大数据智能之道法术

Big Data Intelligence: Tao, Principle and Tactics

张华平 博士

大数据搜索与挖掘实验室

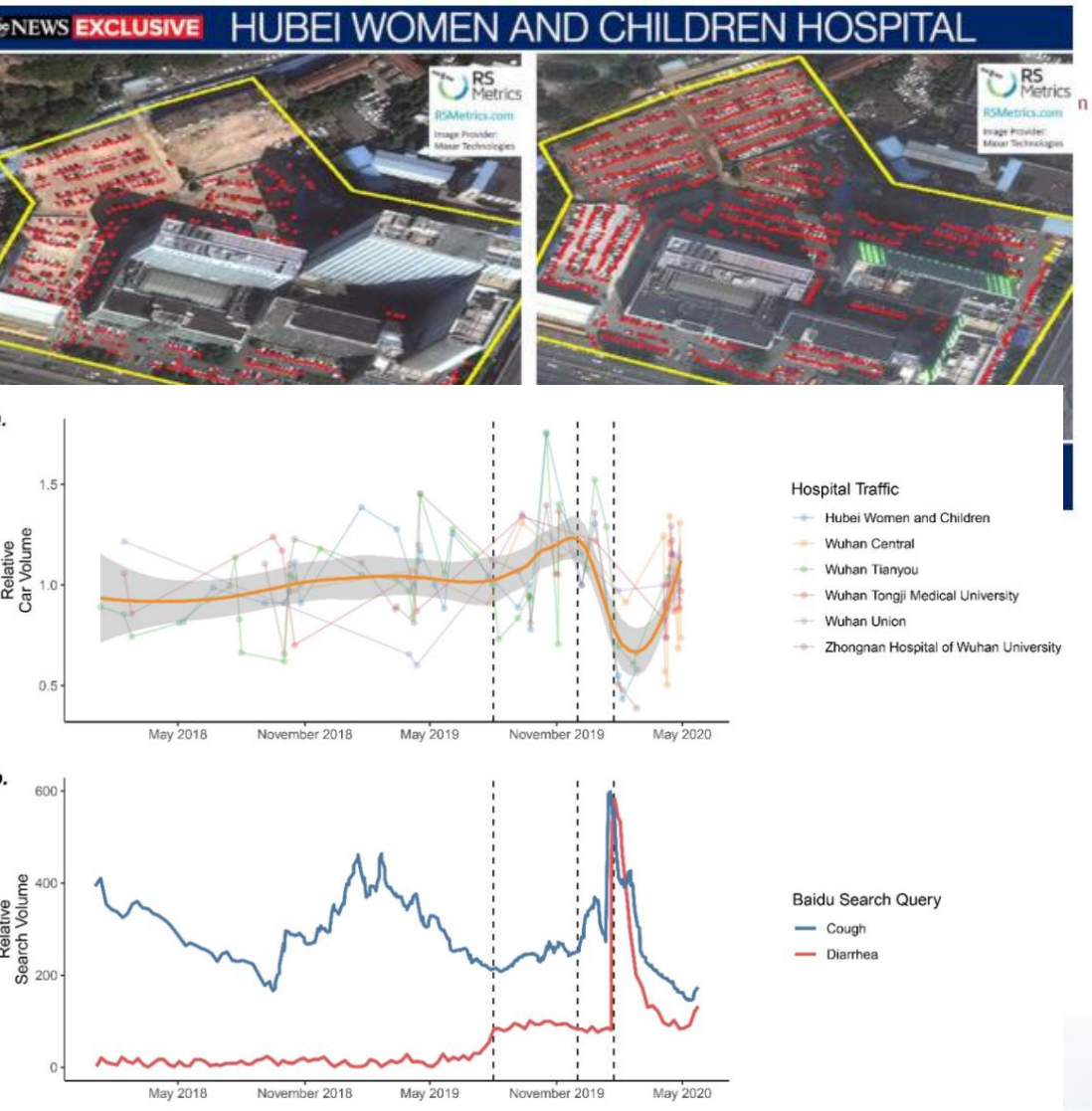
kevinzhang@bit.edu.cn

www.nlpir.org

2021.9

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

大数据智能：新冠预测与五角大楼停车场指数



2012年10月13日这一天，西南侧车辆突然增多约100辆左右

10月13日美海军"蒙彼利埃"号潜艇与提康德罗加级巡洋舰"圣哈辛托"号在东部海域相撞

从棱镜手机监控看大数据...

