



# 微软人工智能平台方案

微软亚洲互联网工程院

2019.12

# 微软亚洲互联网工程院 (STCA) 速览

高精尖人才与技术  
资源雄厚  
值得信赖  
全球化

智能化浪潮=算法+数据+算力  
全球领先的感知认知情感兼具的人工智能技术能力&专利  
基于搜索引擎的海量数据和数据挖掘实战经验  
产品化能力  
多语言，跨语言，多平台，跨平台的实施能力



## 我们的团队：

成立于2005年  
1900+软件开发工程师&产品经理  
1000+人工智能专业技术人员  
微软最大的海外研发团队  
分布北京，苏州，东京&台北

## 我们开发的全球产品：

必应搜索引擎与广告平台(Bing & Bing Ads)  
微软智能云平台Azure人工智能认知服务  
微软 Microsoft 365(Windows & Office)  
微软小冰(Xiaoice)  
微软小娜(Cortana)

## 我们的核心技术：

计算机视觉  
计算机语音  
自然语言处理  
知识图谱  
深度学习  
大数据处理  
搜索&广告

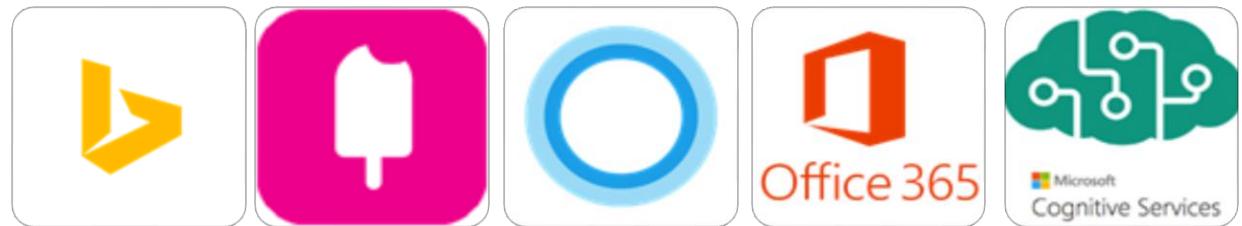
## 我们的中国市场人工智能产品：

微软小冰 - 用于多商业场景下的虚拟对话机器人（智能问答系统与平台/智能客服/智能助手/智能陪伴）  
人工智能内容自动创作（故事，诗歌，以及绘画设计，歌曲创作等）  
智能语音助手（手机，平板，设备，家居等）  
可视化知识图谱及问答  
舆情监测  
计算机语音合成

# STCA自然语言处理团队

微软亚洲互联网工程院自然语言处理团队(STCA NLP Group)  
是微软自然语言处理技术在全球范围内的核心团队。

- 我们负责的人工智能及自然语言处理平台为所有的微软旗舰产品提供服务与解决方案，包括：
- **Microsoft 365 (Office, Windows等):**  
Substrate tool, Search, Speller, Word breaker,
- **必应搜索引擎(Bing):** Relevance, QnA
- **小冰 (Xiaoice):** 1<sup>st</sup> and 3<sup>rd</sup> party KBQA and LU
- **小娜 (Cortana):** LU models
- **Azure认知服务(Cognitive Service):** LUIS models
- 我们的团队主要分布在北京和苏州



人工智能与研究事业部  
AI + Research

体验与设备事业部  
Experiences + Devices

云与人工智能事业部  
Cloud + AI

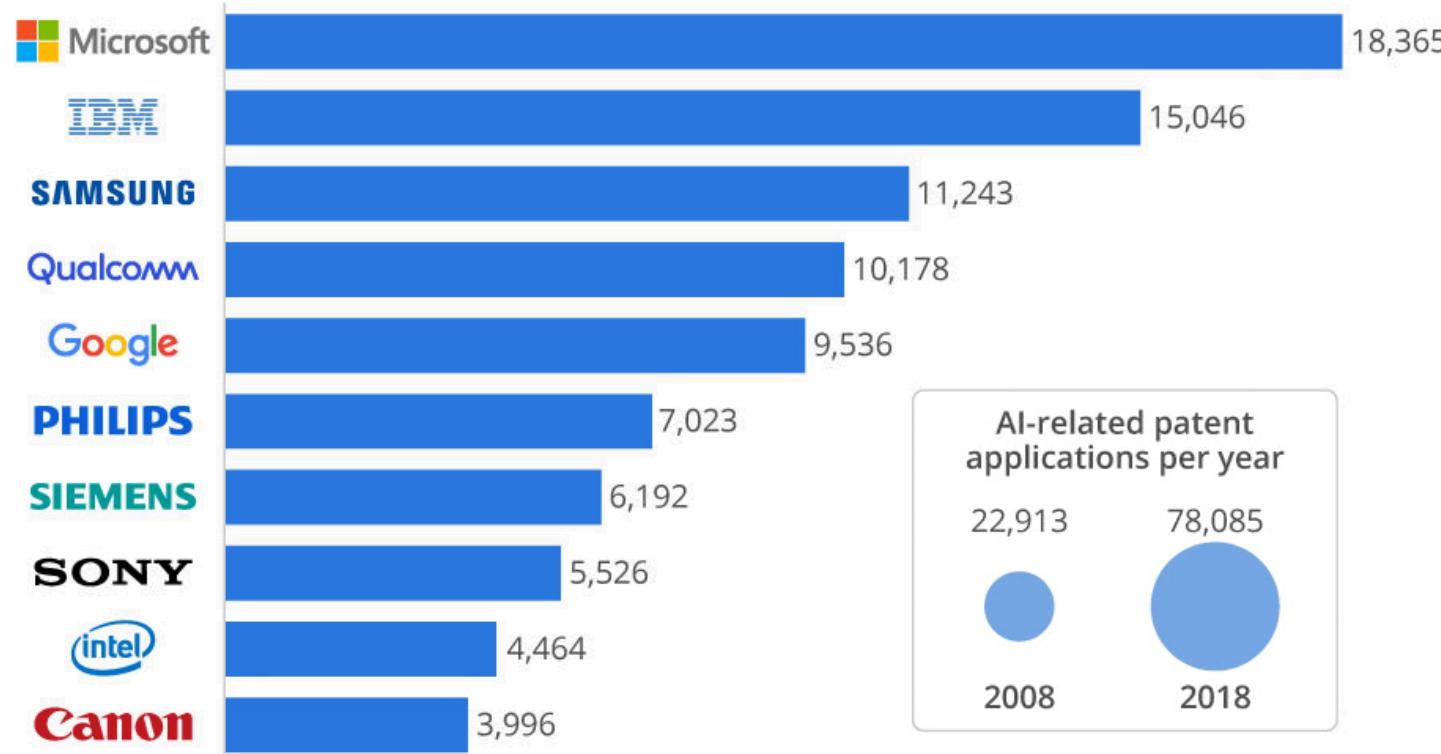
微软亚洲互联网工程院自然语言处理团队

STCA NLP Group

# 人工智能算法和平台技术

## The Companies With the Most AI Patents

Companies with the most artificial intelligence-related patents\*



任务	专利数	任务	专利数
中文分词	8	命名实体识别	7
词性标注	3	语句分类	14
字词向量分析	4	语句分析	3
n-gram分析	1	依存文法分析	3
同义词分析	4	情感分析	8
词频分析	1	文本分类	6
未登录词识别	3	文本聚类	1
语义归一化	2	段落分析	3
文本纠错	16	.....	

任务	专利数	任务	专利数
知识图谱	15	分布式系统	11
对话系统	7	大数据	5
搜索引擎	87	微服务	1

# 微软 AI & NLP 平台

- 13 个不同种类的开发平台
- 84 个预置应用及工作流（实验模板）

## • 必应 Bing

- 50+ 上下文无关语法(CFG)组
- 100+ CFG规则
- 30+ 概率上下文无关文法 (PCFG)

## • 小冰 Xiaoice

- 旅游、诗歌、明星领域知识图谱

## • Azure 认知服务-自然语言理解(LUIS)

- 预定义模型 – 7种语言

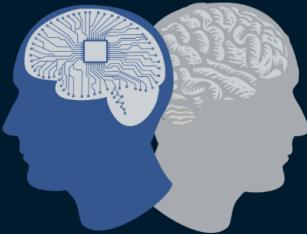
### Recent EQnA Self-Service

Get started with the latest EQnA answer on Bing by walking through the End-to-End workflow.



### Self-Service

Create a powerful NLU model! Give us a few sample utterances, play with our [vending machine], and you'll be all set.



### Creating experiences for Bing

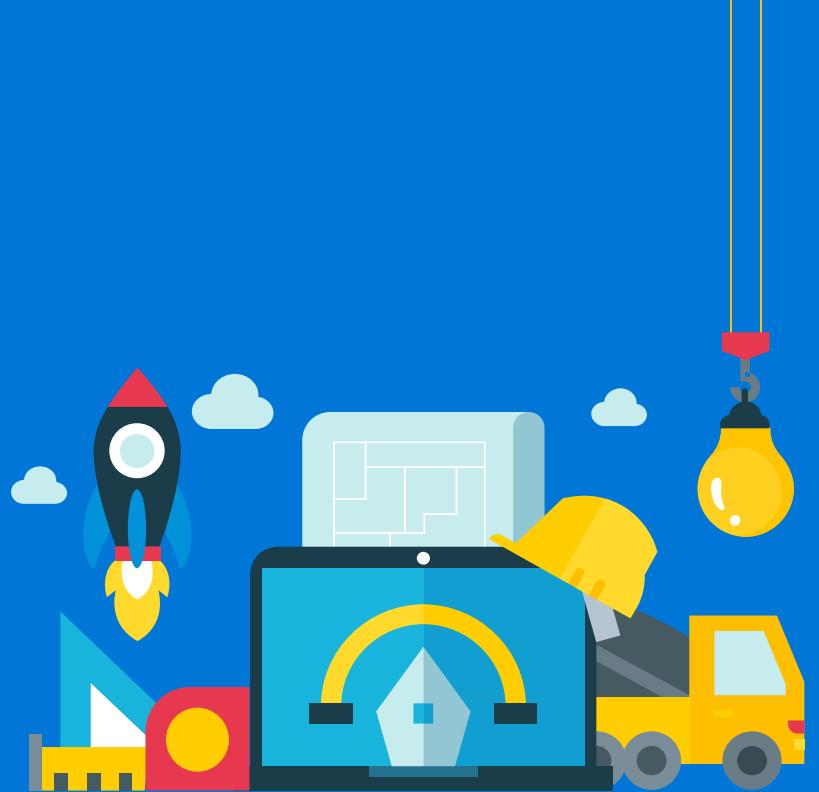
Use LUIS models to understand the user intent and provide the right answer.



### Creating experiences for Cortana

Use LUIS models for your skills.

