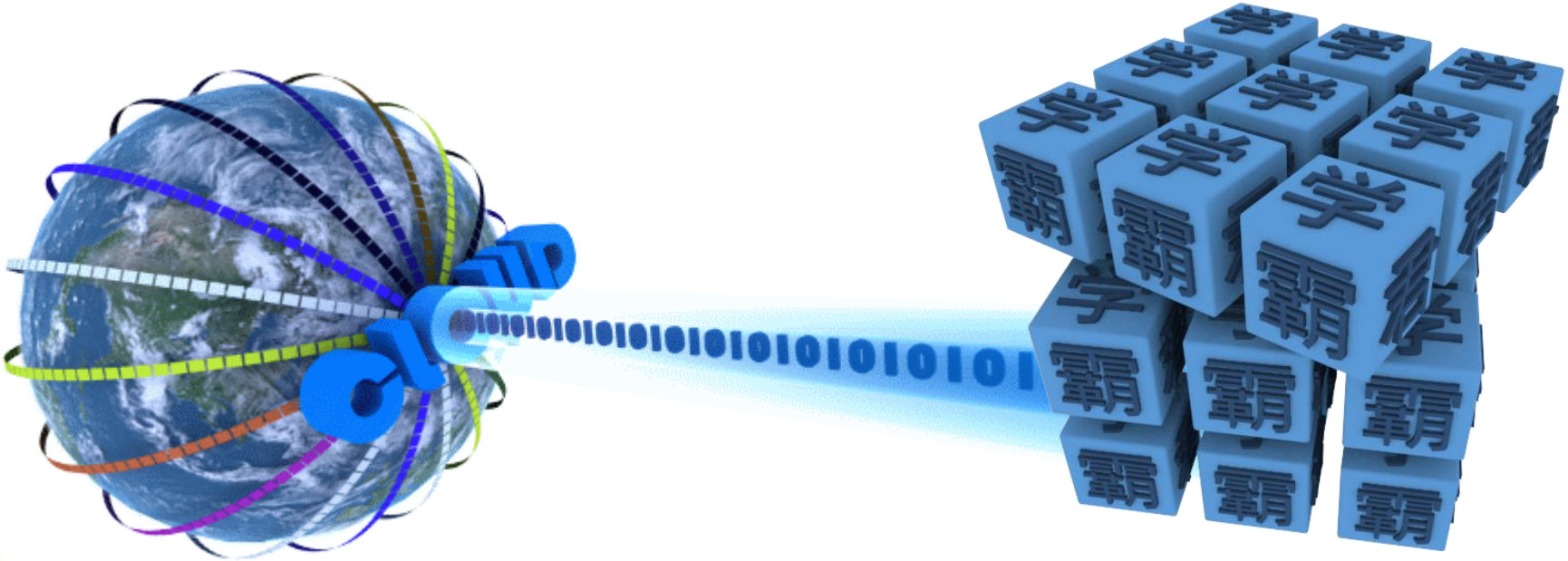
- 经过三年耕耘，学霸君形成的OCR技术核心架构



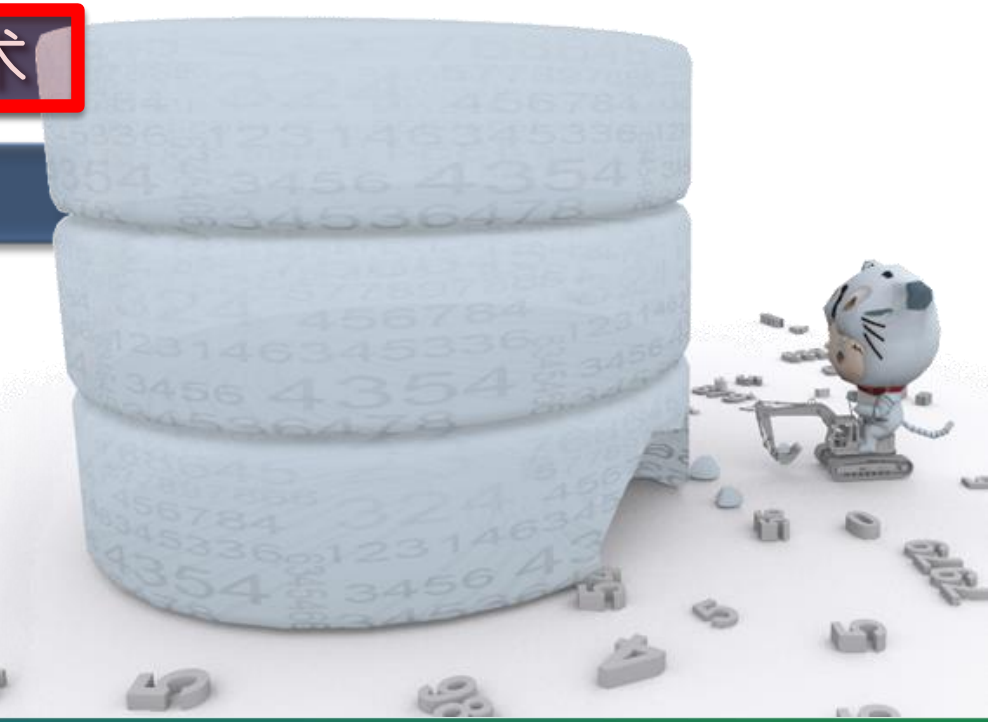
2.7.学霸君的图像云


海量图像处理云
学霸君App
图像处理请求
处理结果
处理作业分发
处理结果收集
作业服务器
和集群主机
图像处理请求
处理结果
第三方平台
采集图片
发送图片，生成处理脚本
生成任务
任务切片
进入队列
计算节点
节点A:
执行任务
节点B:
执行任务
节点C:
执行任务
节点D:
执行任务
节点E:
执行任务
获得处理结果
结果回传
拼接识别结果

- 在OCR技术与大规模移动计算的推进下，学霸君初步完成了大规模的学生拍题数据搜集，并构建自身的知识模块



目录



3.1.最核心技术：分发策略——让最合适的老师给一个学生讲题

- 一个与Uber调度可类比却又很不同的问题，挑战如下：
  - 老师
    - 上线时间不确定
    - 老师擅长版块不一致
    - 各地教纲不同
    - 讲题方式不同
  - 学生
    - 随机发起提问
    - 对价格敏感程度不同
    - 对获取结果期待不同