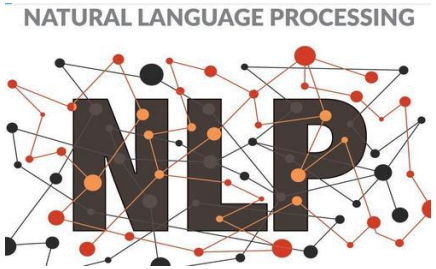
社交媒体大数据背景下的信息爆炸与社群操控

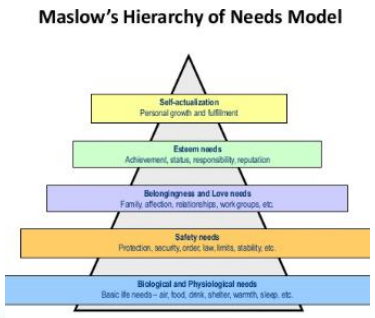
facebook 日活跃19.1亿

日活跃10.90亿

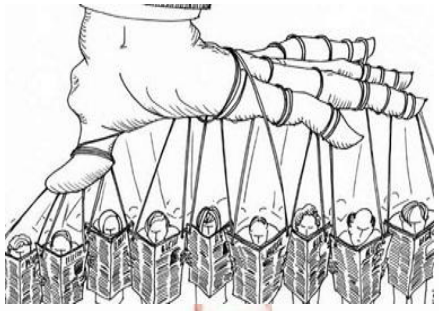
日活跃3亿

日活跃1.45亿





Cambridge Analytica



北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

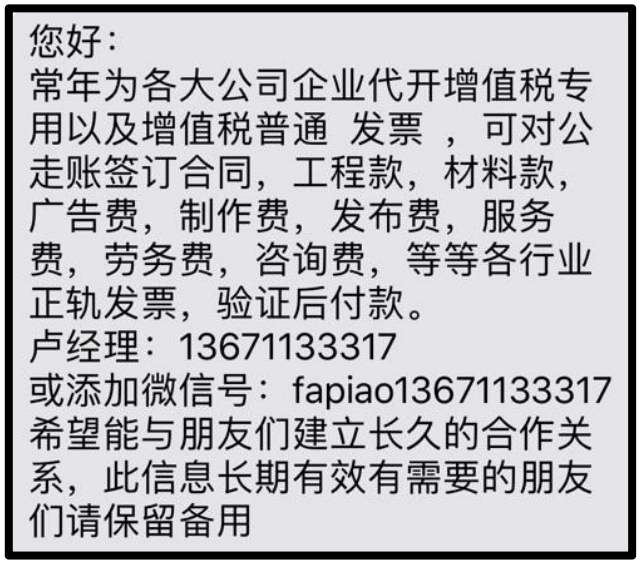
国家安全问题:分裂暴恐势力藏身社交媒体大数据

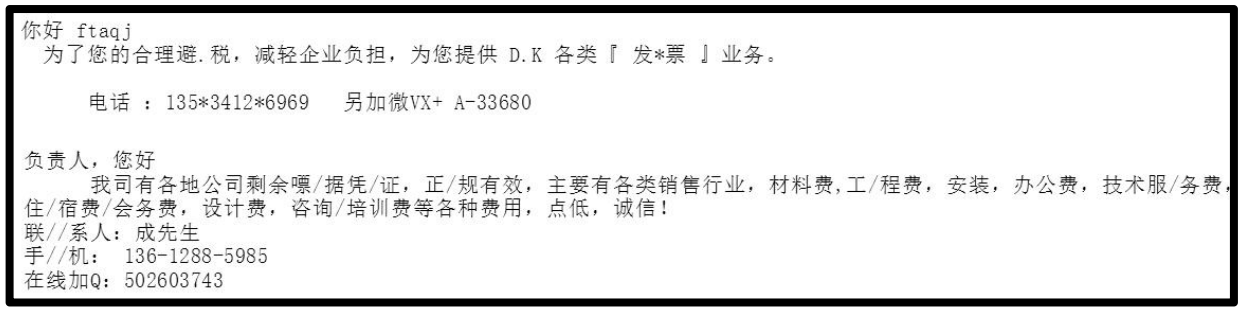
港独台独勾结，危害国家稳定。境外势力插手，妄图独立台湾，

大数据分析与应用案例

工大學
F TECHNOLOGY