- ✓ 人工智能总体解决方案
- ✓ AI+NLP平台核心功能
- ✓ NLP中台介绍
- ✓ 应用搭建示例



# 谁会需要人工智能平台？



## 企业，政府单位...

内部有各种各样AI解决方案，互相割裂，没有一个统一的入口

## 企业，高校，研究所...

内部有AI自研能力，但没有一个统一的开发部署平台

## 高校，孵化器，园区...

希望培植AI产业，但不知从何入手

## 传统企业，政府单位...

面临数据孤岛、存储计算能力瓶颈、缺乏统一规划实施等问题  
但AI能力不足，但希望能拥抱AI，快速上手

# 解决哪些企业痛点？

开发人员成本昂贵



技术门槛高



企业数据出网  
面临合规与安全风险



各AI解决方案，互相割裂  
没有一个统一的入口



技术黑盒子，无法修改



可拓展性，可复用性差



无法与企业自建基础设施  
深度集成



技术迭代快，淘汰速度快

# 微软人工智能平台可支持应用



- **自然语言处理:**

去停词 命名实体识别 情感分类 新闻分类 句法分析  
文本多分类 (中文) 分词 主题分析 摘要生成  
问答对匹配 词槽标记 知识蒸馏 文本向量化

- **计算机视觉:**

OCR 图像识别 视觉搜索

- **计算机语音:**

智能客服 语音质检 智能对话设备 智能外呼 语音翻译  
会议转写 语音评分 语音内容生产 内容合规检查 车载语音

- **其他:**

收入分析 房价预测 销量预测 产量分析 流失率预测  
商品推荐 用户画像 股价分析

# Carina核心功能 - AI 平台 + NLP 中台

上层应用

智能系统

问答系统

智能客服

智能搜索

舆情监测

物体监测

RPA

视频处理

声纹定制

一键朗读

图像生成

机器翻译

能力

## 自然语言处理 NLP

基于搜索的问答系统 IR-QA

基于知识库的问答系统 KB-QA

推荐系统  
Recommendation

搜索引擎  
Search Engine

知识图谱  
Knowledge Graph

对话管理系统  
Dialogue System

自然语言理解  
NLU

文本分析  
Text Analytics

## 计算机视觉 Computer Vision

人脸识别  
Face Recognition

OCR

情绪识别  
Emotion Detection

视觉搜索与辨识  
Visual Search and Recognition

物体识别  
Custom Vision

## 计算机语音 Speech

语音识别  
Speech Recognition

语音定制  
Speech Customization

声纹识别  
Voice Recognition

语音合成  
Neural TTS

情绪识别  
Sentiment Recognition

语音评测  
Speech Rating

机器翻译  
Machine Translation

唤醒词定制  
Keyword Spotting

说话人识别  
Diarization

平台

## 人工智能平台 AI Platform

训练 Training

部署 Serving

资源调度 Resource Management

托管服务 Hosting

算法仓库  
Gallery

大数据平台  
Big Data

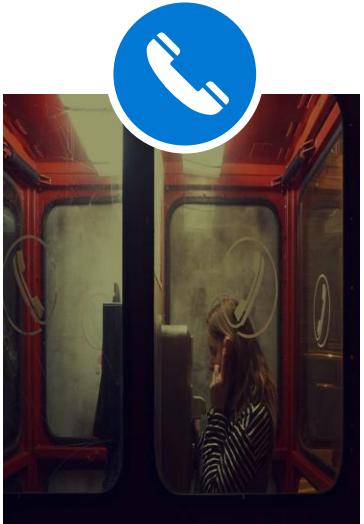
标注平台  
Labelling

集群管理平台  
Cluster Management

运行监控平台  
System Monitoring

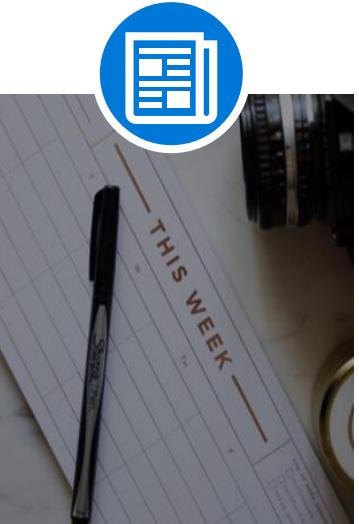
任务编排平台  
Job Orchestration

# 5大能力应用场景



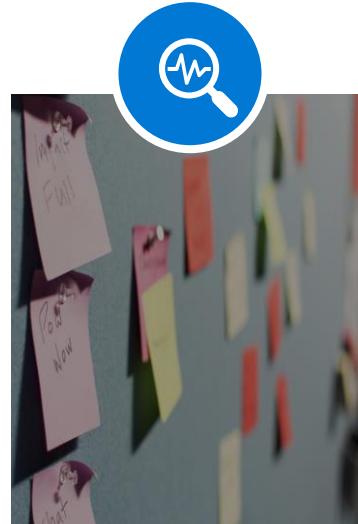
## 对话系统

- ✓ 语音+语义理解
- ✓ 智能客服机器人，智能穿戴设备&语音助手，智能家居



## 舆情监控

- ✓ 故事性追踪，舆情影响力分析，相关搜索词分析，受众画像，简报生成



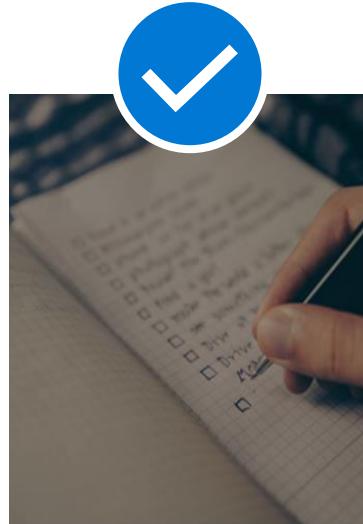
## 智能搜索

- ✓ 模糊搜索，商品搜索，企业内资源信息查询，内容匹配，搜索结果智能排序



## 文本处理

- ✓ 文件信息提取，VOC客户意见分析，合同文档解析，文档分类



## 推荐系统

- ✓ 行为学习排序，个性化商品推荐，相关推荐，用户画像报告

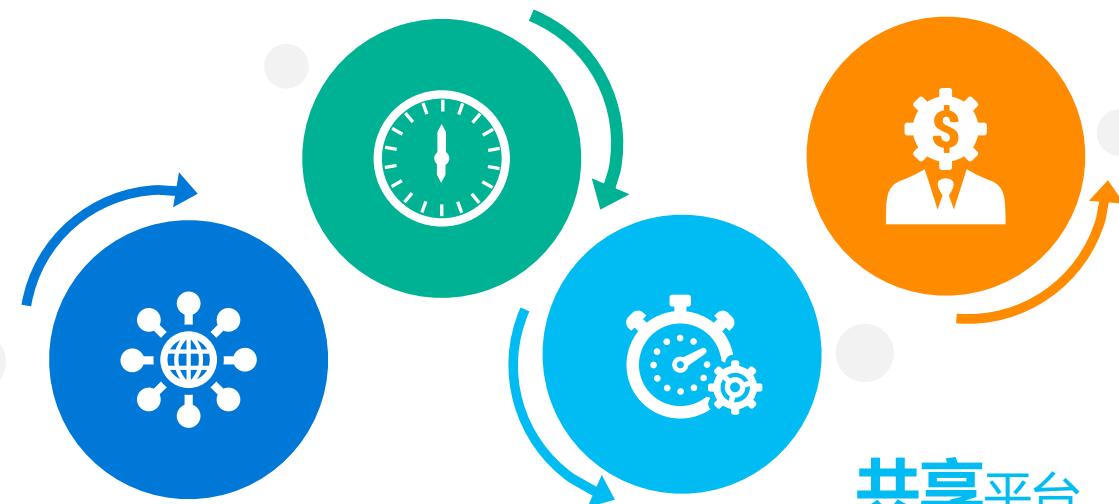
- ✓ 人工智能总体解决方案
- ✓ AI+NLP平台核心功能
- ✓ NLP中台介绍
- ✓ 应用搭建示例



# 什么是人工智能基础平台（AI平台）？

## 部署平台

提供机器学习模型的上线部署，来**使用**人工智能



## 训练平台

提供各种机器学习训练能力，来**获得**人工智能

## 调度平台

为实验或服务调配资源，来**支撑**人工智能

## 共享平台

算法模块，实验模板，预训练模型...  
来**复用**人工智能

# AI平台应该提供哪些功能？

## 预训练模型

使用预训练模型加速开发流程

## 流行框架

构建各类机器学习，深度学习任务

## 开发工具

充分发挥开发团队的创造力，提高生产力

## 强大的算力

加速价值实现

## 灵活的部署方式

私有化，公有云和边缘计算



图像



语音



语言



搜索



Pytorch



TensorFlow



Keras



Onnx



大数据计算



人工智能训练



人工智能发布



CPU



GPU



FPGA



私有化



公有云



边缘计算

# NLP中台在总体架构中的位置

应用

对话系统

舆情监控

文本处理

智能搜索  
(信息, 产品等)

智能推荐  
(信息, 产品等)

中台

对话管理系统  
Dialogue System

问句重写 模型关联  
技能管理 任务定义  
用户会话管理 调用关联

基于信息检索的问答  
IR-QA

问答对导入  
关键词抽取  
关键词扩展

基于知识库的问答  
KB-QA

语义引擎  
规则引擎  
词典和模板管理

知识图谱  
Knowledge Graph

图数据库  
本体关系  
查询服务

搜索引擎  
Search Engine

问句重写 问句理解  
选取服务 排序服务  
现状生成 调用关联

推荐系统  
Recommendation System

用户画像建立  
工作流定义  
查询服务

自然语言理解 Nature Language Understanding

单句语料管理 语料模式管理 模型训练  
模型测试 模型更新 版本管理 内置词槽  
模型导入导出 模型发布

内容分析 Text Analytics

索引生成 内置词典 白名单标记 文本理解引擎  
标注结构定义 文本预选 规则标记 统计标记 数据规范化

平台

人工智能平台 AI Platform

训练 Training

部署 Serving

算法仓库  
Gallery

资源调度 Resource Management

托管服务 Hosting

大数据平台  
Big Data

标注平台  
Labelling

集群管理平台  
Cluster Management

运行监控平台  
System Monitoring

任务编排平台  
Job Orchestration

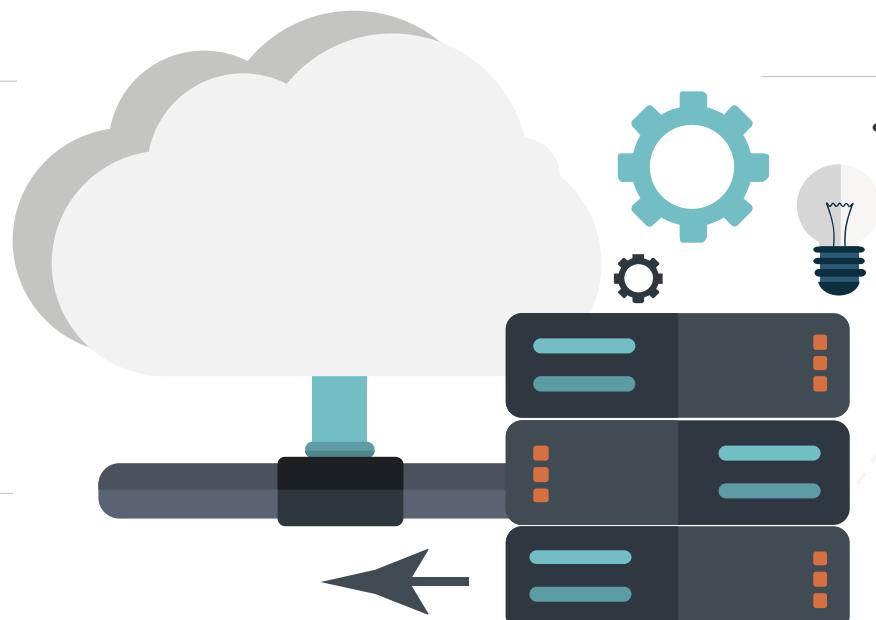
# 企业人工智能的必备工具

## 兼容并蓄

- 与企业现有数据平台与计算引擎 **便捷整合**
- 支持当下所有主流大数据平台和AI架构, e.g.  
Spark, Kubernetes, Linux, TensorFlow ...

## 轻松上手

- 一站式**AI模型设计, 训练, 评估, 生成AI模型
- 一键式**发布AI模型, 上线部署解决业务问题



## 灵活扩展

- 根据企业自身的需求**增减**功能模块和算法模块
  - 尖端的自然语言算法仓库**模块化调用**

## 表现稳定

- 通用的**深度学习&机器学习的服务平台
  - 微软内部上线**8年**
  - 每天提交超过**20,000个**实验

大规模机器学习平台, 覆盖机器学习全流程

# 解决技术人员的实操痛点

## 原有机器学习工作方式

- ✗ 申请资源（机器，IP，存储）
- ✗ 配置基础环境（多版本）
- ✗ 安装第三方算法库
- ✗ 准备数据
- ✗ 选择机器学习模型
- ✗ 单机模式，无法扩容
- ✗ 计算资源抢占
- ✗ 配置复杂
- ✗ 发布流程繁琐易出错
- ✗ 系统监控及维护复杂

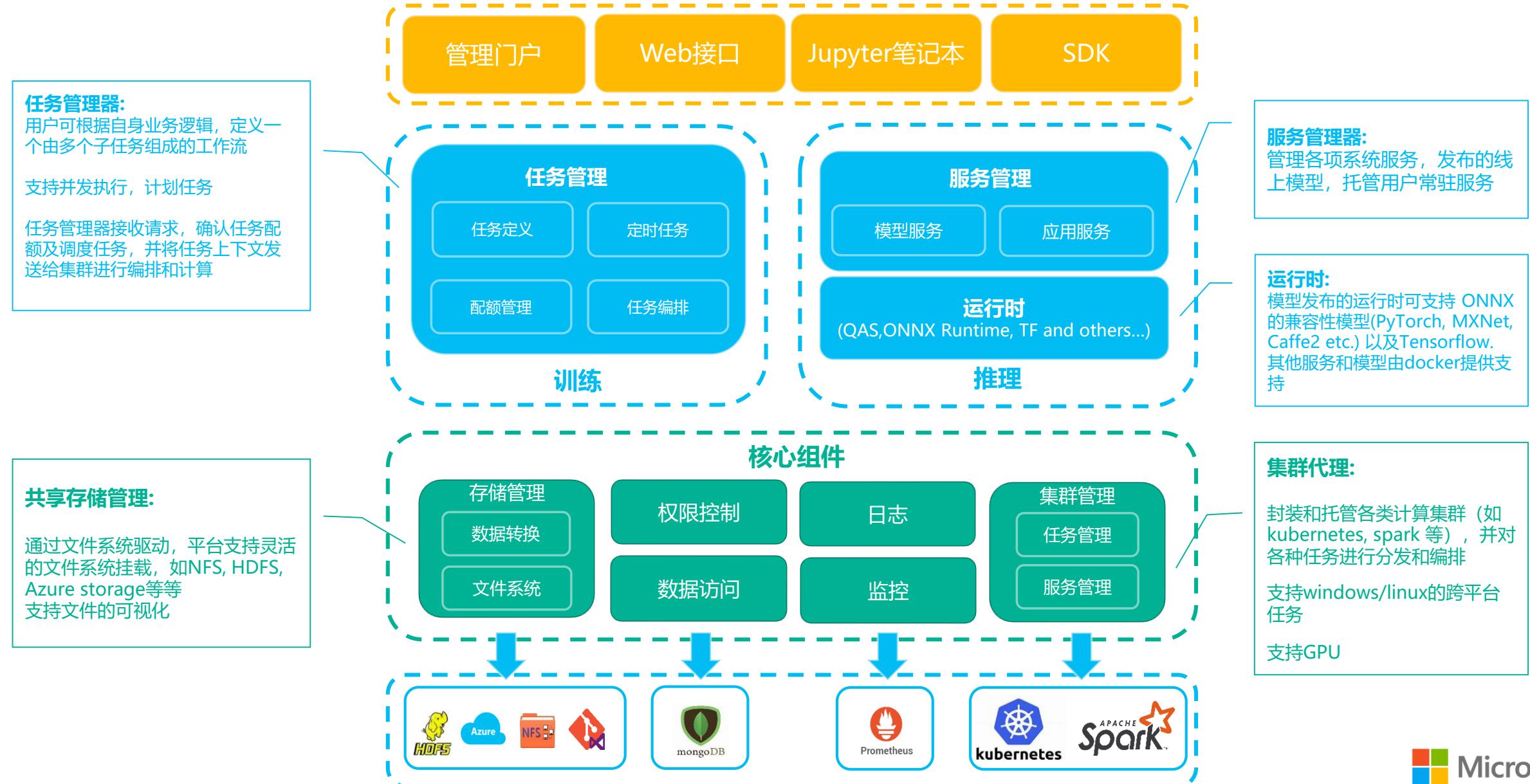


## 实施AI平台后的机器学习工作方式

- 一键环境准备
- 机器学习模型复用
- 机器学习全生命周期管理
- 精准实时的平台监控



# AI平台的架构



# NLP算法仓库



经过微软长期实践得到的各类算法集成化、模块化、积木化，可以随时调用，降低算法技术门坎，提高复用率。

# NLP算法仓库

**NLP Gallery**

Search task by name, task type and description

**Extract Summary** Chinese▼

① Create a summary with the major points of the original document.