学霸君1V1调度算法架构

分发算法

随机建模

最优化

智能调度

预测算法

需求预测

供给预测

用户画像

学生画像

老师画像

知识模型

知识导航

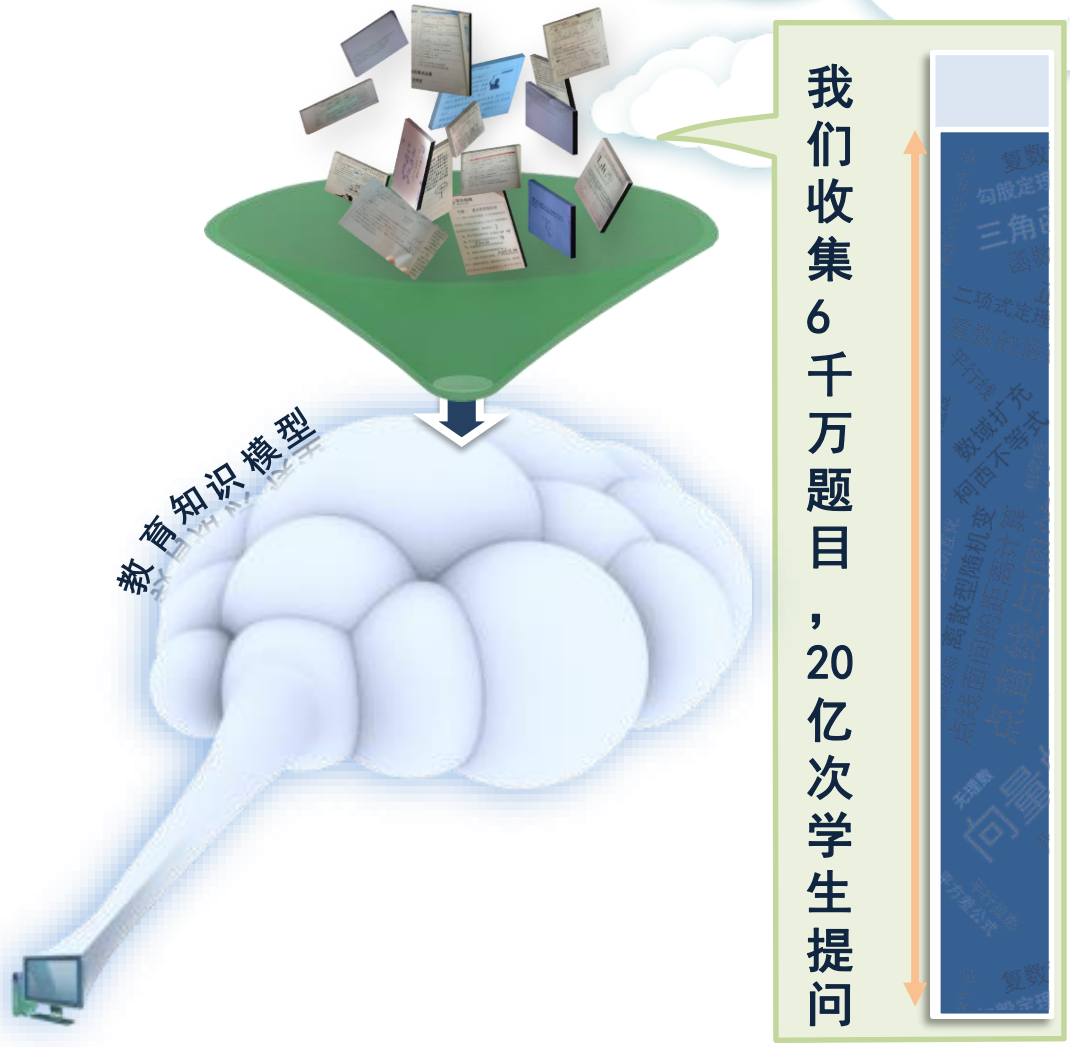
知识图谱

本节分享提纲

3.3. 知识导航体系

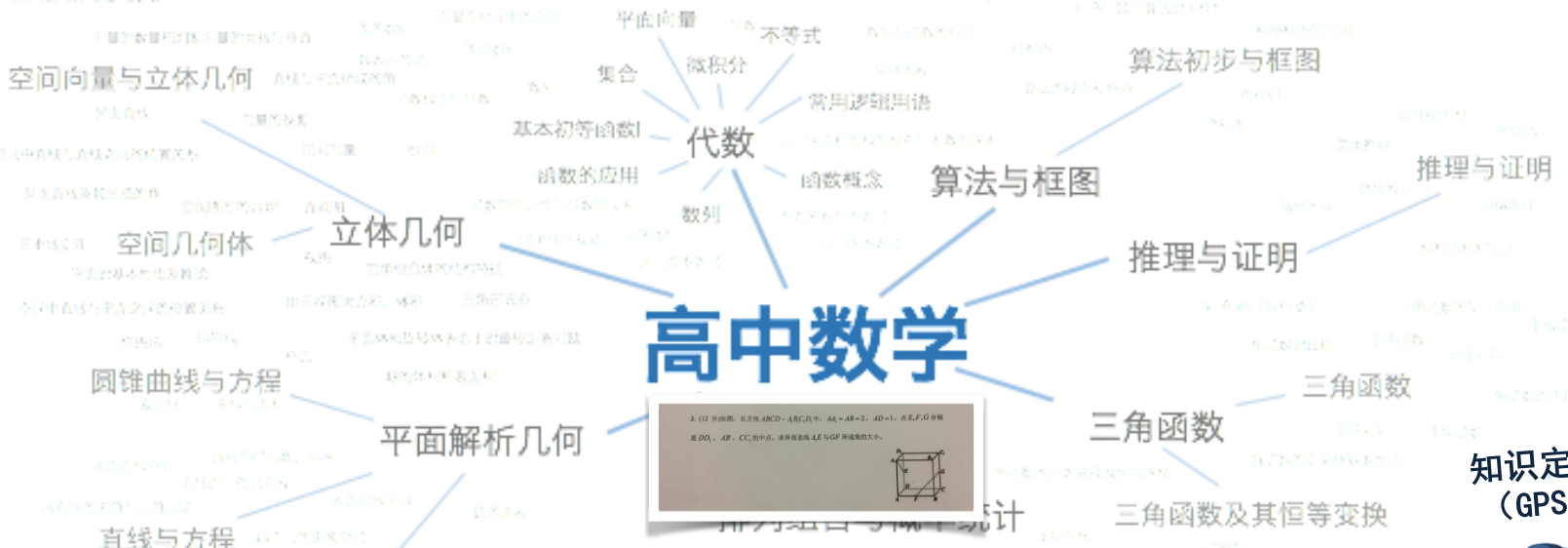
基于自然语言理解和深度学习的数据挖掘

- 通过高性能机器学习技术，我们对数据进行多种维度的挖掘和分析
- 在教育领域，我们搭建了基于海量数据的知识导航系统

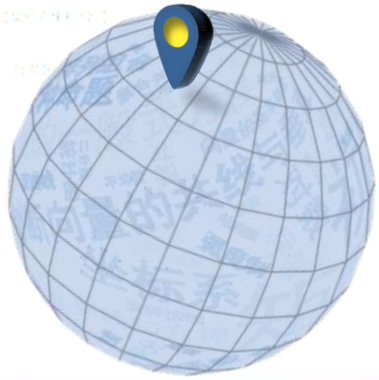


- 对于任意一道K12题目，可以通过数据挖掘的结果自动分析出其知识点，并推荐学习内容。基于大数据的分析对学生精准学习非常有益

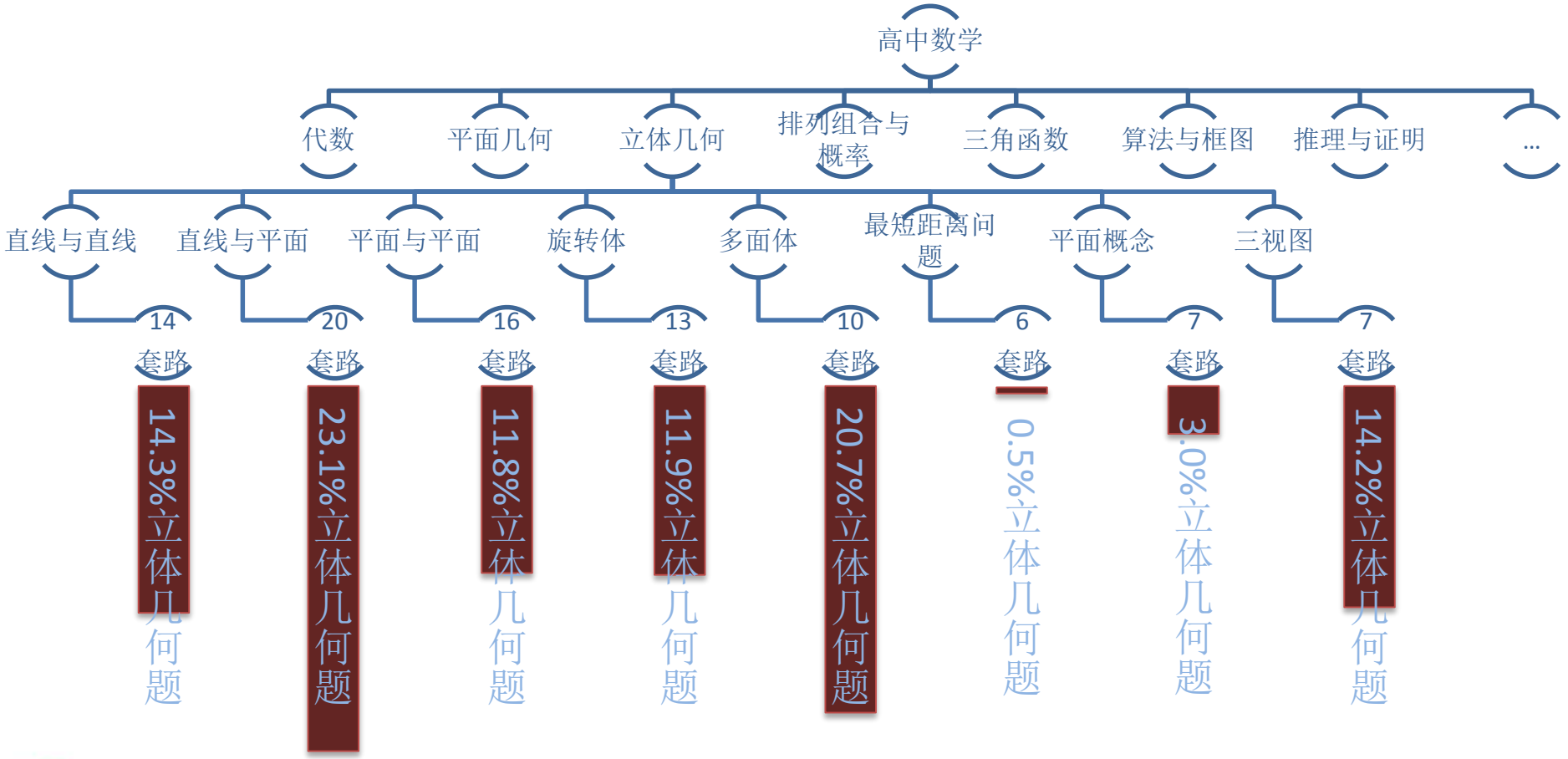
7个板块　　22个章节　　550个知识点　　3529个题型



知识定位（GPS）



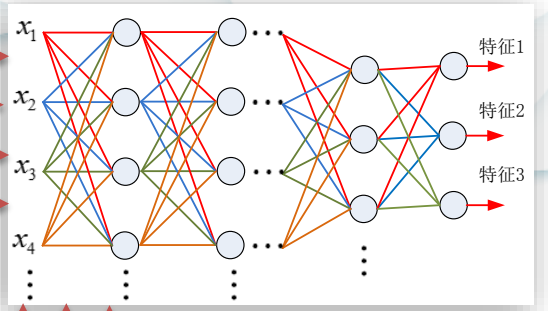
- 通过数据挖掘发现：任何知识点有解题套路，掌握套路， 就能得到好成绩。
- 比如立体几何可以归纳为100来个套路，涵盖所有题目的解法。
- 在实时1V1中，题目画像是老师资源配给的重要依据