社会治理问题：违法犯罪信息变身逃避常规扫描





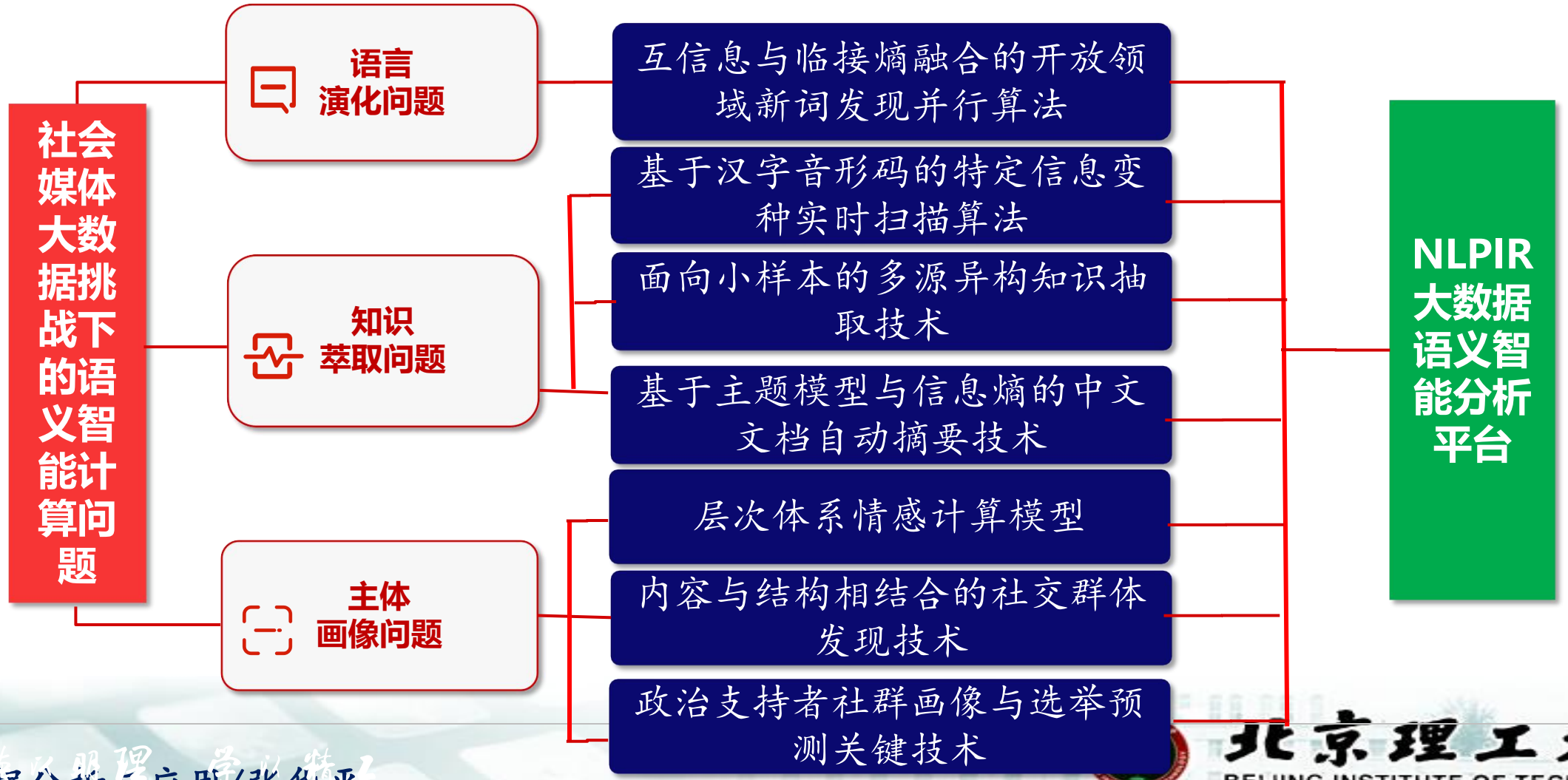
01 **假发票短信泛滥**

02 **涉赌短信屡禁不绝**

03 **假发票邮件不断“变种”**

大数据智能分析架构

一个科学问题　三大技术挑战　七项创新技术　一个集成平台



大数据 人工智能 自然语言处理

推动互联网、**大数据、人工智能**和实体经济深度融合

-十九大报告

人工智能，现代科学皇冠上的明珠；**自然语言处理**则是人工智能皇冠上的明珠。

-周明，ACL主席

人工智能的突破在**自然语言理解**，懂语言者得天下

-沈向洋，原微软全球执行副总裁

认知智能	• 认知能力：与人的交流、交互与交融 • 自然语言处理、知识推理
感知智能	• 感知能力：受限的环境 • 自动驾驶、人脸识别、传感器等
计算智能	• 计算能力：人工定义的严格规则 • AI剪枝优化决策，大数据存储与计算