Hanlp  Default  1.0

习近平指出，中美在促进世界和平和繁荣方面共同肩负着重要责任。一个良好的中美关系符合两国人民根本利益，也是国际社会的普遍期待。合作是中美双方最好的选择。双方要把握好中美关系发展的大方向，推动两国关系长期健康稳定发展，更多更好地造福两国人民和世界各国人民。特朗普赞同习近平对两国关系的评价。特朗普表示，美中关系十分特殊、重要，我们两国都是有重要世界影响的国家，双方保持良好合作关系对两国和世界有利。美方愿同中方通过协商增进两国合作，并就双方存在的问题积极探讨对双方都有利的解决办法。两国元首同意继续通过各种方式保持密切交往，共同引领中美关系发展方向。双方将

Count of Result : 3

Example 1

Hanlp ▾ — Default ▾ — 1.0 ▾ ⓘ

双方保持良好合作关系对两国和世界有利，美方愿同中方通过协商增进两国合作，双方要把握好中美关系发展的大方向

Result Info

The screenshot shows the NLP Gallery interface. On the left, a sidebar lists various NLP tasks: 摘要提取, 命名实体识别, 情感分析, 关键词提取, 语句相似性评判, 智能回复, 否定检测, 相似短语检索, 相似语句生成, 机器阅读理解, 主题分类, 多语言阅读理解, 语句扩展, 文本分类, and 短语扩展. The '摘要提取' task is highlighted with a teal background. The main panel shows the 'Extract Summary' task selected. It includes a search bar, a toggle switch, and a detailed description of the task: 'Create a summary with the major points of the original document.' Below this is a configuration section for 'Hanlp' (selected), 'Default', and '1.0'. A large text area displays a sample document in Chinese. At the bottom, there are buttons for 'Example 1' and 'Refresh All', and a 'Result Info' link.

# 案例一：券商A-人工智能基础平台

目标：为某头部券商A数据处理和离线训练赋能，实现**人工智能+云计算转型**。



## 人工智能基础平台：

- 稳定运行一年
- 四套环境
- 每天上千个任务提交

## 面向业务人员：

- 信息技术部 (200+人)
- 业务部门 (100+人)：证券投资、固定收益、互联网金融

## 应用：

- 统一调度
- 数据科学开发
- 金融分析
- 大数据量化分析：股票策略研究 - 每日交易复盘从4小时缩短为15分钟，每日成交量100亿规模

# 案例一：券商A-人工智能基础平台



01

## 股票策略研究-交易复盘

在Carina平台上创建定时任务。每个交易日闭市后，平台自主地将用户存储的高频因子文件导入，并通过平台内建的spark集群对因子文件和高频因子库进行比对和处理，得出当前交易日的复盘结果。

同时，可对高频因子完成追加，获取，更新，删除等操作。借助平台高效的分布式调度和计算能力，与用户大数据平台深度结合，整个过程仅耗时15分钟左右，且资源复用能力极强。



02

## 大数据平台团队-统一编程平台

在实际开发场景中，研发人员往往需要在特定技术架构和环境下开发和调试。借助Carina平台的 Jupyter Nodebook 和 Python engine服务，用户可以在逻辑独立的运行环境中完成深度学习任务的逻辑实现，进行远程调试，并访问共享容器仓库将代码打包和发布。

# 案例二：汽车制造场景应用

问题一



## 数据孤岛

- 例如售后、生产、采购数据分别存储于各自业务系统中，没有进行有效的整合，导致索赔问题出现的时间段、厂家、批次等没办法做精准定位和分析；

问题二



## 存储计算能力瓶颈

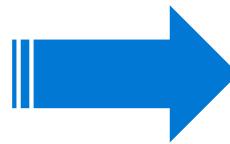
- 存储性能：现有资源存储空间有限，伴随车联网、工业互联网类数据量急剧增长，需要由扩展性强的数据湖来承载海量数据存储；且传统RDBMS不支持非结构化数据存储，不支持无限扩展
- 计算性能：计算资源较为分散，缺乏有效的管理平台，当出现大规模大规模计算场景时，无法统一调配资源

问题三



## 缺乏统一规划实施

- 内部有各种各样AI解决方案，互相割裂，没有一个统一的入口
- 资源调度效率低，应用体验差



## 数据科学平台



数据预处理



用户标签画像



NLP解析



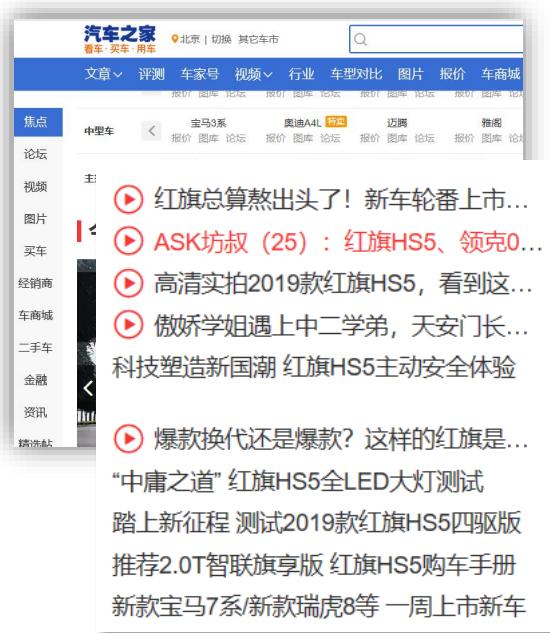
行业知识库

通用人工算法库

数据科学基础平台

AI服务

# 案例二：核心成果



01

## 数据

通过各网络渠道应用关键字抓取、关键字扩展等AI模块汇总有效内容。

02

## 洞察

通过各网络渠道应用关键字抓取、关键字扩展等AI模块汇总有效内容。经语义引擎、本体关系、用户画像建立、索引生成、文本预选、统计标记、数据规范化等AI模块处理后，生成各类报告

03

## 方案

参考所得的有效报告，进行业务布局，优化营销效果，进而提高销量、客户满意度等关键指标。

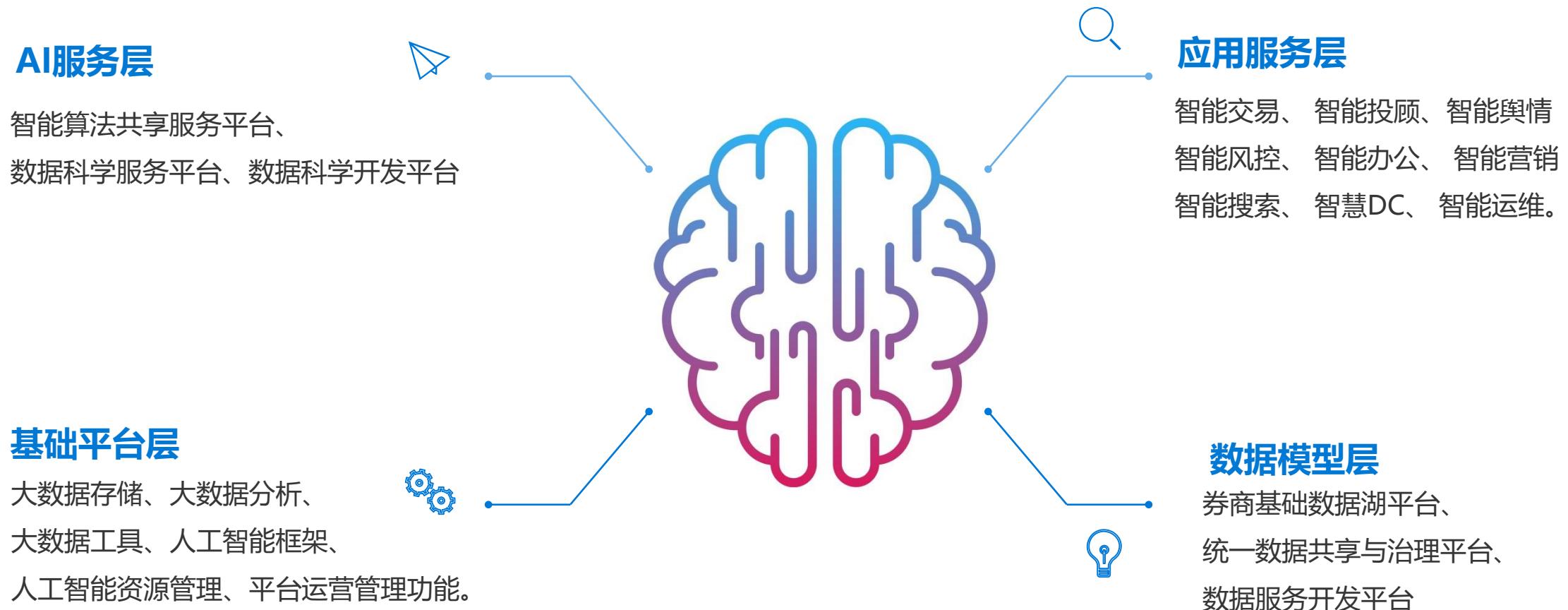
**VOC报告** - 评论分类、问题溯源、业务热点、人群分析、趋势报告

**一群三库** - 情景库、对标库、产品库、用户画像、人群画像

**DMS助手** - 产品优势、产品负面、业务热点、人群分析趋势报告

# 案例三：券商B-人工智能智慧大脑

某头部券商B与微软合作建设人工智能智慧大脑，为其打造智能化证券业务服务核心平台赋能，实现**智能化转型**。



# 案例三：券商B-智能搜索应用场景



办公搜索



智能选股



舆情搜索



行业咨询搜索



事件驱动分析



尽职报告生成



客户行为分析



智能推荐

- ✓ 人工智能总体解决方案
- ✓ AI+NLP平台核心功能
- ✓ NLP中台介绍

**自然语言理解 (Natural Language Understanding)**

内容处理 (Text Analytics)

知识图谱及问答 (Knowledge Graph & QA)

基于检索的问答 (Information Retrieval based QA)

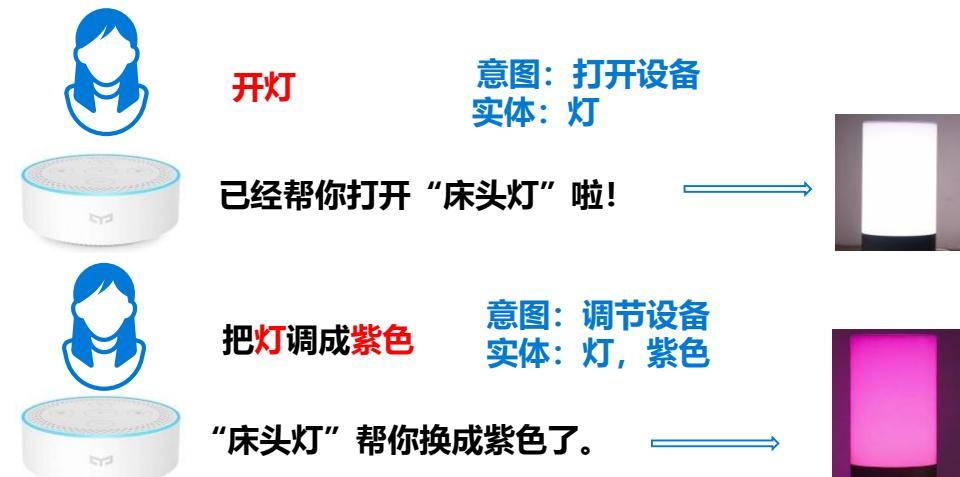
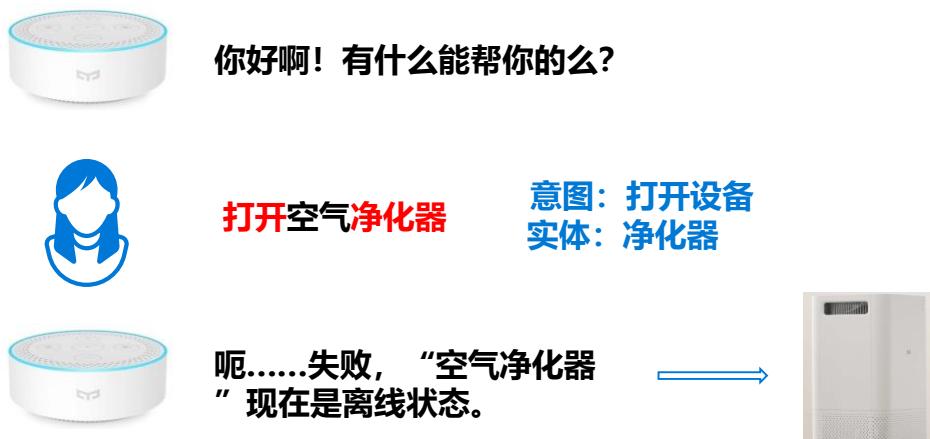
对话系统 (Dialog System)