3.4. 学生画像：一花一世界，一叶一菩提

- 聚拢了约3000万学生用户，每个学生都有不一样的属性：
  - 年级
  - 地区教材
  - 对不同知识点的掌握水平
  - 学习能力
  - 等等

深度神经网络



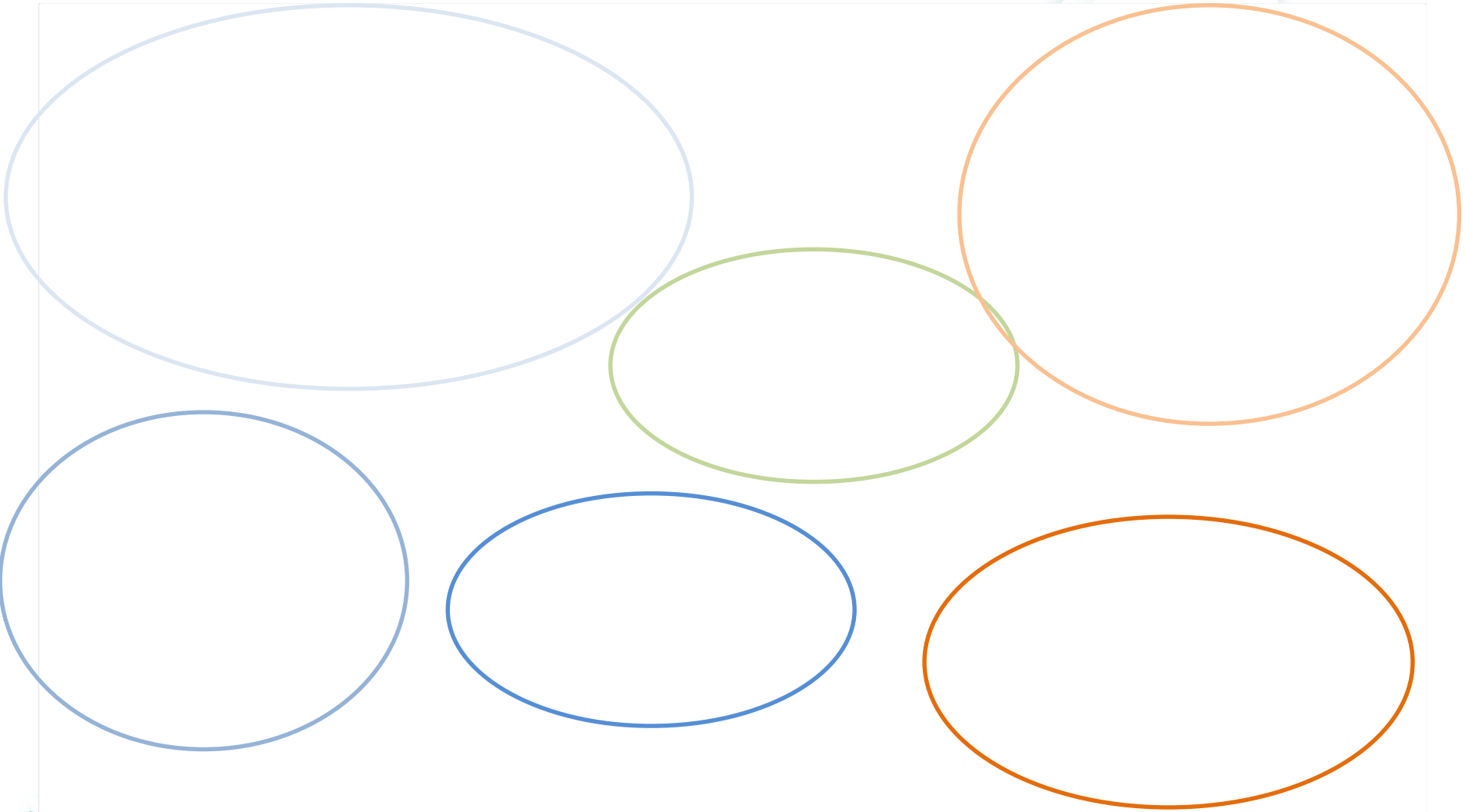
时间

时间

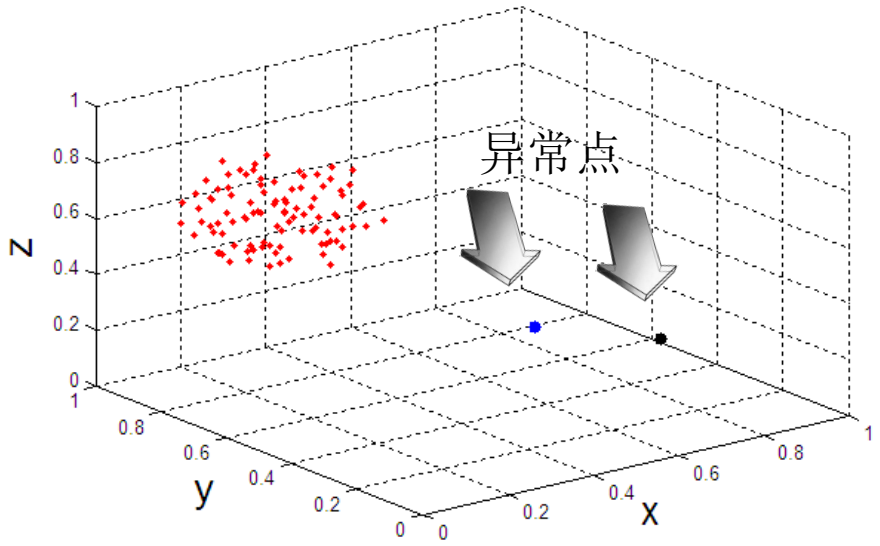
时间

时间

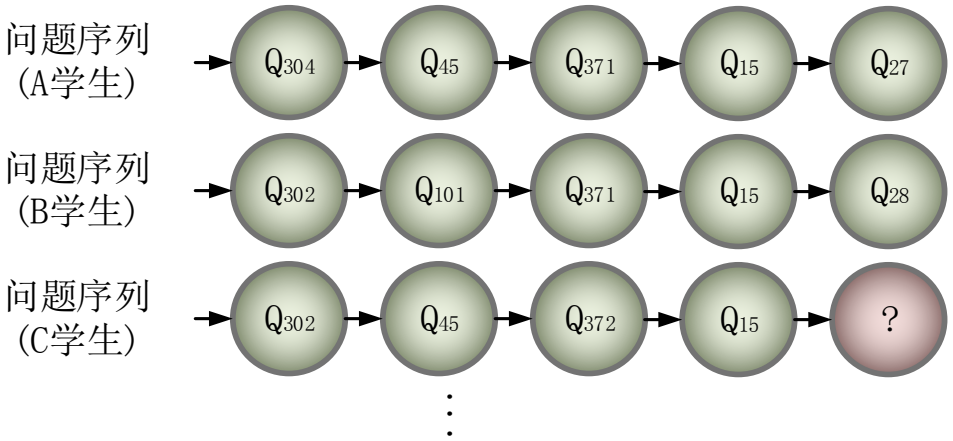
- 学生用户聚类(Clustering)
  - 按照某种相似性将用户分组

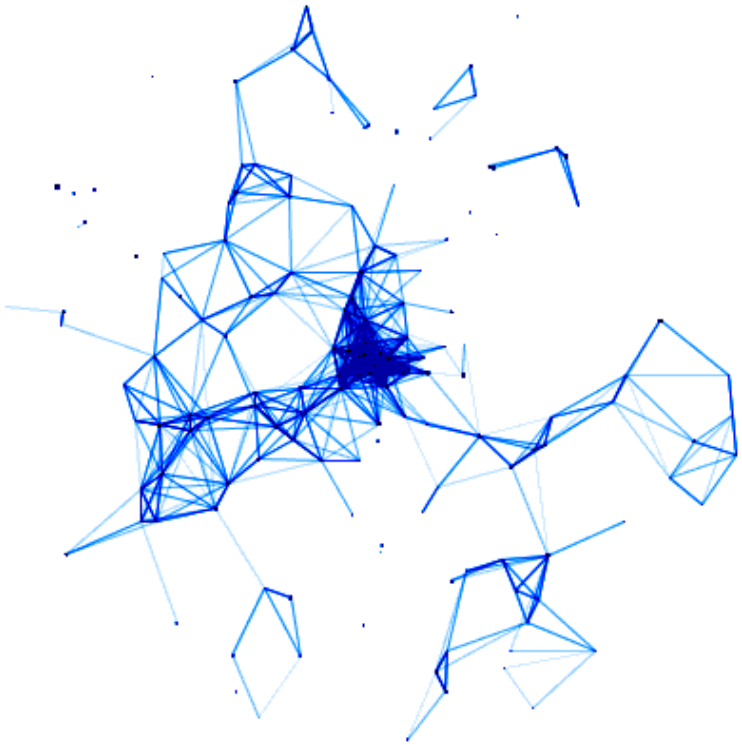


- 异常学生用户检测
  - 恶意破坏
  - 欺诈行为
  - 友商的密集试用



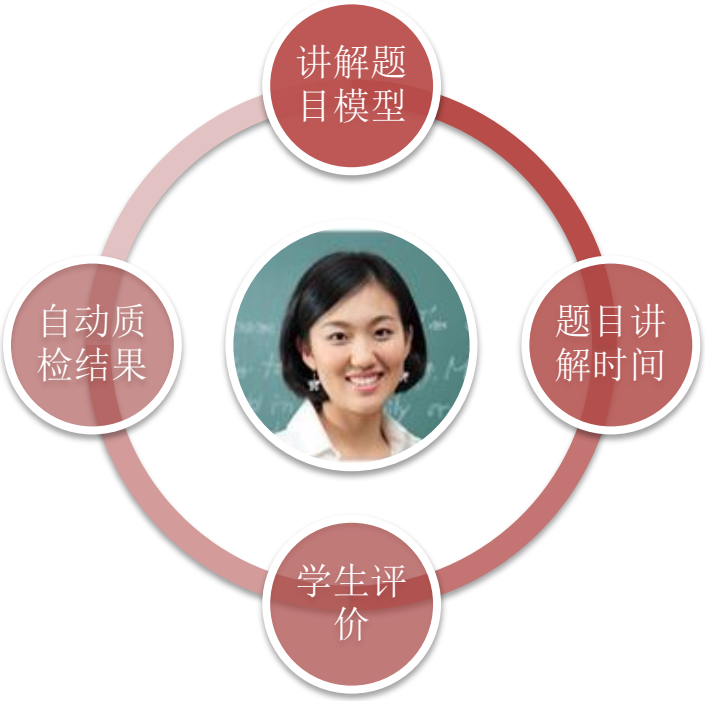