什么是大数据智能

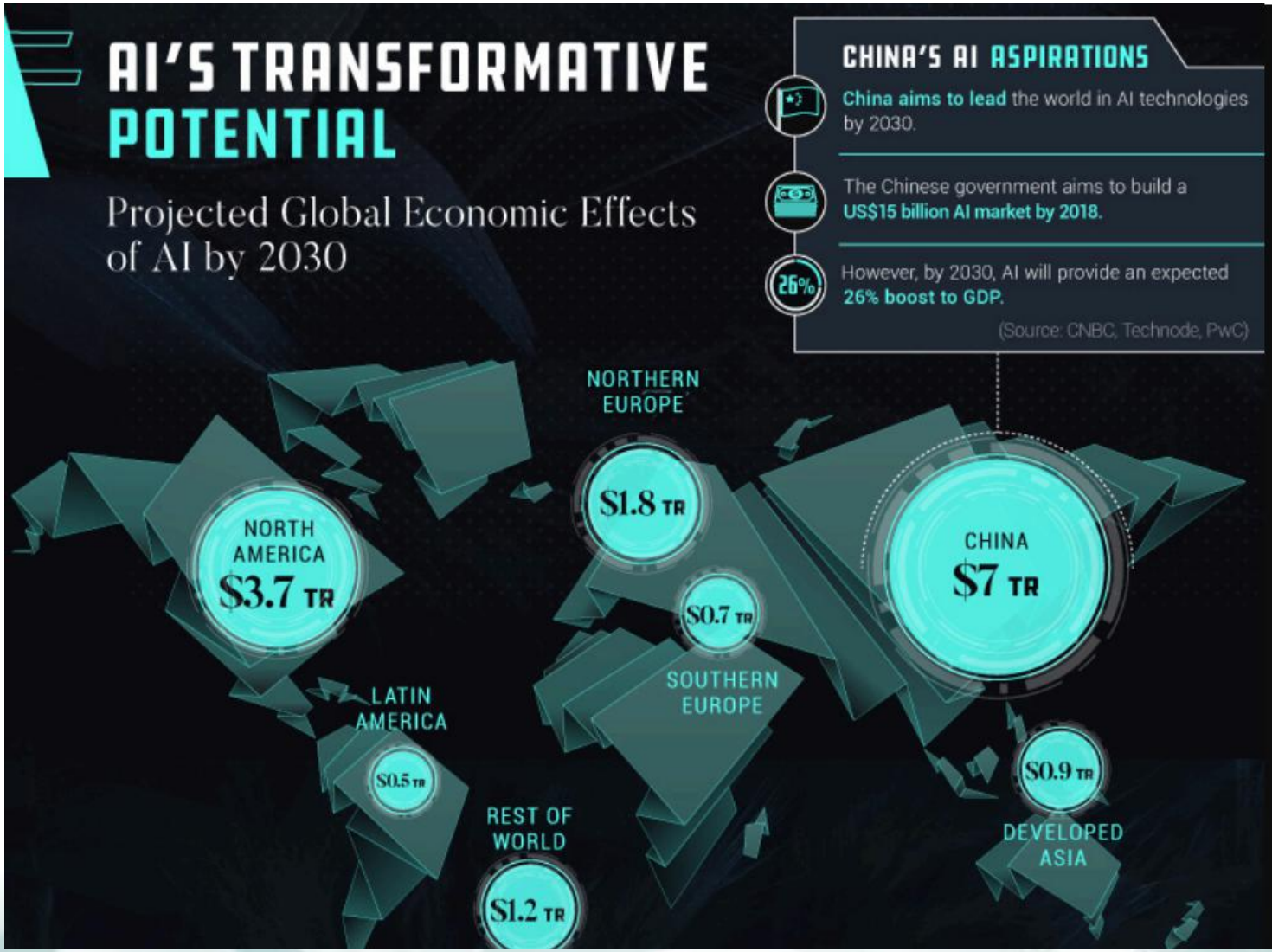
↗我们的见解：

- 大数据智能是指从客观存在的全量超大规模、多源异构、实时变化的微观数据中，利用自然语言处理、信息检索、机器学习等技术抽取知识，转化为智慧的方法学。

- 神即道，道即法，道法自然，如来

- 智能为道、数据为法、语义为术

大数据智能将带来巨大的变革与机会



15.7 万亿美元

2030年

PWC预测的世界AI市场规模

(2017年GDP 中国+印度=15.3万亿美元)


BIG
DATA
SOFTWARE
ANALYTICS
TECHNOLOGIES
INTERNET
PETABYTES
SIZE
NEEDED
PARALLEL
SAN
HUNDREDS
DISK
SHARED
TIME
INCLUDE
SYSTEMS
NETWORKS
INFORMATION
SEARCH
RECORDS
MANAGEMENT
SOCIAL
LARGE
EVERY
STORAGE
RESEARCH
DATABASES
COMPLEX
ABILITY
SETS
EXAMPLES
MASSIVELY
GARTNER
BIOLOGICAL
TOLERABLE
DISTRIBUTED
CAPTURE
MANAGE
GROW
TERABYTES
CONTINUES
CURRENT
APPLIED
ARCHIVES
SENSOR
TARGET
DIFFICULTY
CITATION
COST
WITHIN
PROCESS
DEFINITION
TOOLS
SET
GENOMICS
PERFORMANCE
RELATIONAL
CAPACITY
PRACTITIONERS
PROCESSING
LOGS
STORE
CASE
LARGER
USE
USING
WORKING
BUSINESS
MOVING
MAY

认识大数据

什么是大数据

- **Wiki: Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- 维克托《大数据时代》：大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法。

什么是大数据智能

我们的见解：

- 大数据智能是指从客观存在的全量超大规模、多源异构、实时变化的微观数据中，利用自然语言处理、信息检索、机器学习等技术抽取知识，转化为智慧的方法学。
- 是一场新的科技革命，也是思想方法的革命。（全量分析，让数据说话；承认并客观地认识世界的混杂性； 相关性挖掘替代因果推断）