智能搜索与推荐 (ReSearch)

- ✓ 应用搭建示例



# 自然语言理解基本概念



# 自然语言理解 (NLU) 工作台

帮助开发者快速地创建一个语言理解模型，把人类自然语言转化为机器可以理解的形式



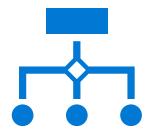
## 主要应用

- 智能问答
- 搜索，查询
- 对话系统



## 主要组件

- 词法、句法、语义分析
- 意图分类
- 词槽标注
- 命名实体识别
- 多轮信息继承



## 核心功能

- 语料管理
- 规则管理
- 语言模型训练
- 数据扩展
- 模型版本控制

# 工作台操作界面

语句样本 ② 语句模板 ② 工具 测试

0/11 个语句样本

添加一个语句样本

语句 意图 来源

苏州明天的天气怎么样 trip\_cn check\_weather Seed

在巴黎找个靠近机场有wifi的酒店 trip\_cn find\_hotel Seed

查询天气 trip\_cn check\_weather Seed

上海天气怎么样 trip\_cn check\_weather Seed

北京的天气 trip\_cn check\_weather Seed

语句样本

意图 ② 意图

check\_weather  
find\_hotel  
book\_flight

词槽 ② 词槽

from\_city  
city  
date

语句样本 ② 语句模板 ② 评估 测试

0/1 个语句模板

用#开头来引用词典非空的词槽或通用短语

批量编辑 | 训练 | 测试

语句 意图 来源

预定从 #city 到 #city 的机票 book\_flight Seed

语句模板

在这里输入您的测试语句 测试

预定从纽约到洛杉矶的机票 1.00

trip\_cn Conf: 1.00 编辑

book\_flight Conf: 1.00 编辑

预定从 纽约 到 洛杉矶 的机票

- ✓ 人工智能总体解决方案
- ✓ AI+NLP平台核心功能
- ✓ NLP中台介绍

自然语言理解 (Natural Language Understanding)

内容处理 (Text Analytics)

知识图谱及问答 (Knowledge Graph & QA)

基于检索的问答 (Information Retrieval based QA)

对话系统 (Dialog System)

智能搜索与推荐 (ReSearch)

- ✓ 应用搭建示例



# 内容分析处理概览



# 内容分析处理模块 - 文本分类

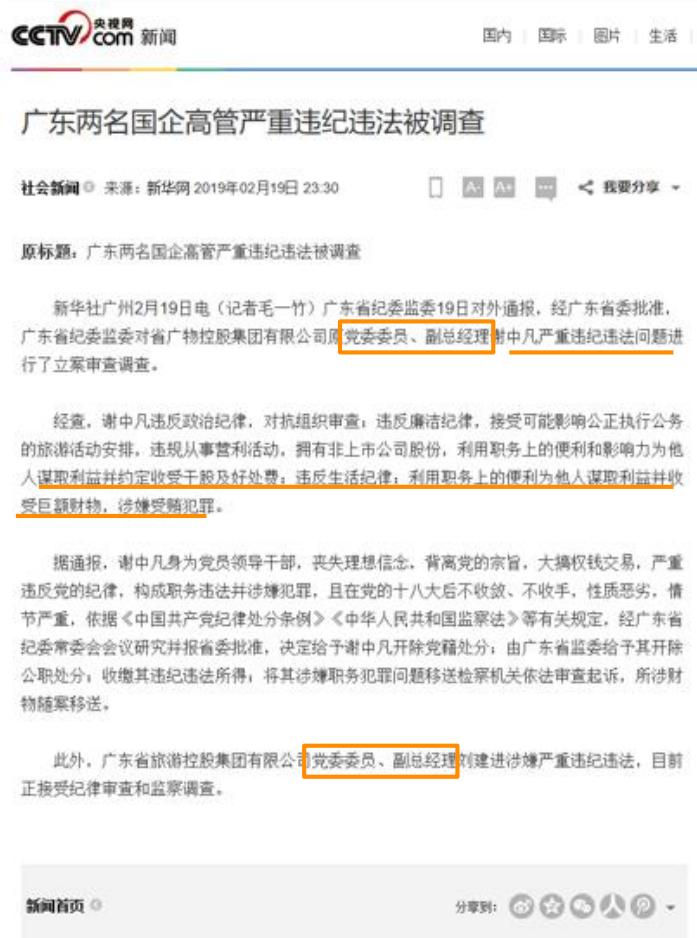
文本分类  
Text Classification

对一段文本根据预定义好的类别进行分类

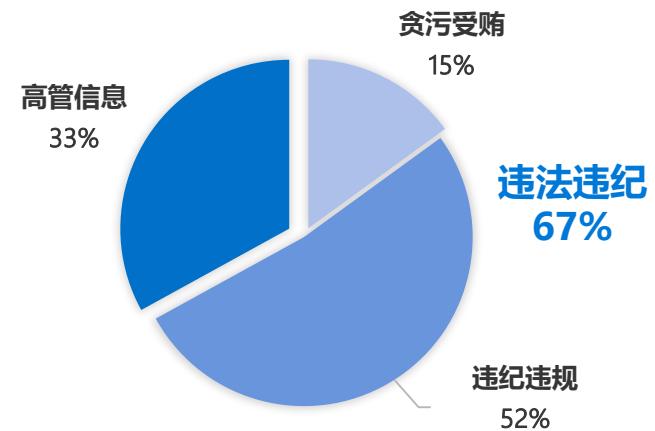
新闻, 研报, 公告

文章分类

- 高管信息
- 人事变动
- 违法违纪
  - 造假欺诈
  - 贪污受贿
  - 违纪违规
  - 垄断
  - 环保处罚
  - 安全事故
  - 司法纠纷
  - 侵权抄袭
  - 偷税漏税
- 公司经营
  - 上市退市
  - 亏损盈利
  - 投融资
  - 收购重组
  - 破产
  - 股权变动
- 信誉成果
- 产品信息



- 违法违纪	score:0.6736842105263158
---- 贪污受贿	score:0.2298634008478568
---- 违纪违规	score:0.7701365991521432
- 高管信息	score:0.3263157894736842



# 内容分析处理模块 - 关键词抽取

关键词抽取  
Keyword Extraction

对一段文本中的关键数字，  
时间，事件等词语进行抽取

新闻，研报，公告

文章信息抽取

类别	关键词
公司名	海尔，360，新浪，美团，阿里，阿里巴巴
产品名	淘宝
组织	保监会，证监会
人	王兴，董明珠
地点	上海，广州，北京
百分数	50%
数额	300万，500股
行业专有名词	短线，大盘，A股，蓝筹，创业板，走势，股指，股东，借壳上市
事件	回落，低开，反弹，增持，冲高，下跌，重组，发行，IPO，违法，违规...

示例：

高鑫零售11月20日公告，阿里以224亿元港币（约28.8亿美元）收购高鑫零售36.16%，收购价格对应高鑫零售估值约619亿港币市值。收购完成后，法国欧尚集团对高鑫持股36.18%，台湾润泰集团仍持股4.67%，未来三方将展开新零售战略合作。

关键字：

高鑫零售，阿里，224亿元港币，28.8亿美元，收购，619亿港币，欧尚集团，持股，36.18%，润泰集团，4.67%，战略合作

# 内容分析处理模块 - 情感分析



商品评价  
舆情监控

## 分类示例

**正面:** 盈利, 高管增持, 买入, 评级上调, 融资成功, 等明显带有正面积极信息的行文, 对公司有利的消息

**中立:** 一般性陈述事件, 不具备导向的事件, 不影响公司经营的事件: 收购/停牌复牌/高管变动, 产品信息, 申请批复, 股权变更等

**负面:** 股价利润下跌, 高管减持, 破产重组, 偷税漏税, 侵权抄袭, 司法纠纷, 各类事故, 行政处罚, 违规违纪等负面新闻

### 受益油价大涨 中石油一季度盈利劲增71%

今年一季度原油价格同比大涨, 直接推升中国石油(601857)业绩。一季度, 中国石油实现归属于母公司股东的净利润324.92亿元, 同比大增71.2%;基本每股收益为0.18元。期内公司上缴的营业税金及附加409.67亿元, 同比提高72.1%, 主要归因于石油特别收益金和消费税增加。1-3月, 中石油的原油平均实现价格为70.01美元, 而去年同期仅为37.1元, 这使得上游业务为主的中石油大为受益。中石油称, 借助国际油价上涨的有利时机, 期内积极组织生产运行, 原油产量比2009年同期提高2.1%, 为2.1亿桶。

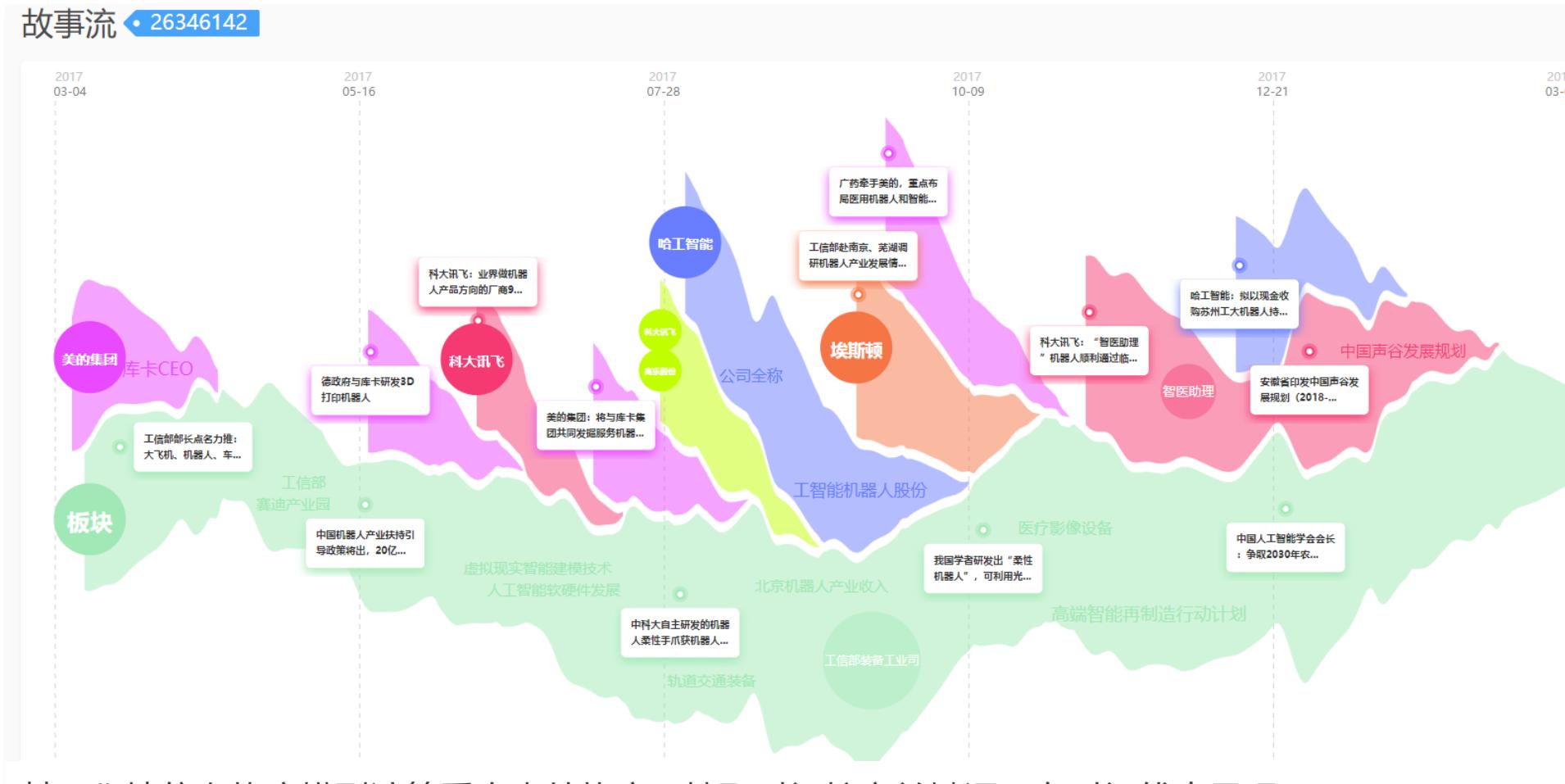
(盈利增加)

### 当当用户信息泄露, 谁之过?

印象里, 京东商城用户被骗的事情好像还没过去多久, 就又有电商平台购物者被骗了。这一次的主角之一, 是正踌躇满志想回归A股的当当网。7月28日, 有当当网用户在当当网下单买书后仅一天, 个人信息就被骗子掌握, 骗子利用详细的订单信息博取用户的信任, 引导自己进入假网站, 输入银行信息后, 被转走约十万元。