- 实际应用中，需要考虑时间轴和相关性，分析中会产生动态学习轨迹
- 对于顺序发生的事件的行为模式进行挖掘和分析





3.5.老师画像

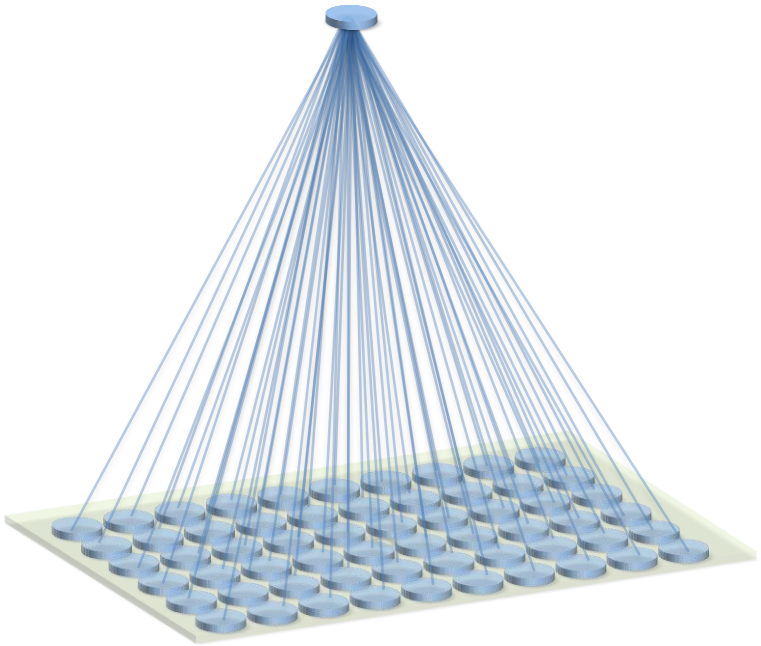
- 每次老师主动讲题，都是一次对老师能力空间的评估
- 积累了一段时间后，每个老师的擅长领域及答疑习惯都能从数据中体现出来



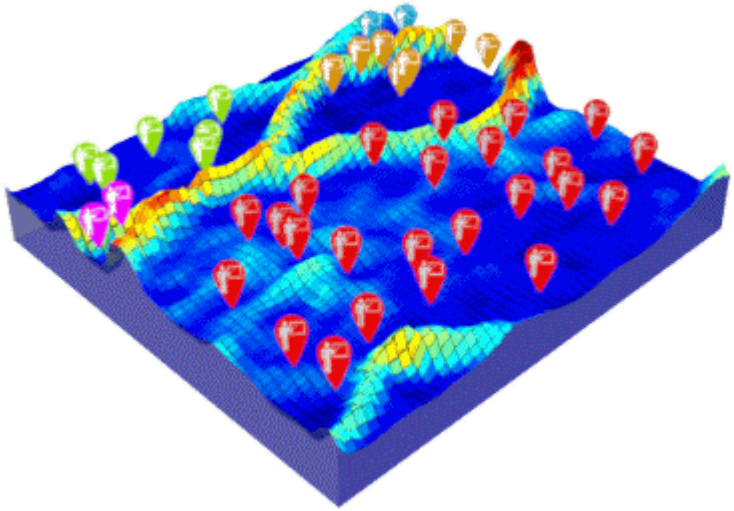
- 老师的分类问题的困难
  - 能力难以量化
  - 区域教纲影响明显
  - 讲法差别很大

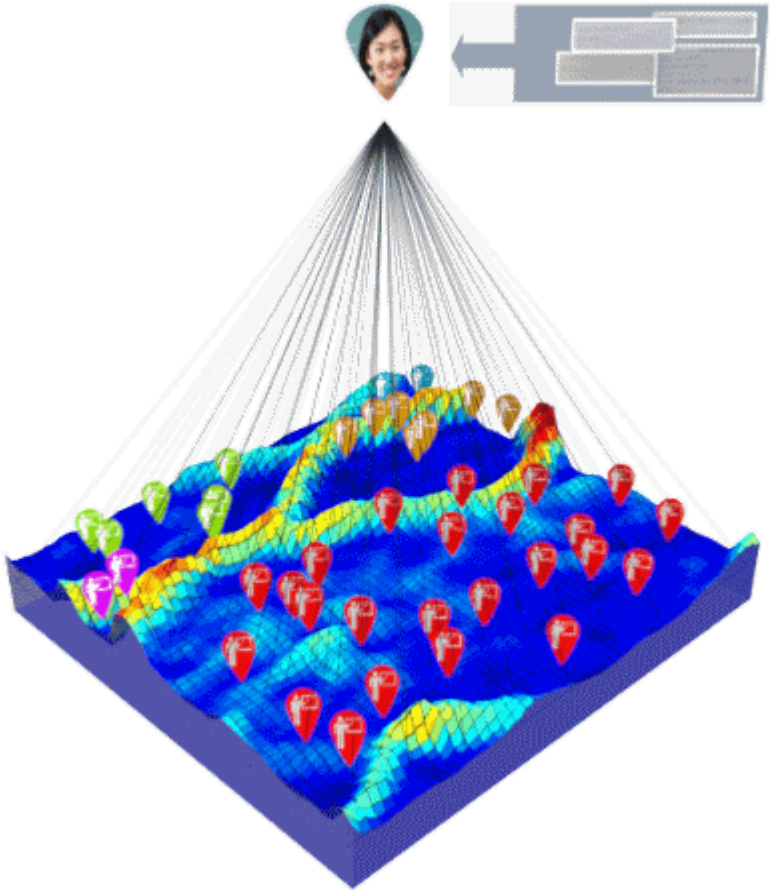
- 解决方法

  自组织地图（竞争神经网络）



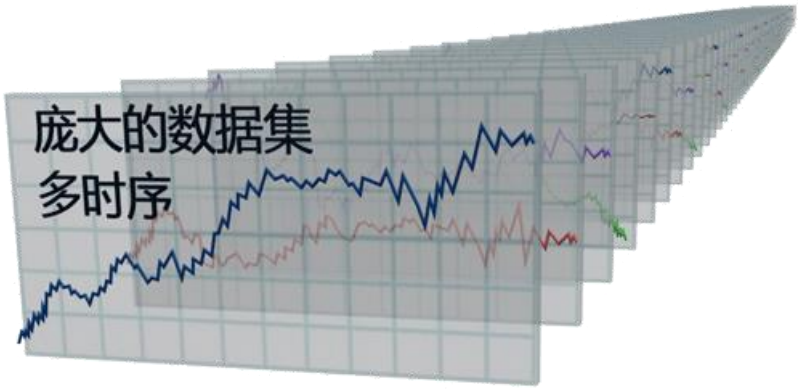
- 按照区域化的划分，每个老师都可以归于不同的组群里
- 当学生提出答疑需求时，学霸君后台会优先选择最为匹配的老师进行答疑服务