这是显微镜下的世界


这是望远镜中的宇宙
© AFP/Getty Images

杨达才启示：1+1>>2才是大数据

近半世纪来的三次革命

计算机时代

20世纪
70年代

计算方式的革命

互联网时代

20世纪
90年代

信息传播方式的革命

大数据时代

21世纪10
年代

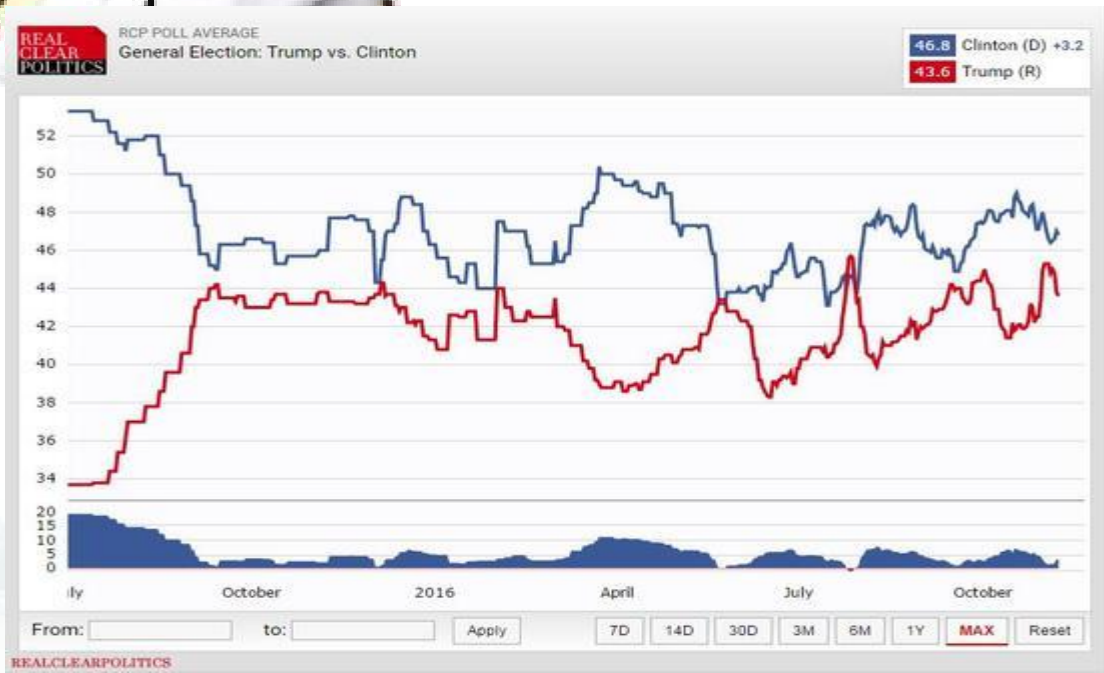
决策方式的革命

大数据颠覆决策模式


富春

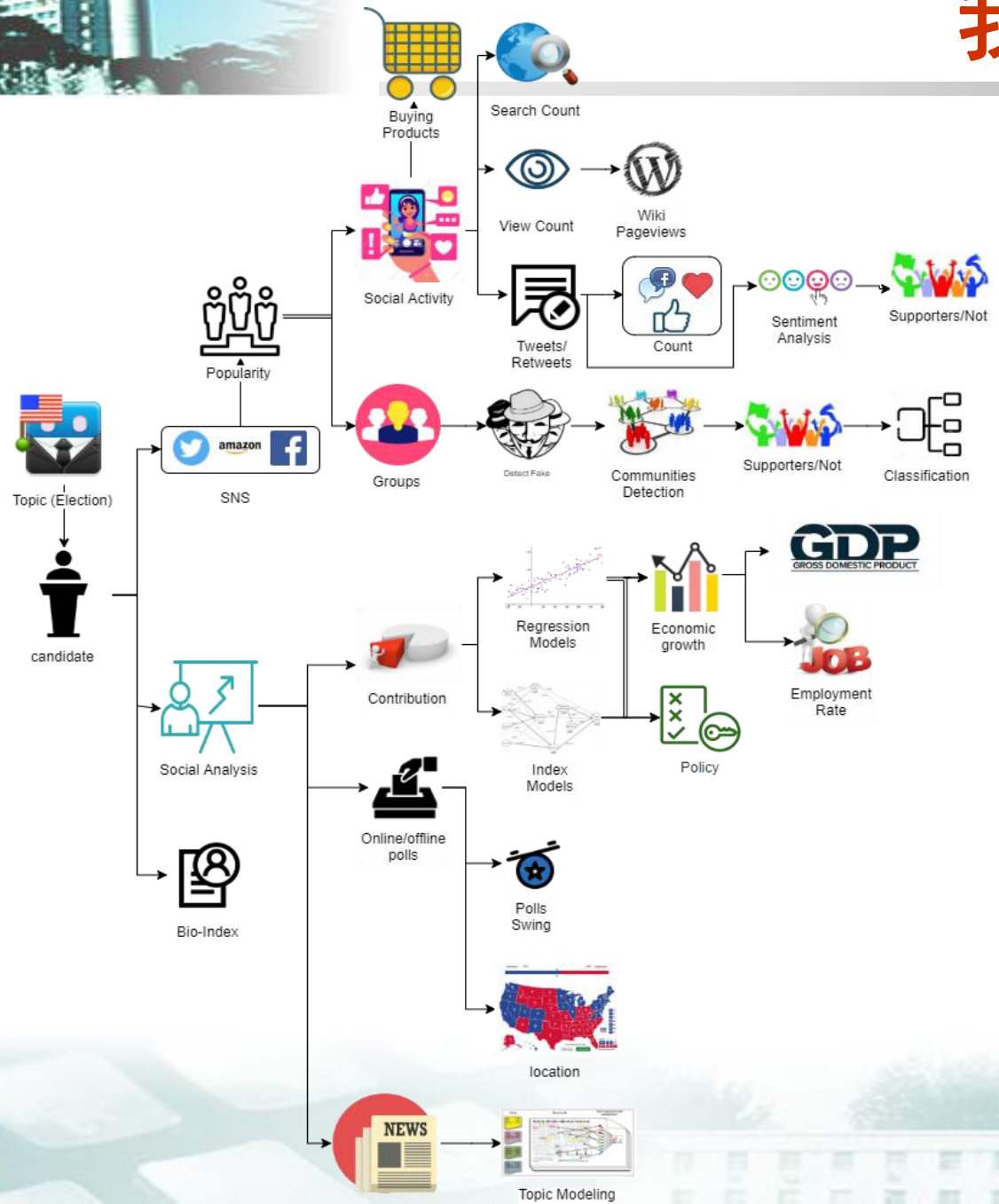

老男孩
肖央
导演作品
7月10日全国上映
人生很搞 梦想不老
PLAZA
USA
领衔主演：肖央 王太利（筷子兄弟） 监制：李仁港 制片人：柯利明
@优酷

小数据精英 VS 大数据庶民



小数据精英 VS 大数据庶民

我们的大数据预测模型



- 2018年提前半年成功预测巴基斯坦大选；
- 2020年初成功预测蔡英文连任，预测的得票率误差在5%；
- 2020年预测美国大选特朗普胜算大于拜登10%，目前失败但仍然颠覆了民调结果。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