(负面报道)

# 内容分析处理模块 - 重要事件热度分析



基于舆情信息热度模型计算重大事件热度，抽取时间核心关键词，在时间线中呈现；

- 支持多公司实体搜索，快速判断公司舆情导向；
- 支持数据源溯源，点击查看原始文章数据；

# 内容分析处理 (Text Analytics) 工作台

从大段文字中抓取有价值的信息



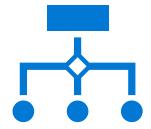
## 主要应用

- 实体抽取
- 关系抽取
- 知识图谱构建
- 舆情监测



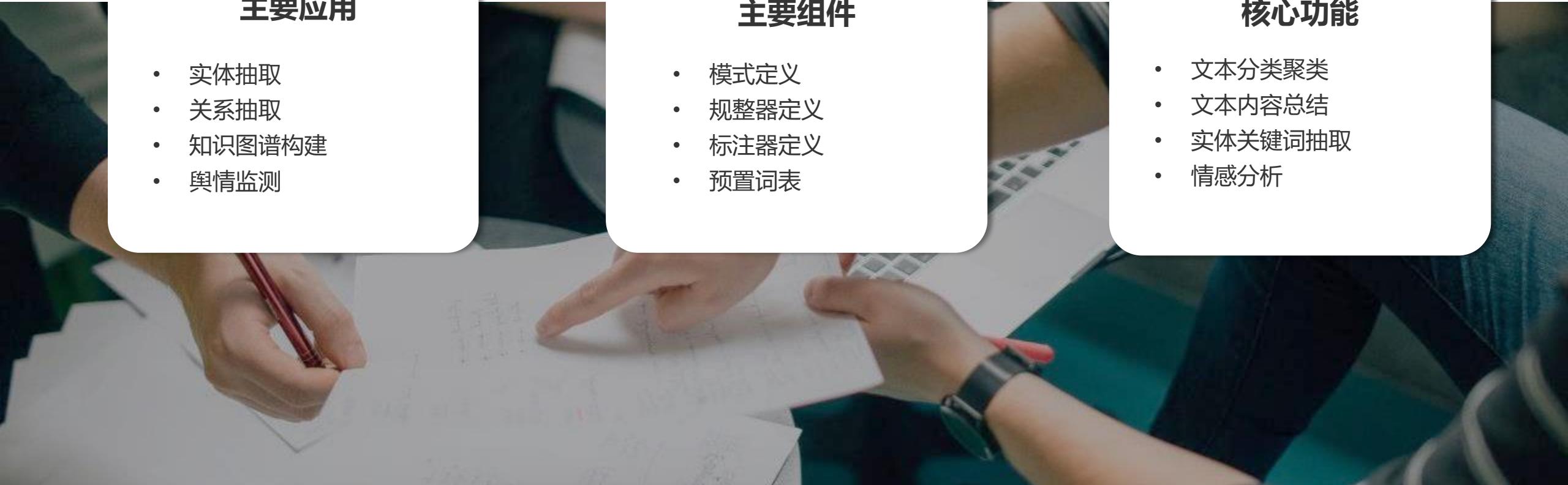
## 主要组件

- 模式定义
- 规整器定义
- 标注器定义
- 预置词表



## 核心功能

- 文本分类聚类
- 文本内容总结
- 实体关键词抽取
- 情感分析



- ✓ 人工智能总体解决方案
- ✓ AI+NLP平台核心功能
- ✓ **NLP中台介绍**

自然语言理解 (Natural Language Understanding)

内容处理 (Text Analytics)

**知识图谱及问答 (Knowledge Graph & QA)**

基于检索的问答 (Information Retrieval based QA)

对话系统 (Dialog System)

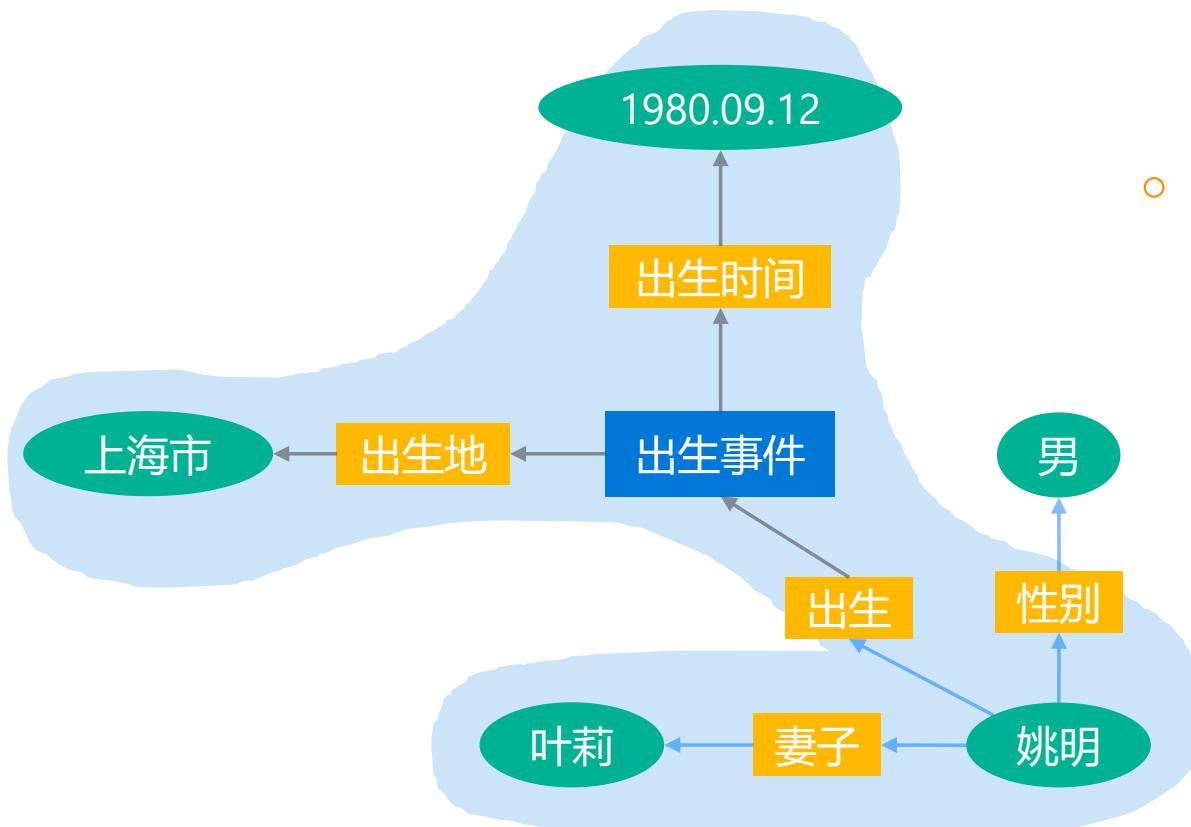
智能搜索与推荐 (ReSearch)

- ✓ 应用搭建示例



# 知识图谱的概念

知识图谱是用点和边组成的图来表示知识的形式



## 知识图谱中的一些常见概念：

- **实体 (Entity)**
- **谓词 (Predicate)**
  - 两个相连实体间的关系
- **复合值类型 (Compound Value Type)**
  - 用来描述一个多属性的事件的虚拟实体
- **事实 (Fact)**
  - 三元组 (Triple), 用来连接两个实体
  - 事件 (Event), 通过复合值类型来连接多个实体

# 知识图谱应用的技术难点

从“关系”的维度深度整合现有数据，形成结构化的知识，进而挖掘实体间显性和隐性的联系。

- **构建**

- 本体定义，特定情况需领域专家参与
- 表格、数据库、半结构化文本的数据难以融合
- 数据的一致性与准确性

实体抽取  
关系抽取 (实体、属性、值)  
属性抽取 三元组

- **存储**

- 支持低延迟、高可靠的管理要求



- **查询**

- 自然语言本身的复杂性、模糊性、多义性

# 知识图谱+问答 (KBQA) 工作台

建立知识库，理解复杂自然语言问句，并基于知识库给出答案



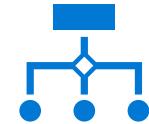
## 主要应用

- 知识存储
- 知识展示
- 知识问答



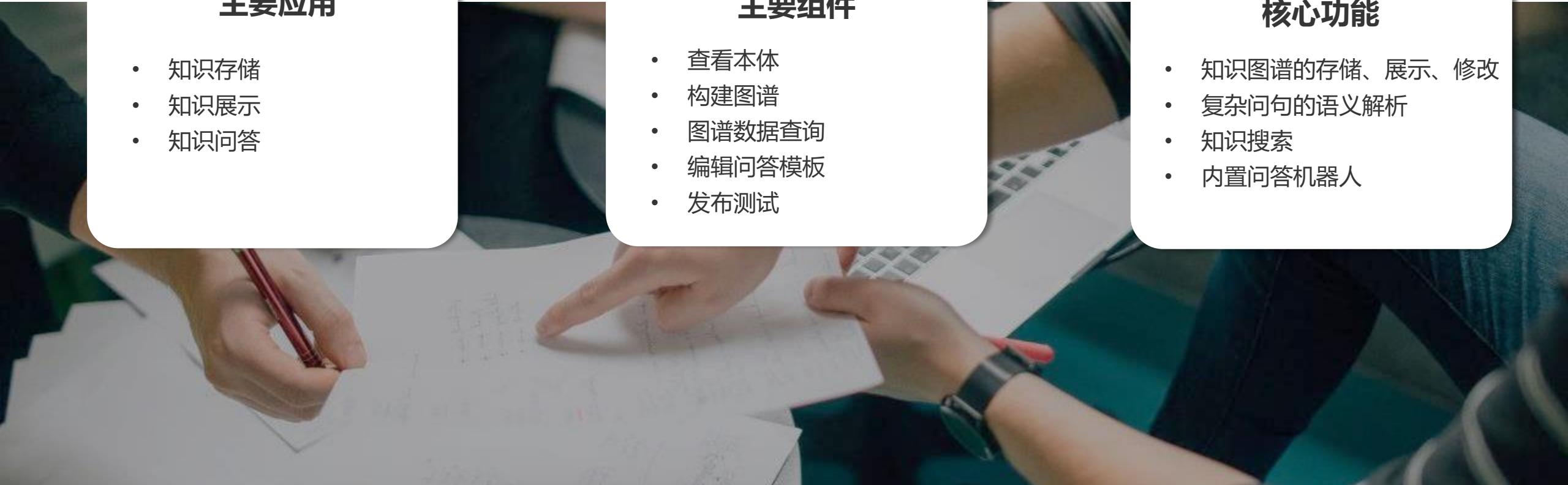
## 主要组件

- 查看本体
- 构建图谱
- 图谱数据查询
- 编辑问答模板
- 发布测试



## 核心功能

- 知识图谱的存储、展示、修改
- 复杂问句的语义解析
- 知识搜索
- 内置问答机器人



# 知识图谱+问答 (KBQA) 工作台

KBQA

领域 汽车 本体结构 数据导入 图谱数据 词典设置 语法模板 公司 党史 地理 诗歌 本体结构 数据导入 图谱数据 词典设置 语法模板 + 添加领域 通用文件 日志数据 服务终端 帮助文档

丰富的知识图谱  
基于丰富的知识图谱，构建智能问答

第一步 查看本体  
通过查看本体结构来理解如何组织和展示数据。

第二步 导入数据  
上传文件导入数据，文件数据内容需符合本体的定义。

第三步 图谱数据  
查看当前图谱数据，图谱数据分为节点、边和属性，点击显示具体的数据。

第四步 更改词典和模板  
查看或修改词典和模板来定制问句

第五步 保存重建与测试  
保存词典和模板配置的更改，通过与内置的测试机器人聊天来进行测试。

第六步 发布与重置  
发布配置好的服务或将服务重置至上次发布的版本。

# 案例一：金融机构知识构建

The screenshot displays the Microsoft Knowledge Platform interface, specifically the '基本信息' (Basic Information) section for Tencent Technology (Shenzhen) Co., Ltd. The page includes the following details:

- 公司名称:** 腾讯科技（深圳）有限公司
- 公司Logo:** Tencent 腾讯
- 成立时间:** 1998年11月
- 创始人:** 马化腾、张志东、许晨晔、陈一丹、曾李青五位创始人共同创立。
- 服务范围:** 深圳市腾讯计算机系统有限公司是中国最大的互联网综合服务提供商之一，也是中国服务用户最多的互联网企业之一。腾讯多元化的服务包括：社交和通信服务QQ及微信/WeChat、社交网络平台QQ空间、腾讯游戏旗下QQ游戏平台、门户网站腾讯网、腾讯新闻客户端和网络视频服务腾讯视频等。
- 上市信息:** 2004年腾讯公司在香港联交所主板公开上市（股票代码00700），董事会主席兼首席执行官是马化腾。
- 重要事件:** 2017年11月23日，腾讯公司与香港铁路有限公司正式签署合作协议，双方就微信支付和WeChat Pay HK（微信香港钱包）在港铁的移动支付业务展开合作。
- 股价:** 2017年7月17日，《福布斯富豪榜》发布，马化腾以净资产324亿美元排名第23位。2017年8月7日，腾讯股价盘中再创历史新高320.6港元，马化腾身家361亿美元成为中国首富。
- 注册信息:** 注册工商号: 440301503270924, 组织机构代码: 71526726X
- 董事会:** 法人代表: 马化腾 (职务: 腾讯公司董事会主席兼首席执行官, 出生: 1971年10月29日 (46岁))  
法人代表: 刘炽平 (职务: 腾讯公司执行董事兼总裁, 出生: 1973年)  
法人代表: Charles St Leger Searle (职务: 非执行董事, 出生: 1973年)  
法人代表: Jacobus Petrus (Koos) Bekker (职务: 腾讯公司首席战略官, 集团高级执行副总裁, 出生: 1971年10月29日 (46岁))  
法人代表: 许晨晔 (职务: 腾讯公司首席信息官兼创办人, 出生: 1973年)
- 员工人数:** 40,678人 (2017年上半年)
- 分红详情:** 公告日期: 2017-05-05, 分红方案 (每10股): 送股 (股): 0, 转增 (股): 0, 派息 (税前) (元): 1.3, 进度: 实施, 除权除息日: 2017-05-11, 股权登记日: 2017-05-10, 红股上市日: 2017-05-11
- 股票代号:** 港交所: 700