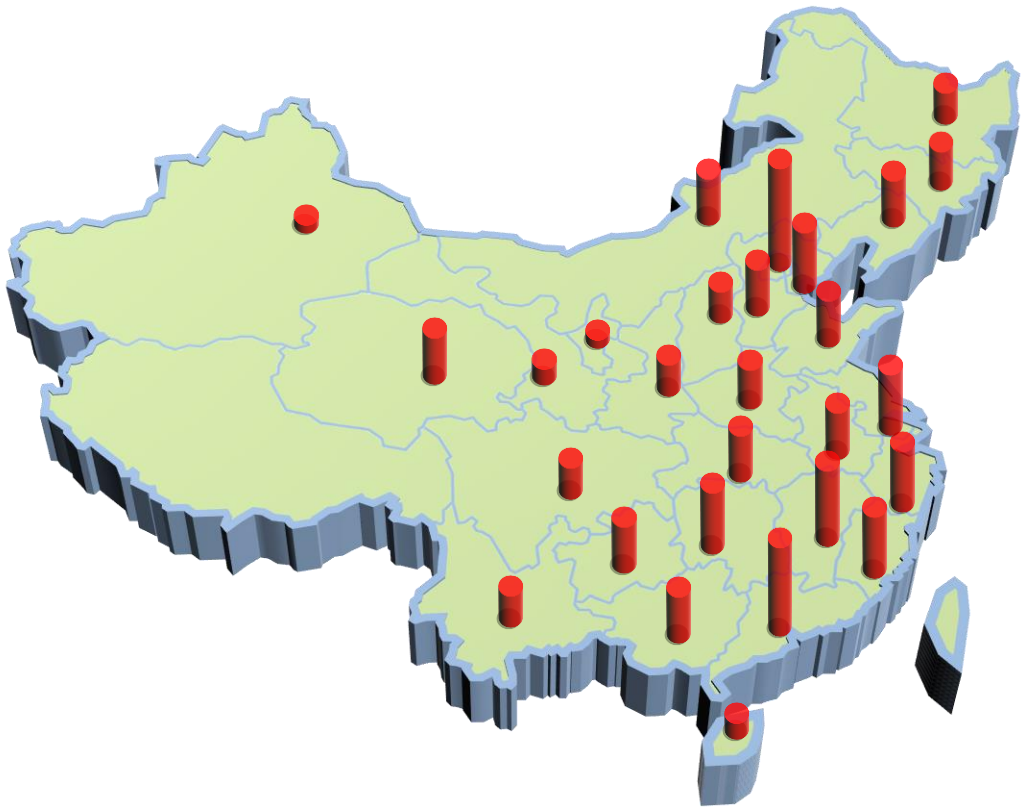




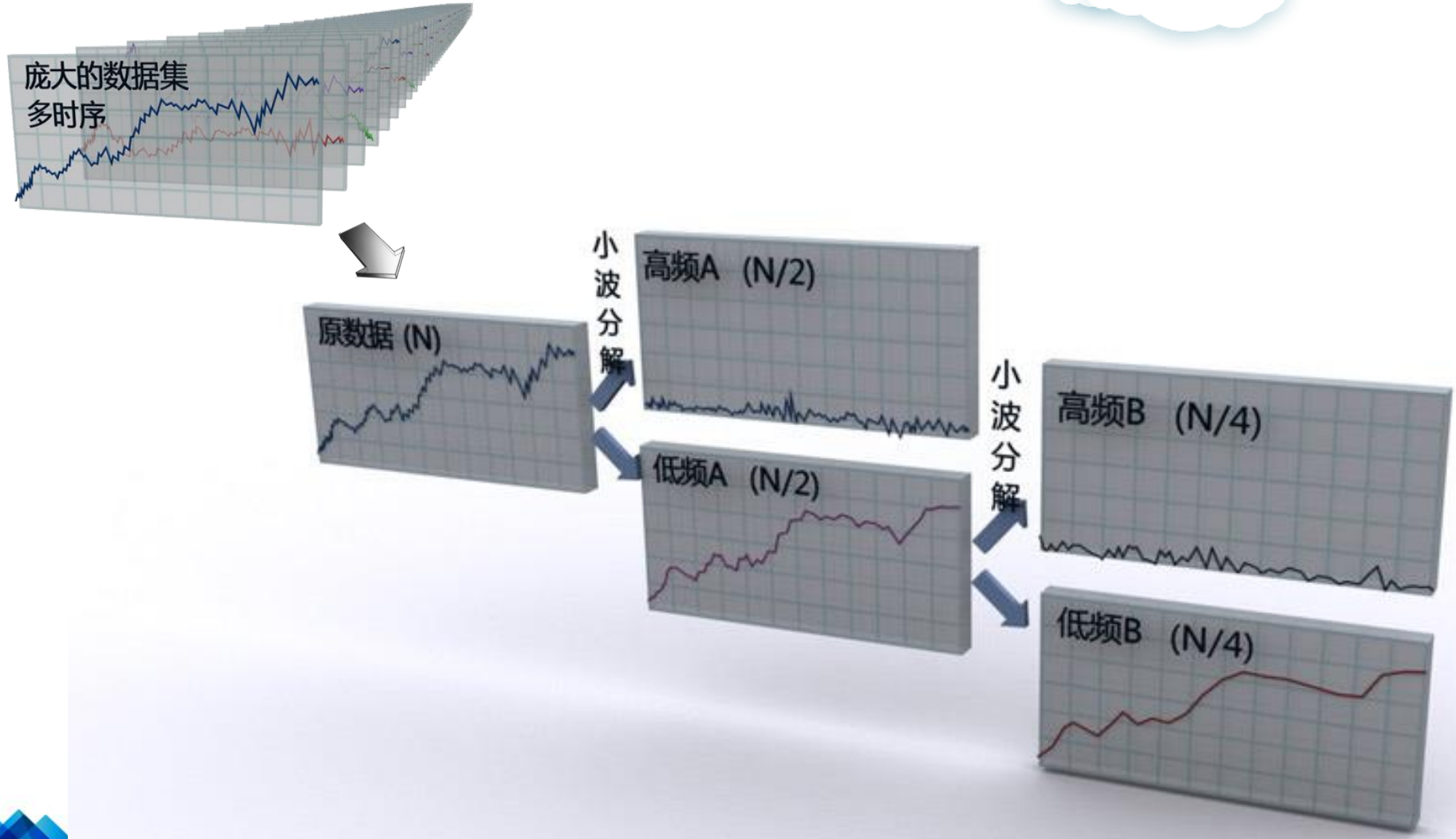
3.6.调度的基石——供应预测

- 每个省份，每个知识点对应的老师的上线时间都具有强烈的随机形态。
- 从不同尺度对老师答疑服务量进行评估，可以得到：
  - 以省份为颗粒度的供应时序
  - 以知识点为颗粒度的供应时序
  - 以个人为颗粒度的供应时序





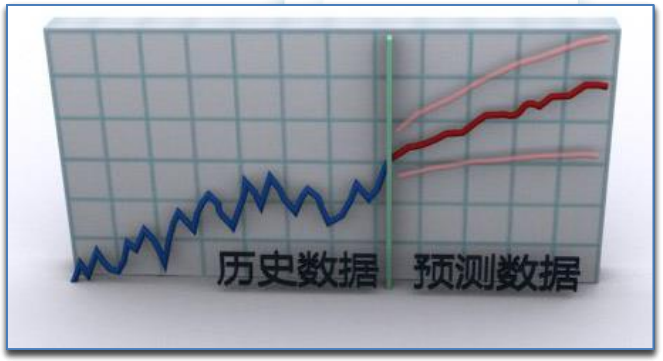
- 学霸君时序处理技术：基于多分辨率(小波变换)的时间序列分析
  - 大幅度降低数据量
  - 提高分析实时性
  - 提高抗噪声能力