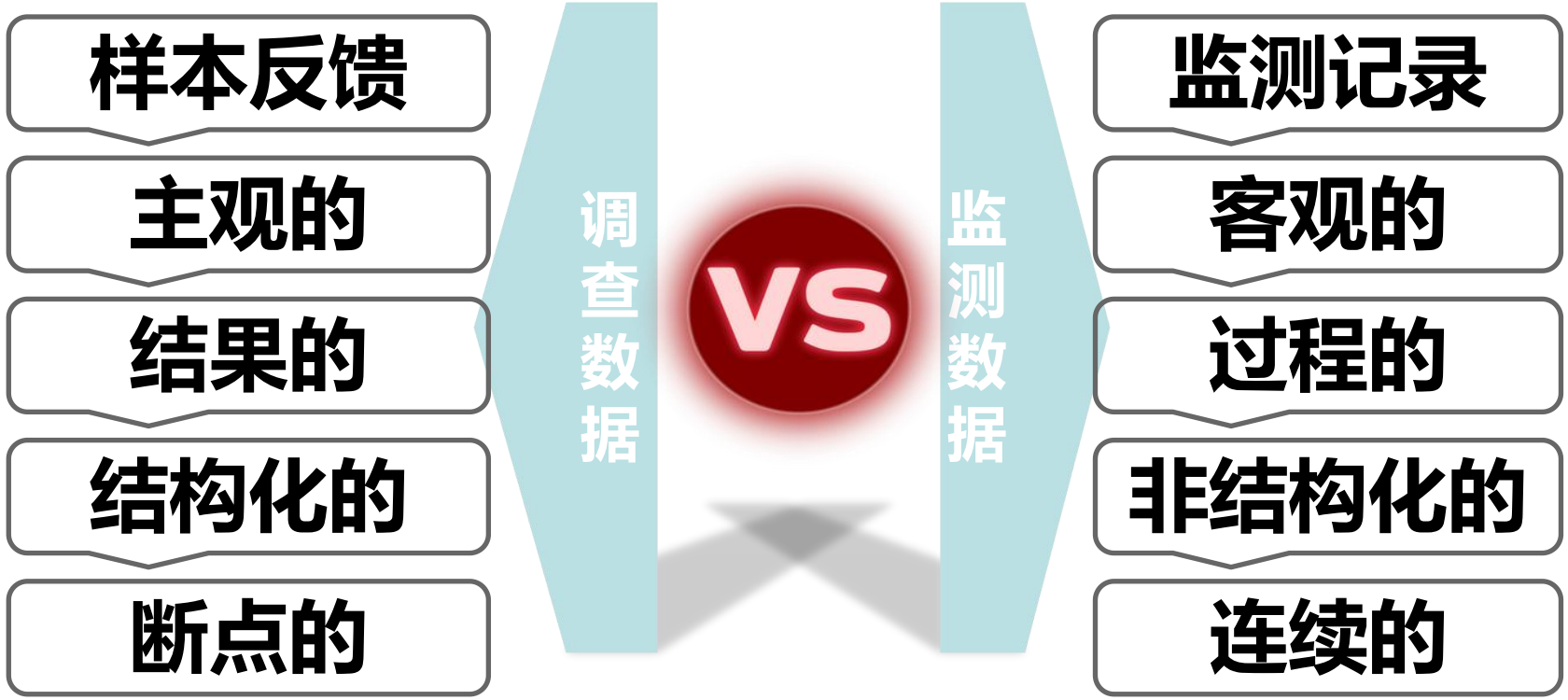
幸存者偏差：你不了解的中国

- 10亿人没有坐过飞机；4亿人没有使用过抽水马桶
- 6亿人次出国，出国的人数只占5%
- 2019年6.18，消费升级，市场下沉，四五线贡献率超过7成
- 618奇瑞小蚂蚁400 1分钟销量387台，比亚迪7805台
- 7.14高炮背后的庞大群体

大数据时代的特征


Variety
Velocity
Volume
Structured & Unstructured
Batch
Structured
Streaming Data
Zettabytes
Terabytes
省
好
快
多