# 案例一：金融机构知识问答

基于知识图谱挖掘数据关系，基于知识图谱进行问答

通过知识图谱实现数据推理的问答

实现对比、排序、筛选、多实体等复杂查询

Question Type	Sample
Single-constraint	万科股票的价格
Constraint(implicit )	万科股票最高点绿地的股价（小米上市当天苹果的股价）
Ordinal	XX基金公司规模(收益率/净值)最大(第二大/第三大)的基金
Aggregation	XX基金公司一共有多少基金
List	今天发年报的公司有哪些
Multi-predicate	X股票今天的开盘价和收盘价（最高价和最低价）
Multi-hop	xx公司的母公司的股价
Multi-entity	x股票和y股票的股价
Multi-constraint + aggregation	过去一周（月，年）总增长超过10%的股票（基金）

## 案例二：券商多因子选股

连续1周上涨的股票  
今天涨停并且没有开板的股票  
哪些股票发生了MACD金叉

最近有哪些妖股  
近期比较热门的股票  
高成长低估值的股票

股票简称	主力资金流向(亿元)	买入日期
五粮液 000858	8.80亿	2019.08.12 买入 2019.
京东方A 000725	6.86亿	行情收
中国平安 601318	5.60亿	行情收
东方财富 300059	5.28亿	mtm金
立讯精密 002475	3.88亿	mtm金
中信证券 600030	3.46亿	mtm金
中兴通讯 000063	3.44亿	bias买入

标准属性问题理解

智能搜索

非标准属性问题理解

- ✓ 人工智能总体解决方案
- ✓ AI+NLP平台核心功能
- ✓ NLP中台介绍

自然语言理解 (Natural Language Understanding)

内容处理 (Text Analytics)

知识图谱及问答 (Knowledge Graph & QA)

**基于检索的问答 (Information Retrieval based QA)**

对话系统 (Dialog System)

智能搜索与推荐 (ReSearch)

- ✓ 应用搭建示例



# 智能问答

## 基于信息检索

### IR-QA

- 基于已预置的问答对或从文本中抽取段落
- 常见场景：FAQ、党章问答
- 主要工作：找到最相似的问句或最相关的段落
- 优势：前期投入小，上手方便快捷
- 局限性：已有问答对的数量和段落的数量

问句：如何开户？

问句：MACD金叉是什么？

## 基于知识库

### KB-QA

- 基于预先构建的知识库
- 主要工作：自然语言问句转化为结构化的数据库查询
- 优势：问句可以更为多样化
- 局限性：需要搭建结构化数据库，前期投入大

问句：方正证券的第三大股东是谁？

问句：万科的最大的股东还参股了哪些企业？

# 企业规章智能问答查询案例



# IR-QA工作台

IRQA

创建应用 我的应用 我的服务 预置应用 知识库

可定制化

一个完全可定制知识领域的Studio，为用户提供了极大的灵活性、可扩展性和稳定性。

1. 准备一个问答知识库

上传您自己的问答表，以及相关领域的关键词以创建知识库

2. 创建新的问答应用

提交您的配置以创建新的问答应用程序，训练过程大约需要5分钟

3. 测试您的问答应用

向机器人提问以测试您训练好的的问答应用

4. 调整问答知识库

调整您的问答知识库，以适应复杂的问答应用场景

5. 调整问答特征

在您的应用中引入预置问答服务，或添加预置的领域关键词

6. 将问答应用发布为云服务

将您的应用程序发布到Azure，即可获得您自己的问答服务！

- ✓ 人工智能总体解决方案
- ✓ AI+NLP平台核心功能
- ✓ **NLP中台介绍**

自然语言理解 (Natural Language Understanding)

内容处理 (Text Analytics)

知识图谱及问答 (Knowledge Graph & QA)

基于检索的问答 (Information Retrieval based QA)

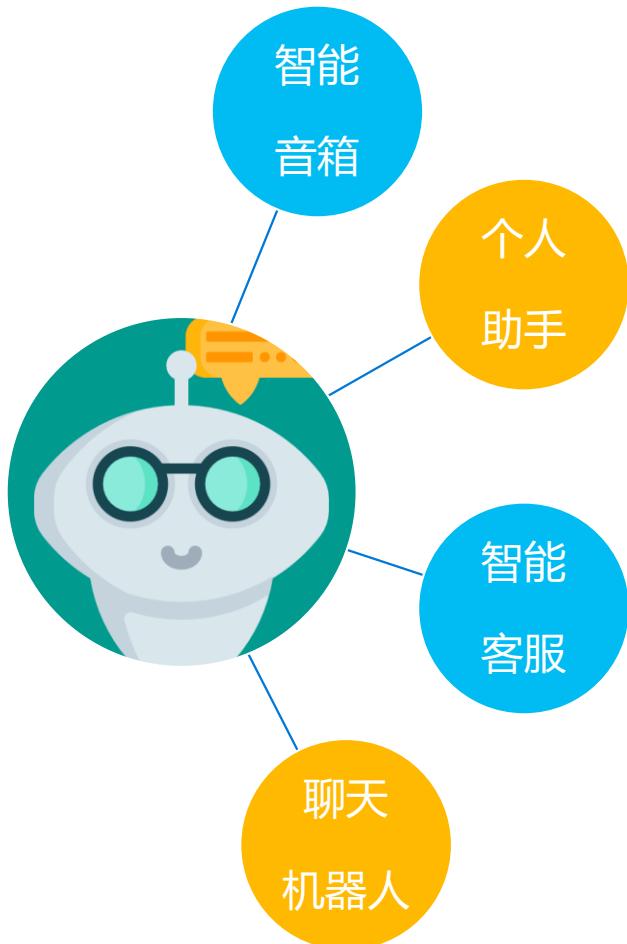
**对话系统 (Dialog System)**

智能搜索与推荐 (ReSearch)

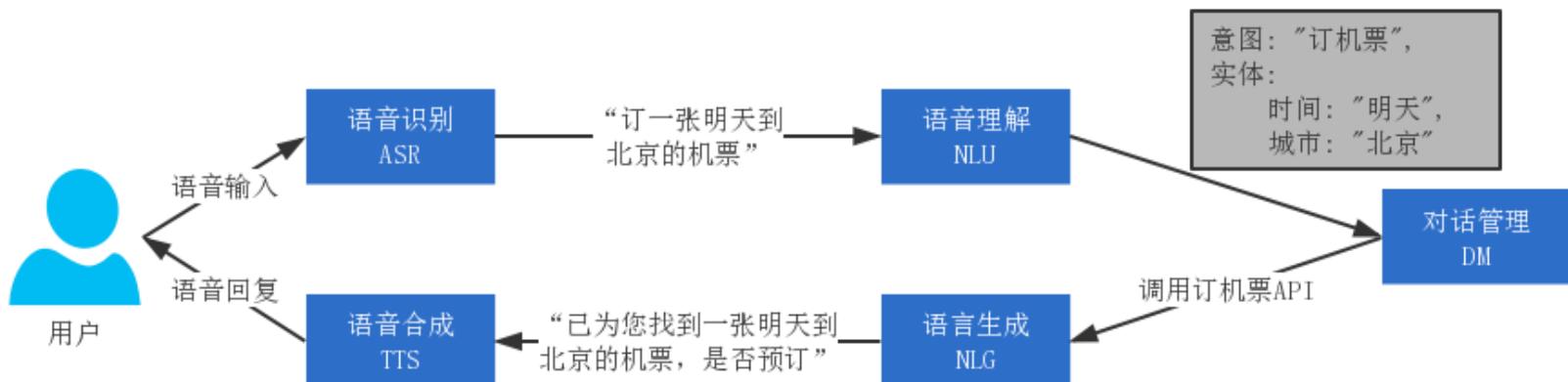
- ✓ 应用搭建示例



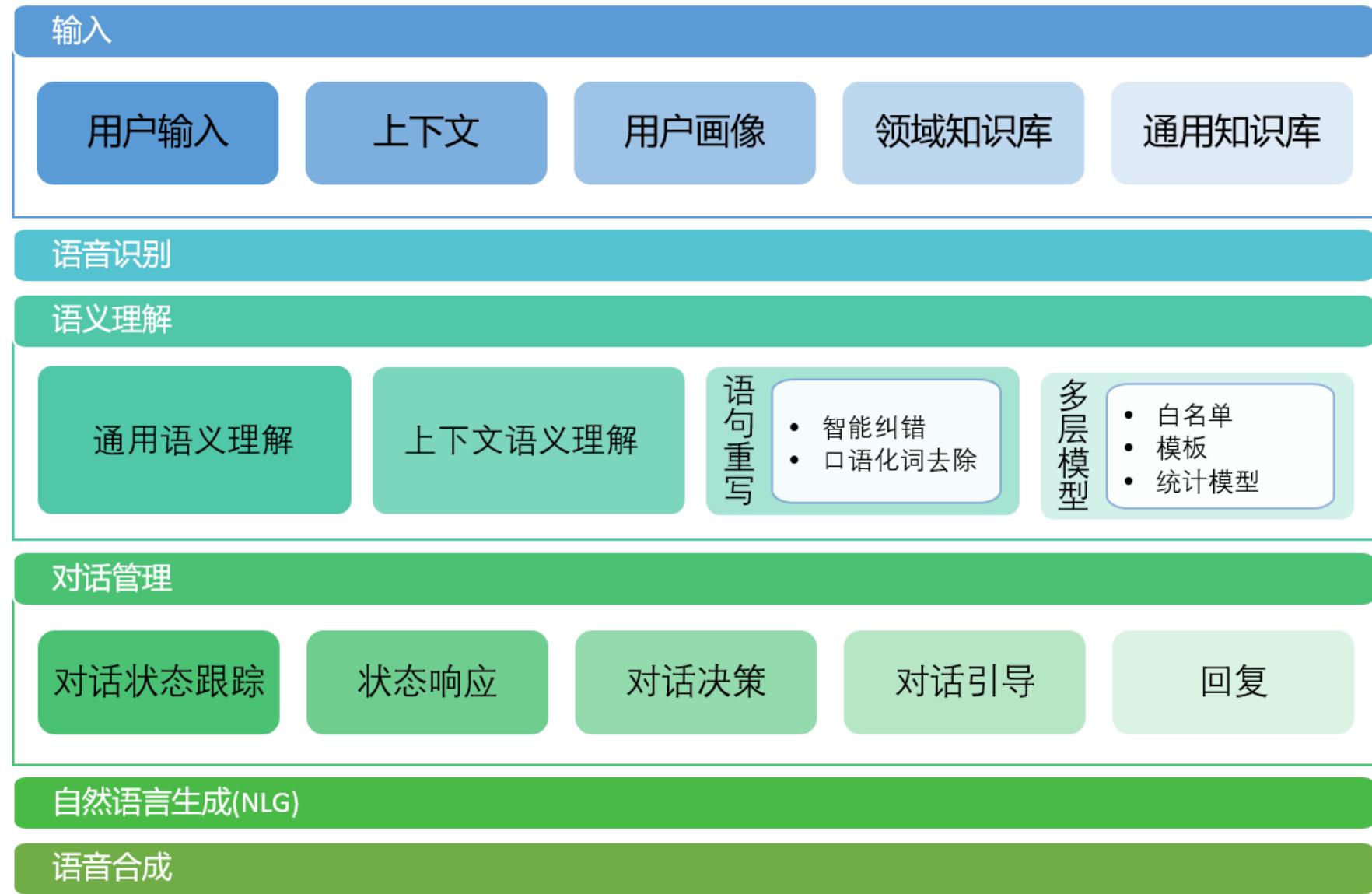
# 产品概述



- 对话系统通过构建基于语音和文本的对话系统  
(例如：聊天机器人)，为用户提供了新的与产品交互的方法。
- 提供以下信息可以帮助用户轻松的构建一个机器人：
  - 能够触发机器人的表达式语句(如“订机票”)
  - 机器人需要知道的参数(如“城市” 和 “日期” )
  - 作为对用户响应的API或回复文本(如“订机票” API)



# 我们的解决方案



# 对话管理

主动响应 →

Trip\_Bot  
今天是你的生日, Happy birthday!

我想去北京

请求澄清 →

请问您想乘火车还是乘飞机?

北京有哪几个机场

整合KBQA →

北京有首都机场和南苑机场

坐飞机有哪些注意事项

整合IRQA →

乘坐飞机需要您提前两小时值机

那我乘飞机吧

技能切换 →

请问您打算什么时候出发?

那儿的天气怎么样?

中断继续 →

北京明天预计温度30°C, 天气晴

订票还没完成, 你想继续吗?