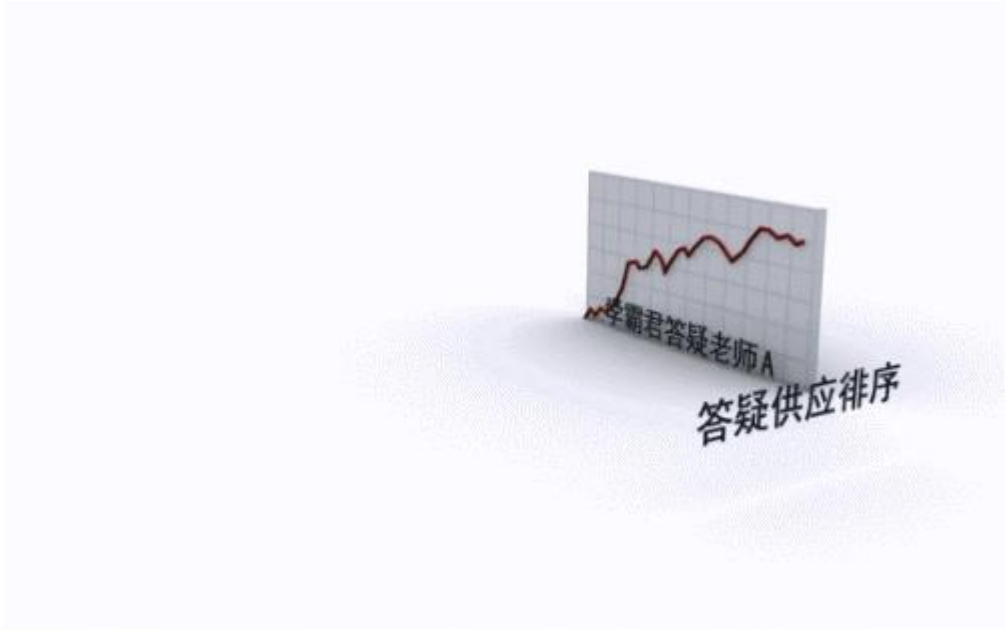


预测每个老师个体的上线时间

统计每个知识点的答疑供应能力

根据历史信息评估老师未来的服务能力





3.7.基于精益工程的老师答疑供给库存模型MTO（Make-To-Order）

调度协调(自动+人工)

老师答疑供给库存

学生答疑需求

简化的数学模型：

- $S=(x, \gamma_1, \gamma_2)$

x：老师服务库存队列

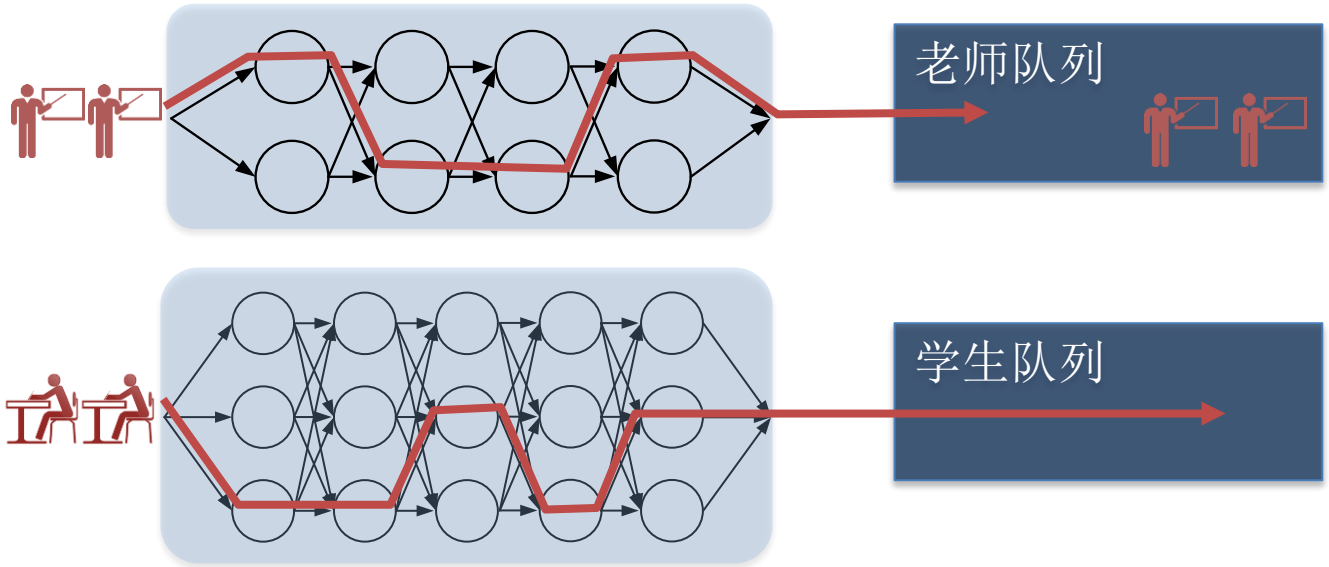
γ_1：老师上线事件

γ_2：学生提问事件

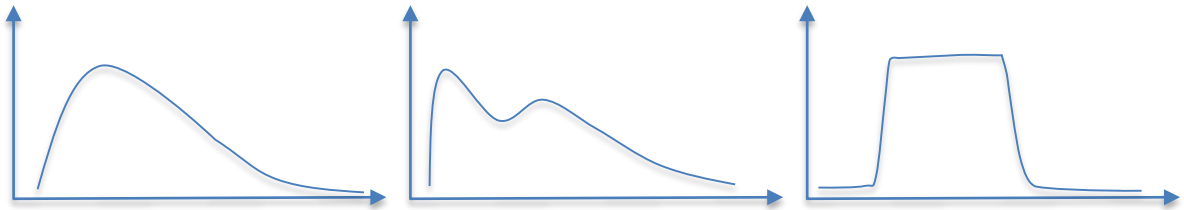
值得一提的是：

γ_1，γ_2服从复杂的Hyper-hypo-exponential随机分布

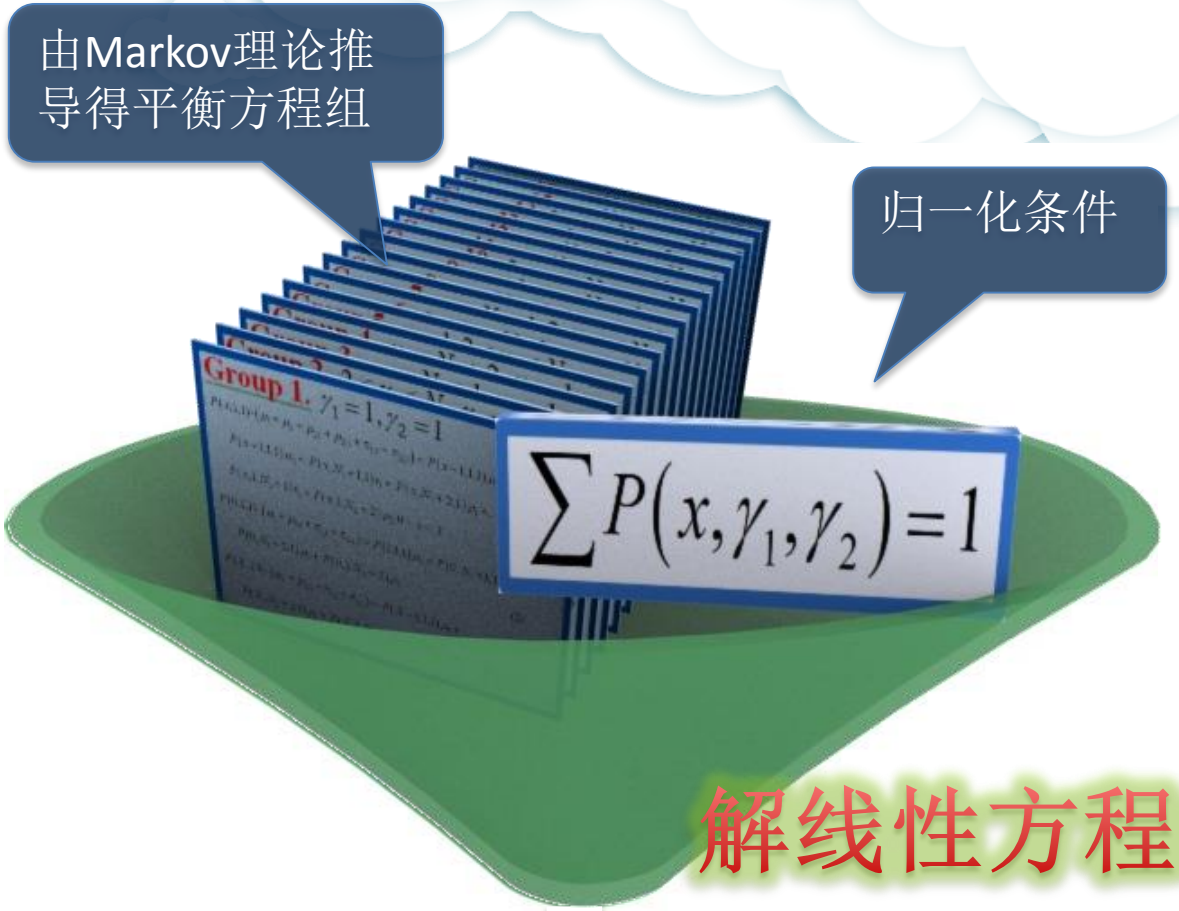
- 学霸君1V1答疑系统的Markov过程



Hyper-hypo-exponential随机分布，可以拟合各类随机分布



- 数学计算过程



求解得到系统稳态分布:$P(x, \gamma_1, \gamma_2)$

能预测出:
老师队列长度及等待时间
学生队列长度及等待时间

运筹学模型(极度简化版):

Maxmize: $\alpha \cdot$ 答疑总量 $+ \beta \cdot$ 答疑评分 $- \gamma \cdot$ 提问流失率 $- \delta \cdot$ 老师闲置率 $-$ 其他成本

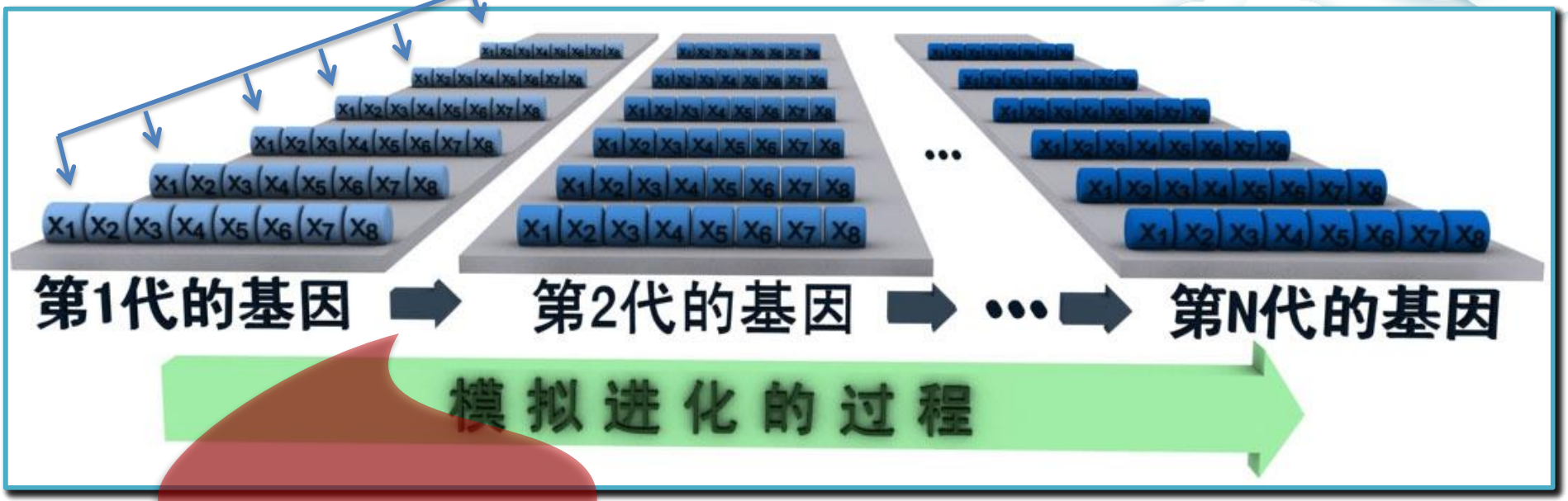
Subject to: 老师实际调配量x(t) $<$ 最大老师量MaxSupply(t)

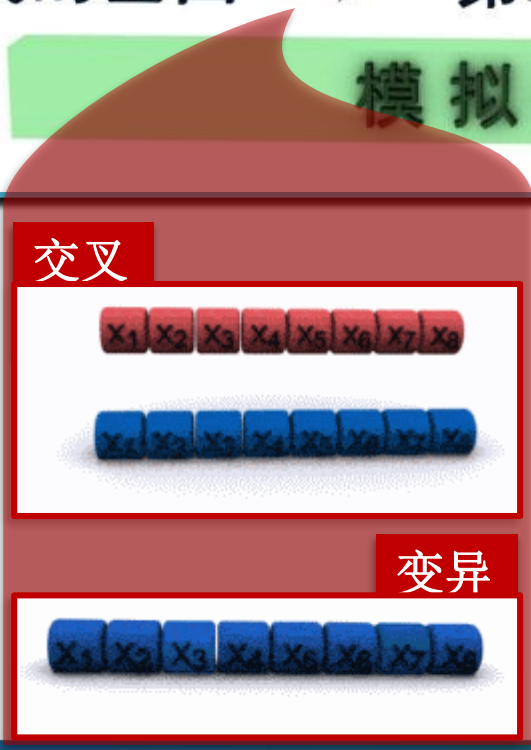
- 调度决策

  高效地自动排班，保障答疑时效的同时又节约了老师人力成本。

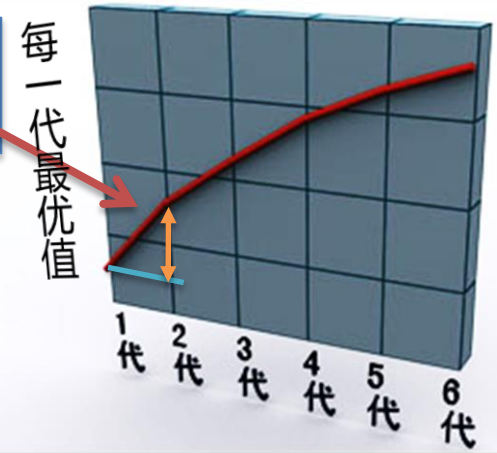
排班优化实战：遗传算法 (GA)