我想明天出发

子对话 →

好的, 将要给您预定明天到达北京机票,  
是否确认

确认

机票预定成功, 具体信息将通过短信进行  
通知

对话

重置

对话状态跟踪

- 系统状态
- 用户状态

状态响应

- 主动响应
- 中断继续

对话决策

- 智能消歧
- 预执行
- 技能切换

对话引导

- 链式技能
- 技能推荐

回复

- 预置文本
- 用户接口

指令控制

- 取消
- 重复
- 转人工

# 对话系统应用案例-智慧客服系统

## 微软智慧客服系统



在线服务前

在线服务时

在线服务后

### 【数据准备】

借助智能展示管理平台进行知识与业务管理，为构建智能考培助手、智能客服机器人与智能客服助手提供数据支持。

### 【客服培训】

智能考培助手模拟用户与客服人员对话，帮助客服人员梳理与掌握知识点，并对客服人员的回复进行纠错与打分。

### 【在线服务】

智能客服机器人 $7\times24$ 小时响应客户，支持对话分拨。智能客服助手根据客服人员与客户的对话自动产生参考回复/操作，也支持知识/业务流程的智能检索与推荐。

### 【后台管理】

智能展示管理平台实时展示日志分析与系统监控数据，提供对话质检、数据挖掘等功能，掌握客户最新需求与满意度、提升服务质量。

## 以客服场景为例：

### 1. 智能考培助手

功能：对客服人员进行自动化培训与考核  
价值：减少培训成本，缩短上岗时间

### 2. 智能客服助手

功能：辅助客服人员在线回答与操作  
价值：提高工作效率，减少人力成本

### 3. 智能客服机器人

功能： $7\times24$ 小时实时响应客户需求  
价值：提高响应速度，减少人力成本

### 4. 智能分析质检

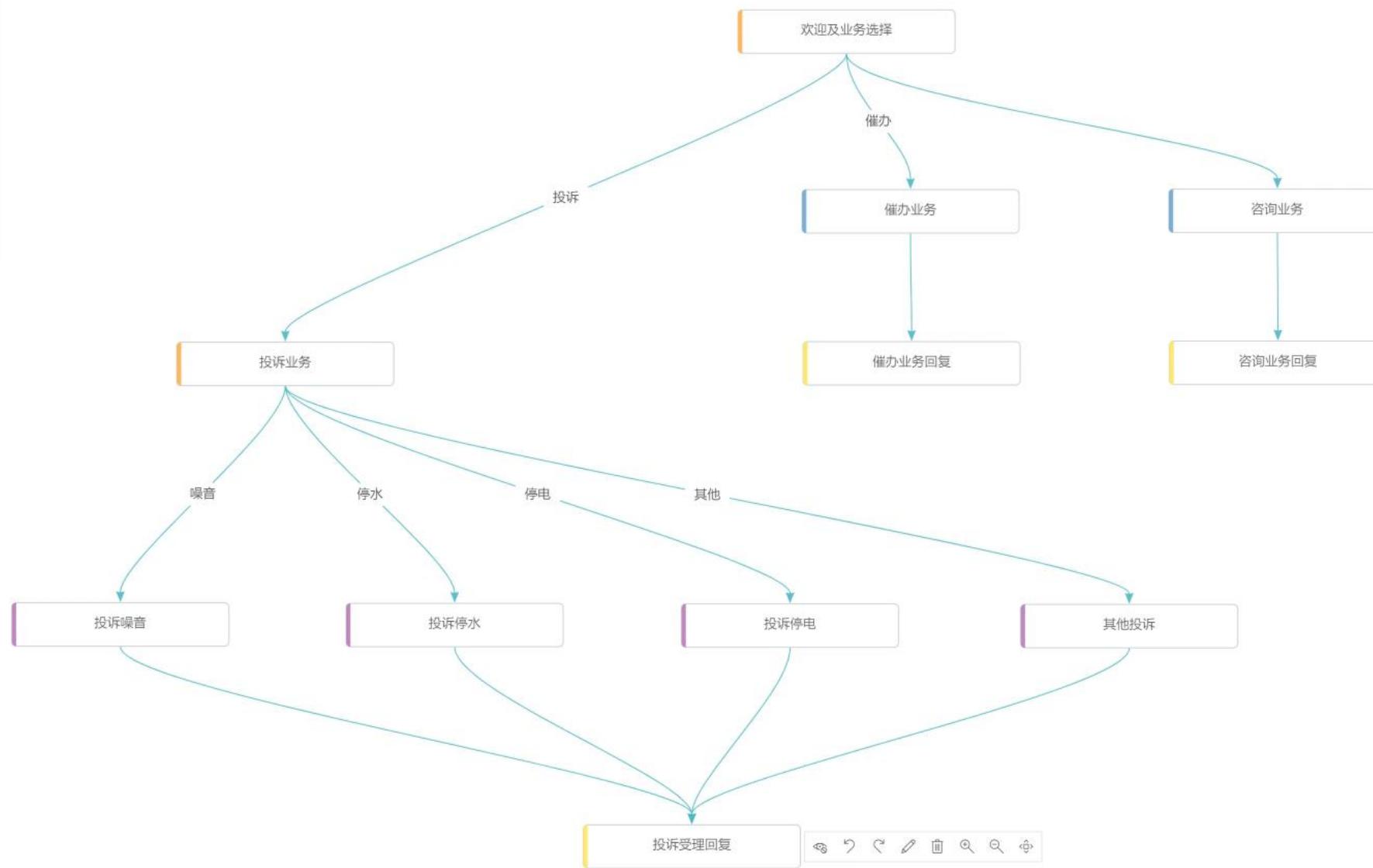
功能：实时检验对话规范、质量  
价值：减少质检成本，提高覆盖量

# 对话系统案例分享展示(2min)

## 某省会城市市长热线·智能客服（对话引擎展示）

- **任务办理 场景 - (任务驱动 | 事件驱动 | 模型驱动)**
  - 和用户交互地完成一个任务（工单）
  - Example: “我要投诉我们小区旁边的KTV，太吵了”
- **问答类 场景 - (知识图谱 | 问答库 | 模型驱动)**
  - 信息检索型
    - 从预置的问答对中匹配问题，提供答案
    - Example: “市房管局的联系方式”
  - 知识图谱型
    - 构建知识图谱，转换成图查询
    - Example: “今天可以去哪里办理个人公积金提取”
- **智能客服 上线成效**
  - **1000+** 日均接入量
  - **40%+** 成功转化率（无需转接人工）





- ✓ 人工智能总体解决方案
- ✓ AI+NLP平台核心功能
- ✓ NLP中台介绍

自然语言理解 (Natural Language Understanding)

内容处理 (Text Analytics)

知识图谱及问答 (Knowledge Graph & QA)

基于检索的问答 (Information Retrieval based QA)

对话系统 (Dialog System)

**智能搜索与推荐 (ReSearch)**

- ✓ 应用搭建示例



# 搜索系统 & 推荐系统

## · 搜索 —

- 百度/谷歌/必应
- 电商内的搜索功能
- 企业内部办公自动化文件/信息搜索
- 基于问答的信息检索

## · 推荐 —

- 电商系统的商品推荐：
  - 货品的相似推荐
  - 猜你喜欢
- 亚马逊销售额**30%**来自于推荐
- 类今日头条的信息推荐

# 好的搜索和推荐的体验涉及 ——

- 庞大的运算量：在线服务+离线服务
- 提供整套的方案 或者 部分子系统
- 自然语言理解、机器学习、图像及视频理解、深度学习、搜索和推荐优化、概率与统计、  
大数据处理...
- 集群管理、算法训练与部署；索引和排序系统、日志系统、文档处理系统、大数据处理  
系统、任务编排系统...
- 数据标注系统、指标追踪系统...

ady gaga



拼写检查

搜索 "ady gaga"

视频

Lady Gaga

<http://www.ladygaga.com>

OVERVIEW

TOP SONGS

ALBUMS



Wikipedia



Instagram



Twitter



Facebook



YouTube



IMDb



MySpace

videos of ady gaga



Lady Gaga, Bradley Cooper - Shallow (From A Star Is Born/Live From The Oscars)

178M views · 6 months ago  
[YouTube](#) · Lady Gaga



Lady Gaga, Bradley Cooper - I'll Never Love Again

177M views · 10 months ago  
[YouTube](#) · Lady Gaga



Lady Gaga - Always Remember Us This Way (From A Star Is Born)

158M views · 9 months ago  
[YouTube](#) · Lady Gaga

News about Lady Gaga



Lady Gaga is working on new music

Lady Gaga is working on new music. The 33-year-old singer has been dating the 37-year-old audio technician since

Lady Gaga's Latest Cryptic Tweet Has Her Fans Freaking The 'F' Out

Lady Gaga shared just one single letter in a new tweet: the letter 'F'. Fans responded with a flurry of excitement, hoping it means songs from her new album are ...

[Women's Health](#) · 8h



See all images

Lady Gaga

Musician

Stefani Joanne Angelina Germanotta, known professionally as Lady Gaga, is an American singer, songwriter and actress. She is known for her unconventional, provocative work and visual experimentation. Gaga began performing as a teenager, singing at open mic nights and acting in school plays. She studied at Collaborative Arts Project 21, through Ne...

Born: Mar 28, 1986 (age 33) · New York, NY

Height: 5 feet 1 inches

Net worth: \$320 million USD (2019)

Member of: SGBand

Awards: Grammy Award for Best Pop Duo/Group Performance (2019) · Grammy Award for Best Pop Solo Performance (2019) ·

Nominations: Academy Award for Best Original Song (2016) · Grammy Award for Best Pop Vocal Album (2018, 2012) ·

Interesting stories



Faces of Lady Gaga's beauty campaign on importance of brand's inclusive message

[YAHOO! News](#) · 2 weeks ago



Lady Gaga's New York

[beyondmidtown.com](#) · 2 months ago



Lady Gaga's Top 10 Rules For Success (@ladygaga)