一个基因序列代表一种排班策略

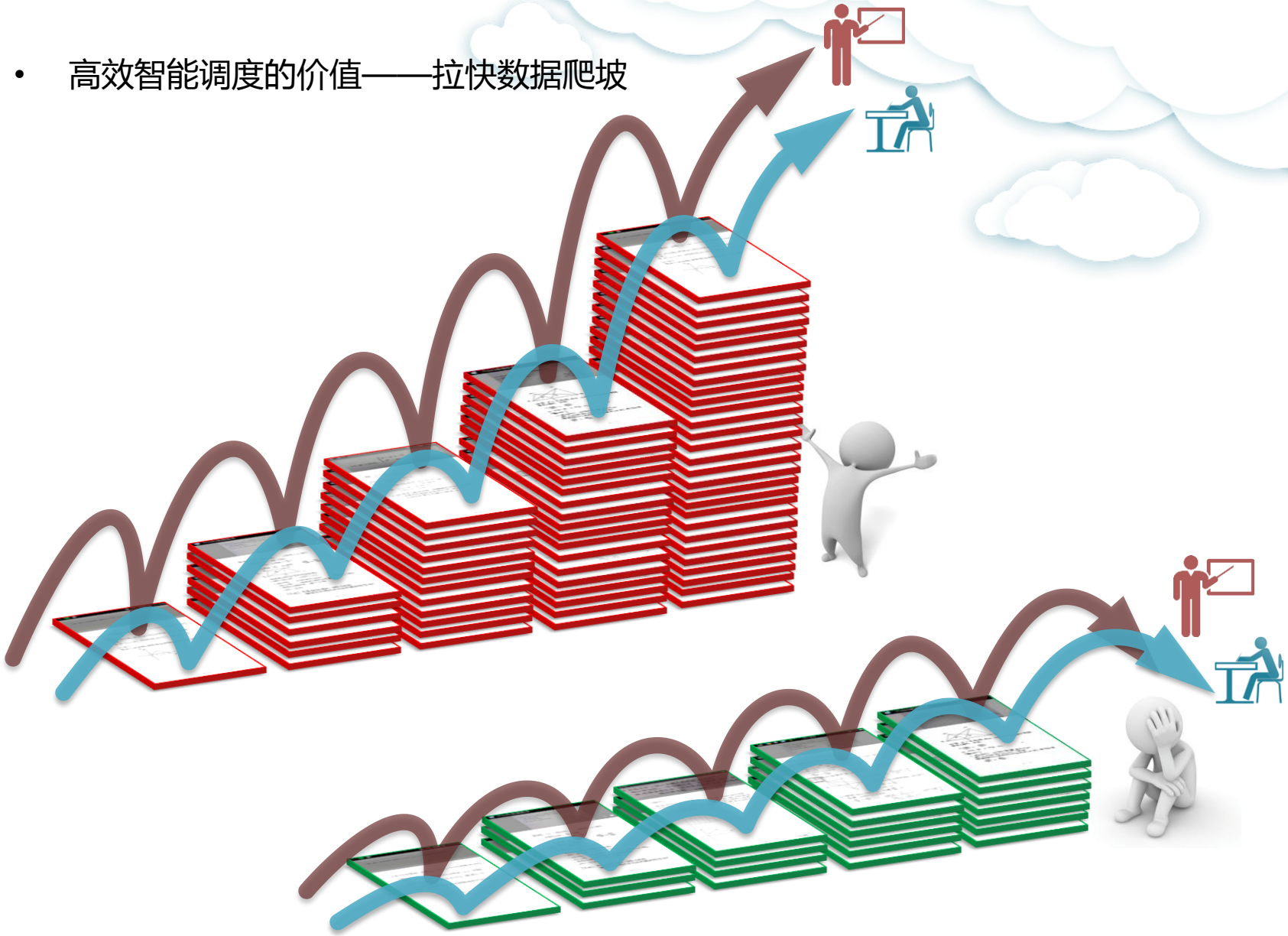




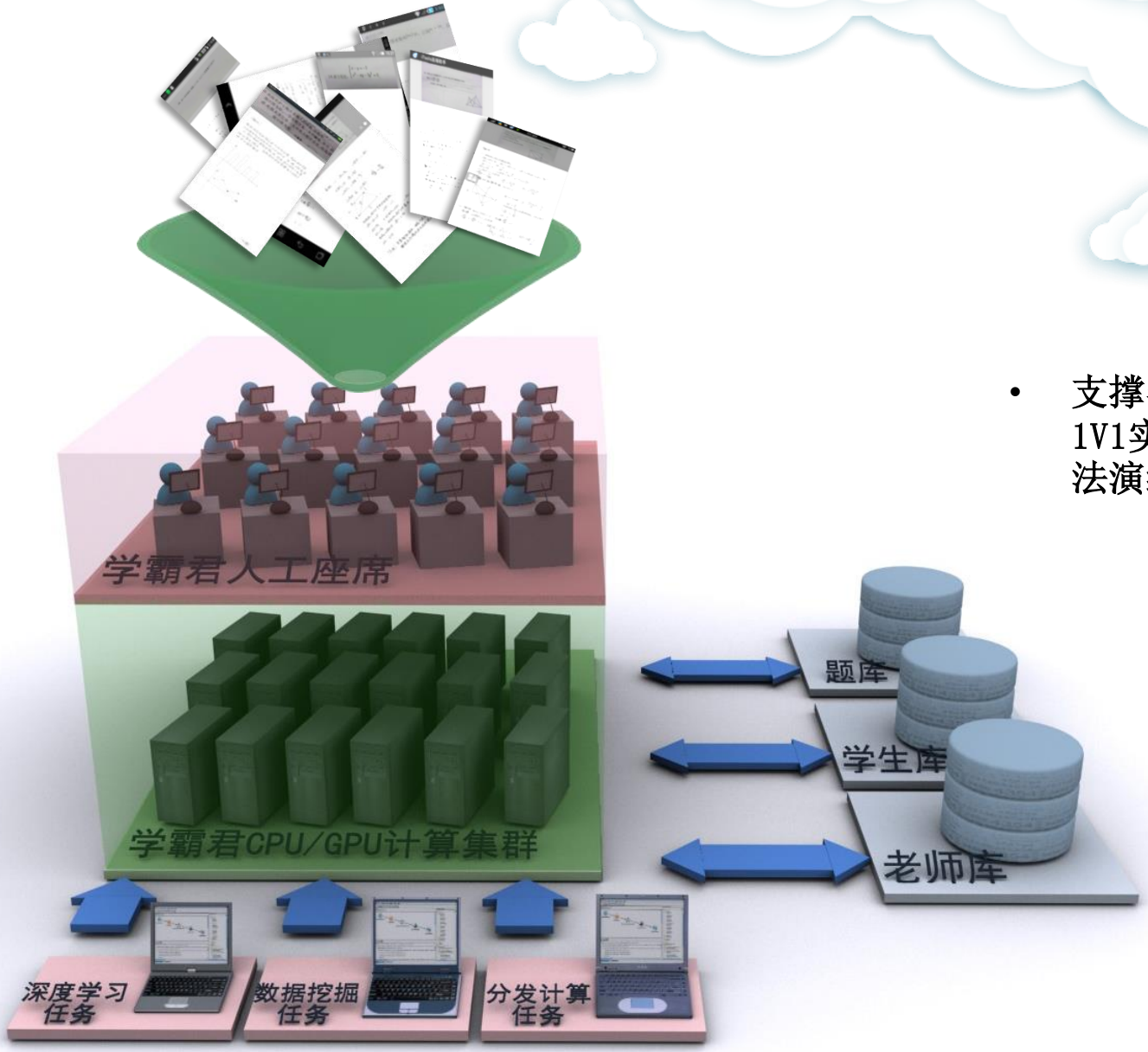
每一代对应的最优排班策略
预估收益比上一代有所优化



- 高效智能调度的价值——拉快数据爬坡



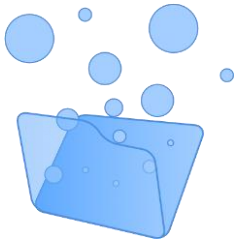
学霸君下一个数据采集目标：1000万高质量1V1视频样本



- 支撑学霸君拍照搜题、1V1实时答疑等业务的算法演练场

目录

- 学霸君的教育业务是以数据及分析为支撑的



收集

图像识别

手写识别

文档布局分析

分析

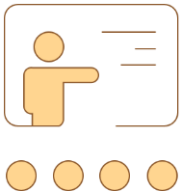
ML/DL（机器学习/深度学习）

NLP

DM

多维标签题库

知识图谱



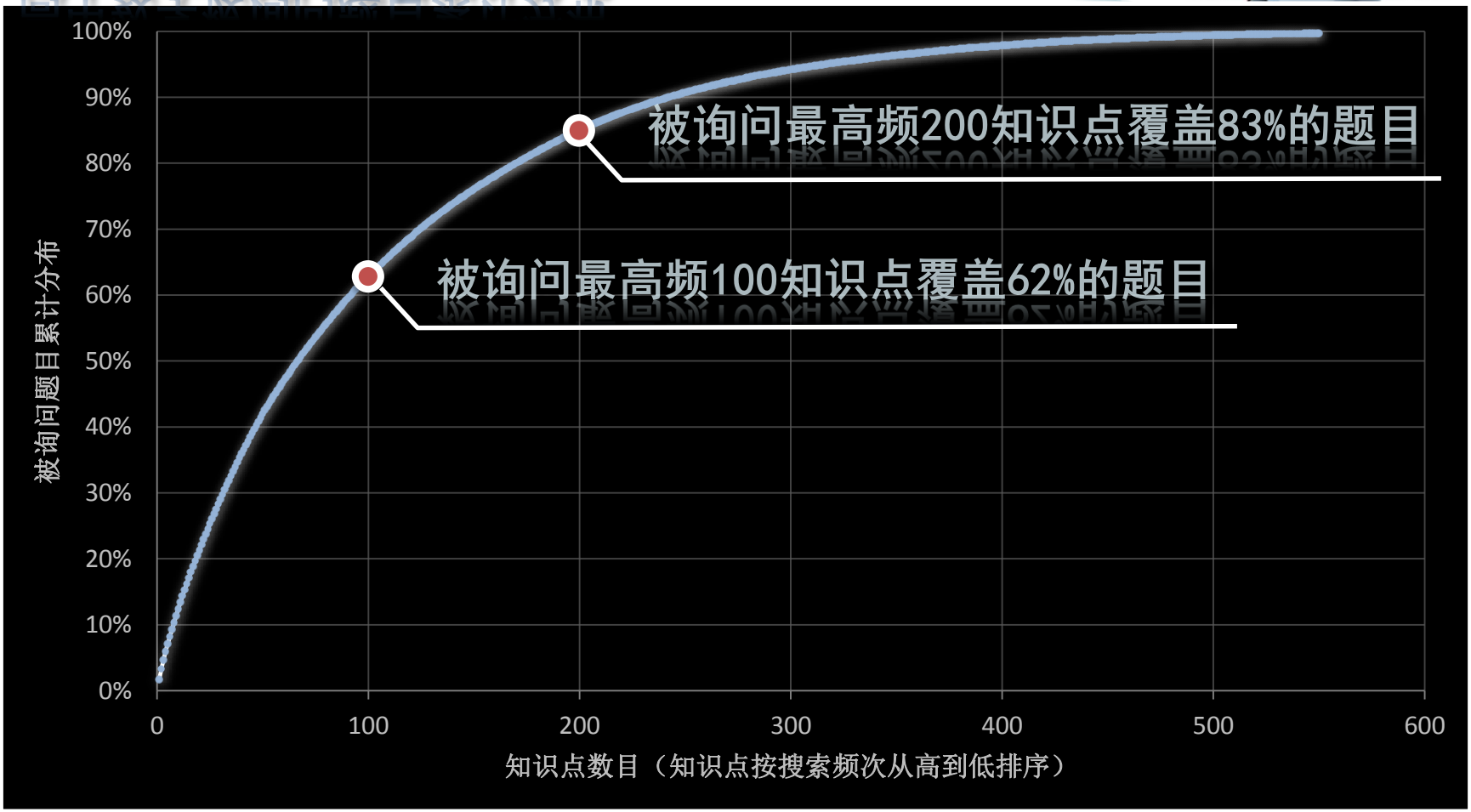
培训

实时答疑

学习内容推荐

高效的数据可以优化教育，有的放矢，学习贵在精而不在多

高中数学被询问题目累计分布



Thank You!