Evan Carmichael · 2 years ago

Listen or buy



Apple Music  
Buy



YouTube  
Play

实体

新闻

Microsoft

- ✓ 人工智能总体解决方案
- ✓ AI+NLP平台核心功能
- ✓ NLP中台介绍
- ✓ 应用搭建示例



# 平台应用示例（知识图谱的构建）

平台



1 原始数据



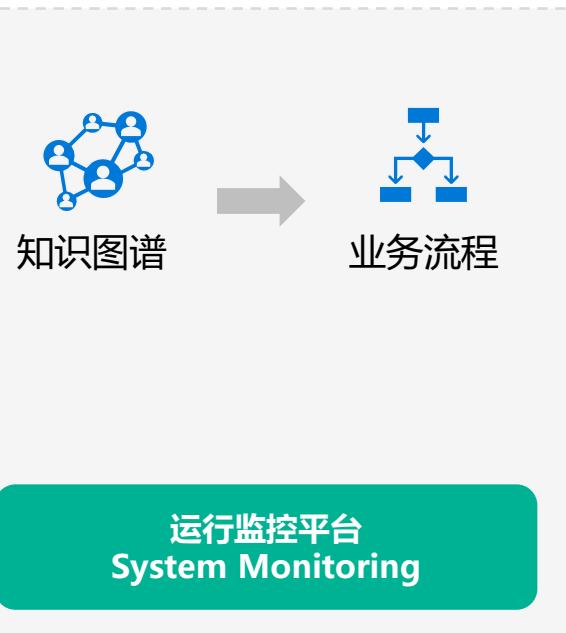
2 获取数据



3 数据处理



4 发布



5

不同数据源

不同获取方式和更新频率

不同数据处理要求

不同运行状态

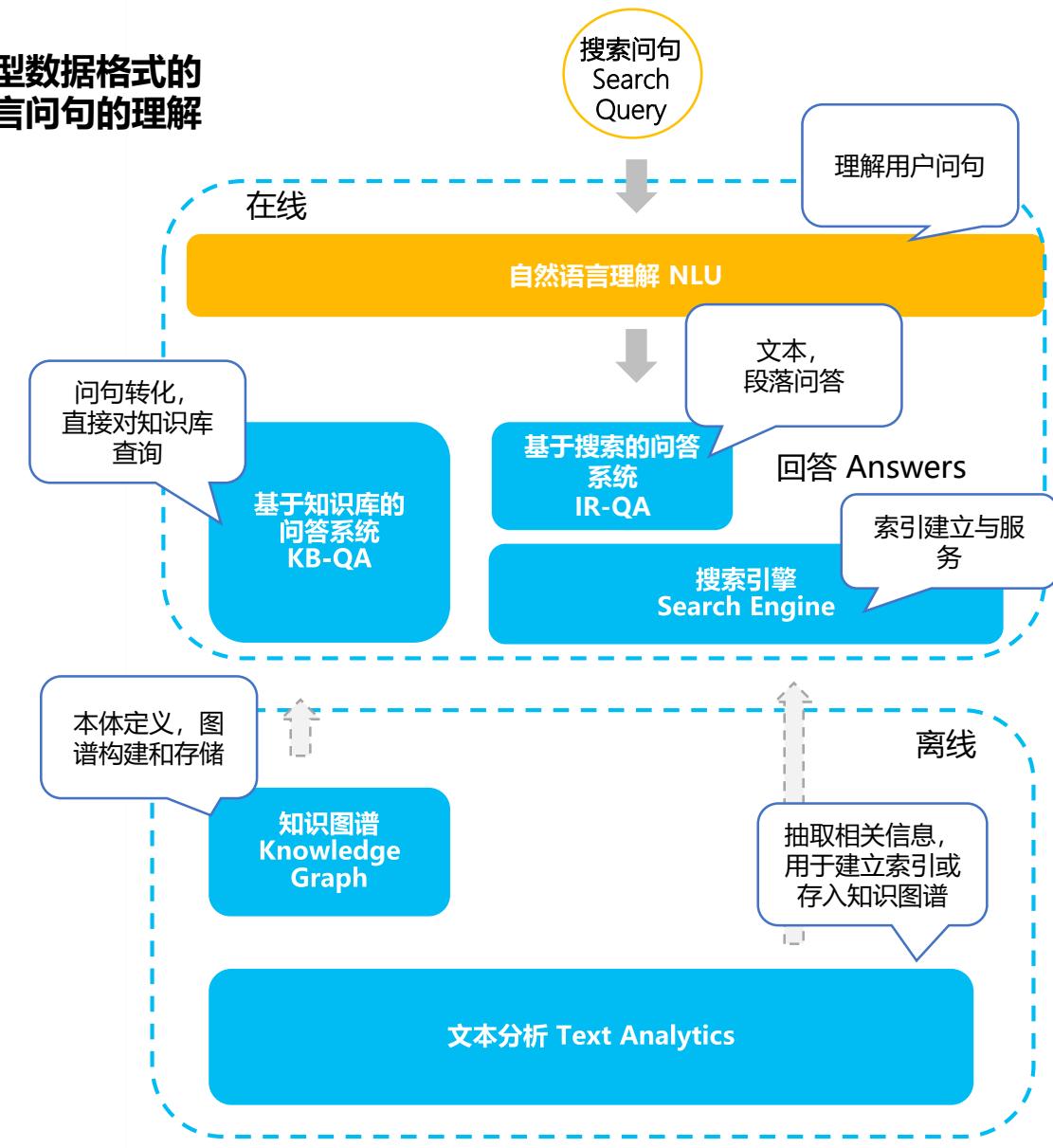
任务编排平台  
Job Orchestration

# 智能搜索

- **常见需求：**
- 办公搜索（人，文件，网站，流程等）
- 智能投研（金融领域的资讯，文件搜索）

The screenshot shows the Office 365 SharePoint interface. At the top, there's a navigation bar with 'Office 365' and 'SharePoint'. Below it is a search bar with 'Marketing analysis'. Underneath are tabs for 'All', 'Sites', 'People', 'Files', and 'News'.  
**Instant answers:** This section displays a 'Title and Content Layout' chart with four categories (Category 1, Category 2, Category 3, Category 4) each containing three series (Series 1, Series 2, Series 3).  
**People search:** This section shows profiles for Arjuna Mishra, Christos Eliopoulos, Hedi Peetre, Joao Mello, and Mia Wilkins, categorized by skill or department.  
**Whole page ranked multiple entity types from across provenances:** This section shows a list of items including 'What markets to target the next year?', 'Business-to-Business Marketing: A Strategic Approach', 'Marketing Analysis tools', and 'Marketing department operations', each with a brief description and stats like views and last update date.

- **功能描述：**
- 在传统搜索之上增加了不同类型数据格式的搜索体验，提高对用户自然语言问句的理解以及回答的多样性

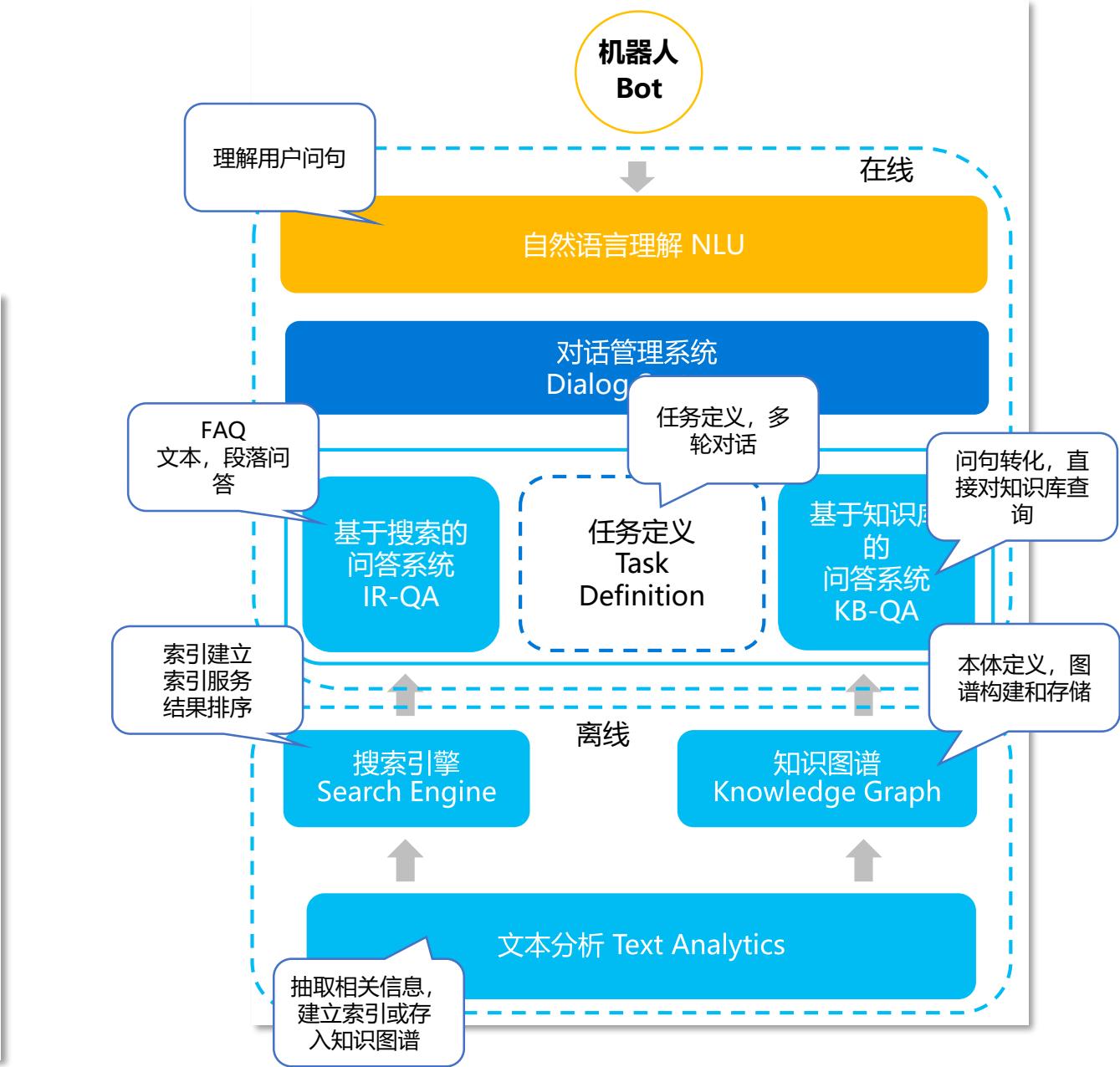
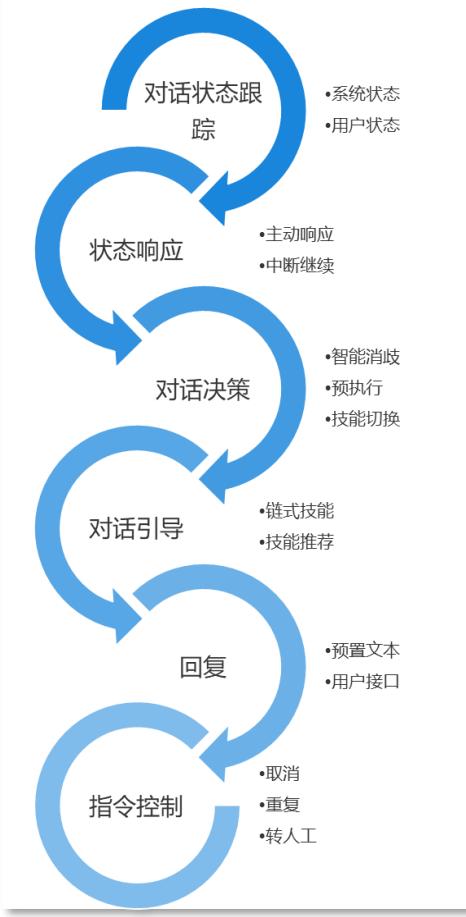


# 对话机器人

- **典型案例:**
  - 市长热线
  - 智慧客服

- **功能描述:**

有任务完成，问答等形式的对话机器人的需求

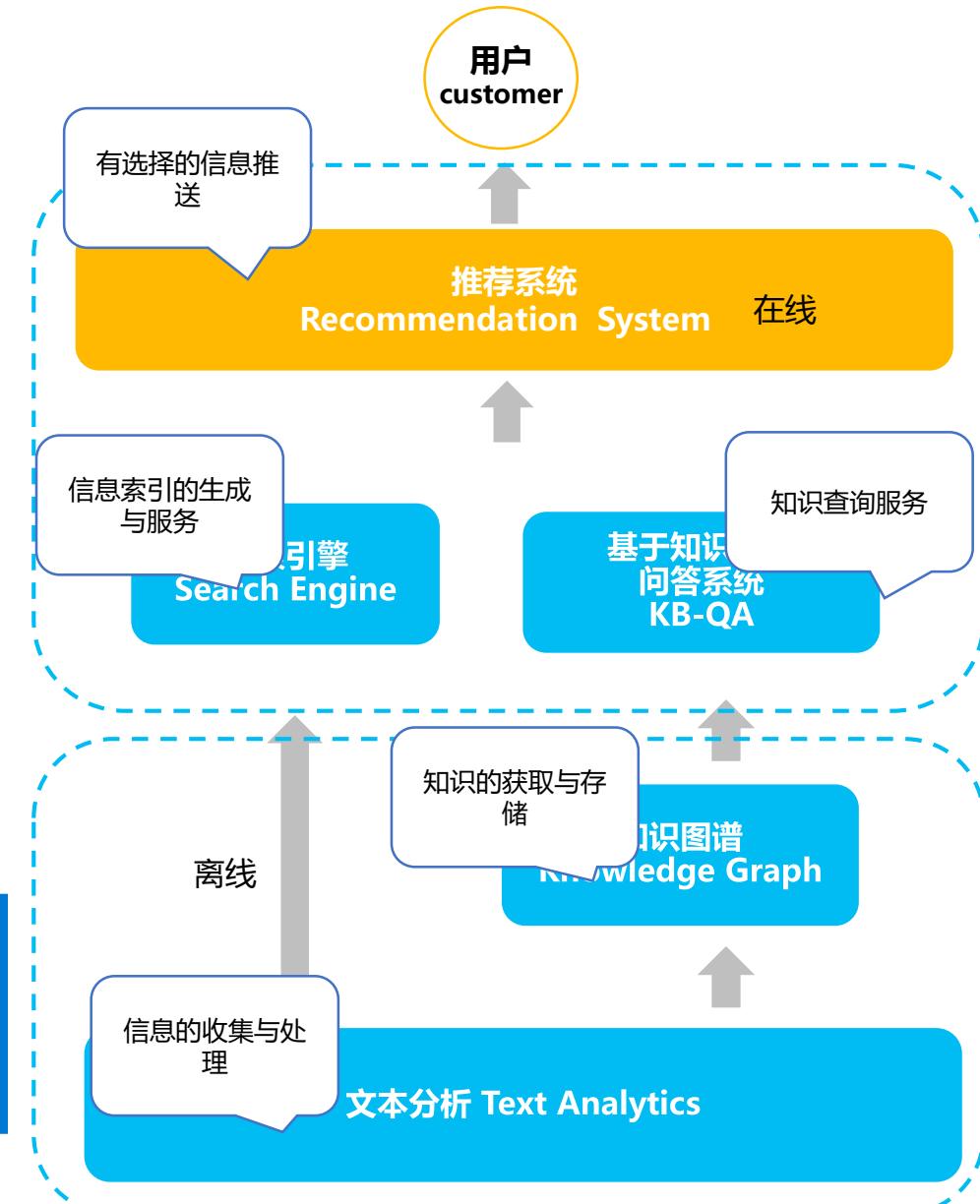


# 舆情监控

- **典型案例:**
  - VoC 客户声音(一汽)
  - 新闻资讯

## • 功能描述:

为有信息收集, 分析, 推送, 提醒 (报警) 需求的客户提供基础能力



# 为什么选择微软？



## 已在微软内部达到大规模应用

上线8年

每周有~**1800**用户访问

提交超过**300,000**个模块

每天提交超过**20,000**个实验

至今共提交了超过**1500万**个实验

## 已被验证的机器学习能力

### 超级云计算平台

01

- 微软Azure支持机器学习和深度学习在大数据上的应用。技术在必应搜索引擎，小娜及小冰等亿级用户应用上得以验证和发展

### 硬核机器学习算法

02

- 深度基础研究以及所拥有的世界最多的AI相关技术专利成就了微软的在语音，图像、自然语言处理等AI基础领域拥有尖端算法。

### 丰富的场景和海量数据

03

- 微软丰富的产品线（Office365, Dynamics, Bing, Gaming等）全面覆盖各种类型的机器学习模型，并积累了来自Web以及企业端的海量数据。



## 联系我们

✉ 邮箱: [stcaai@Microsoft.com](mailto:stcaai@Microsoft.com)

“我们的使命是予力地球上每一人每一组织成就不凡。”

纳德拉 萨提亚，微软CEO