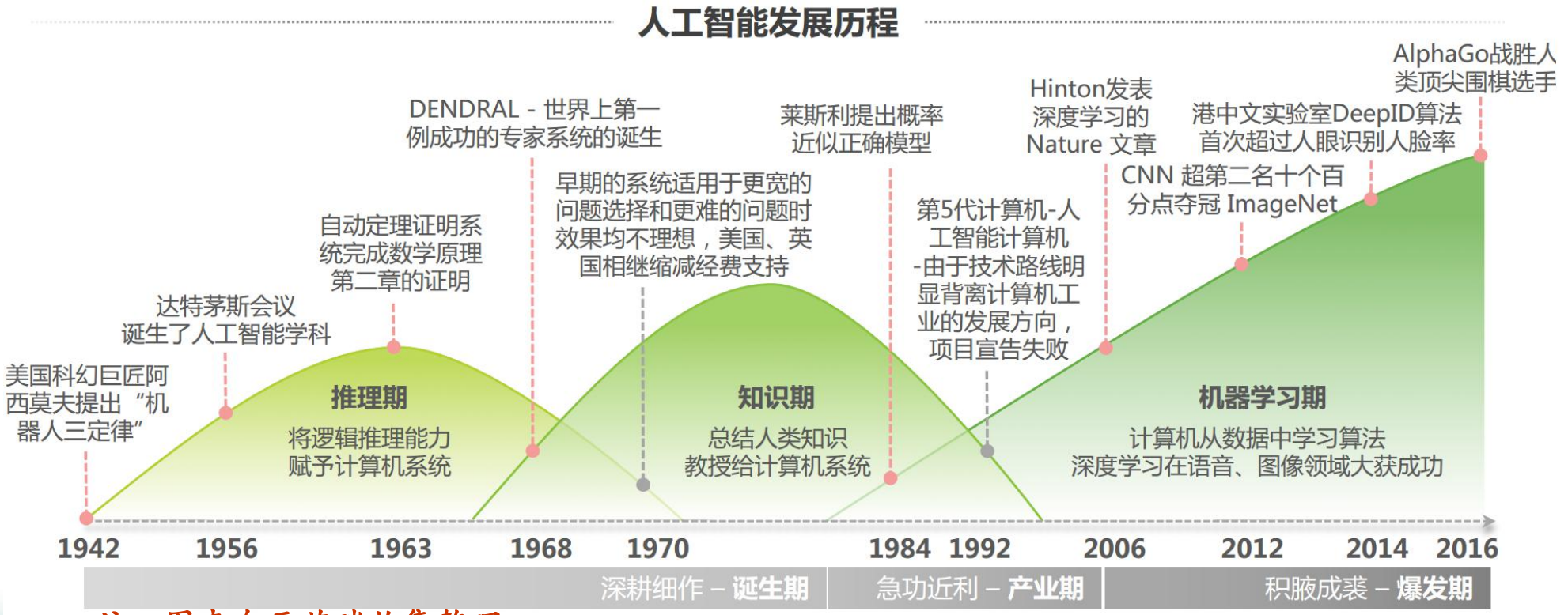
小数据和大数据的区别



认识人工智能

人工智能定义与发展历程

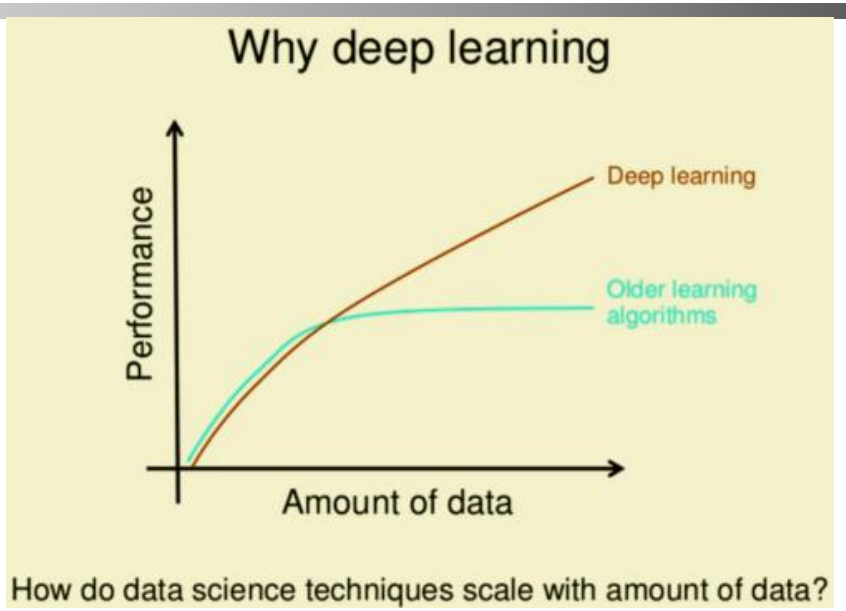
- 1956年，约翰麦卡锡(John McCathy)在达特茅斯会议上首次提出人工智能（Artificial Intelligence: AI）的定义：使一部机器的反应方式像一个人在行动时所依据的智能。

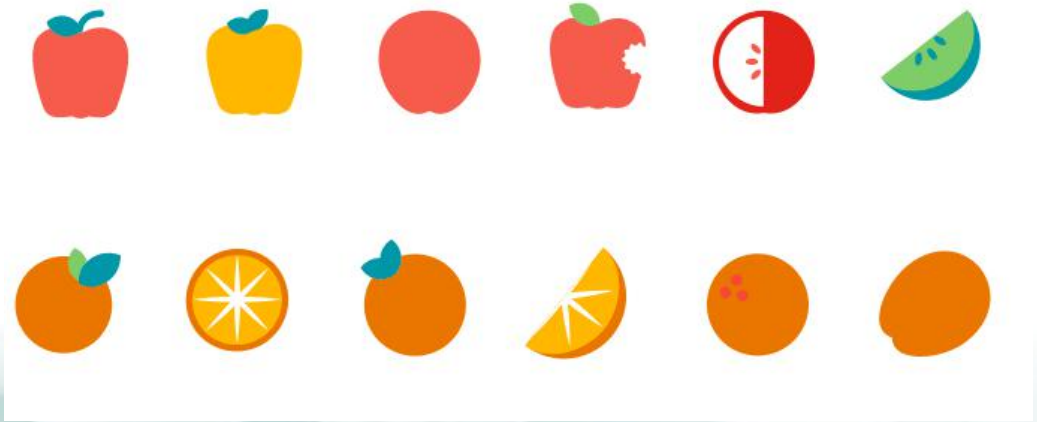


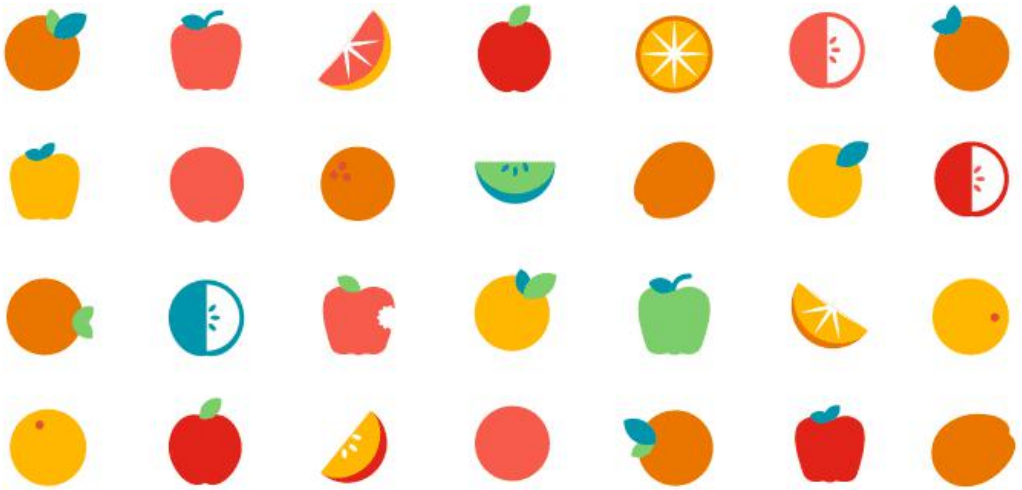
注：图来自于艾瑞收集整理

机器学习与深度学习


人工智能
早期的人工智能令人兴奋不已
机器学习
机器学习开始兴起
深度学习
深度学习取得突破
驱动人工智能蓬勃发展
1950's
1960's
1970's
1980's
1990's
2000's
2010's


Why deep learning
Performance
Amount of data
Deep learning
Older learning algorithms
How do data science techniques scale with amount of data?





人工智能的三个境界

超人工智能
- 认知智能：与人的交流、交互与交融
- 自然语言理解、知识推理

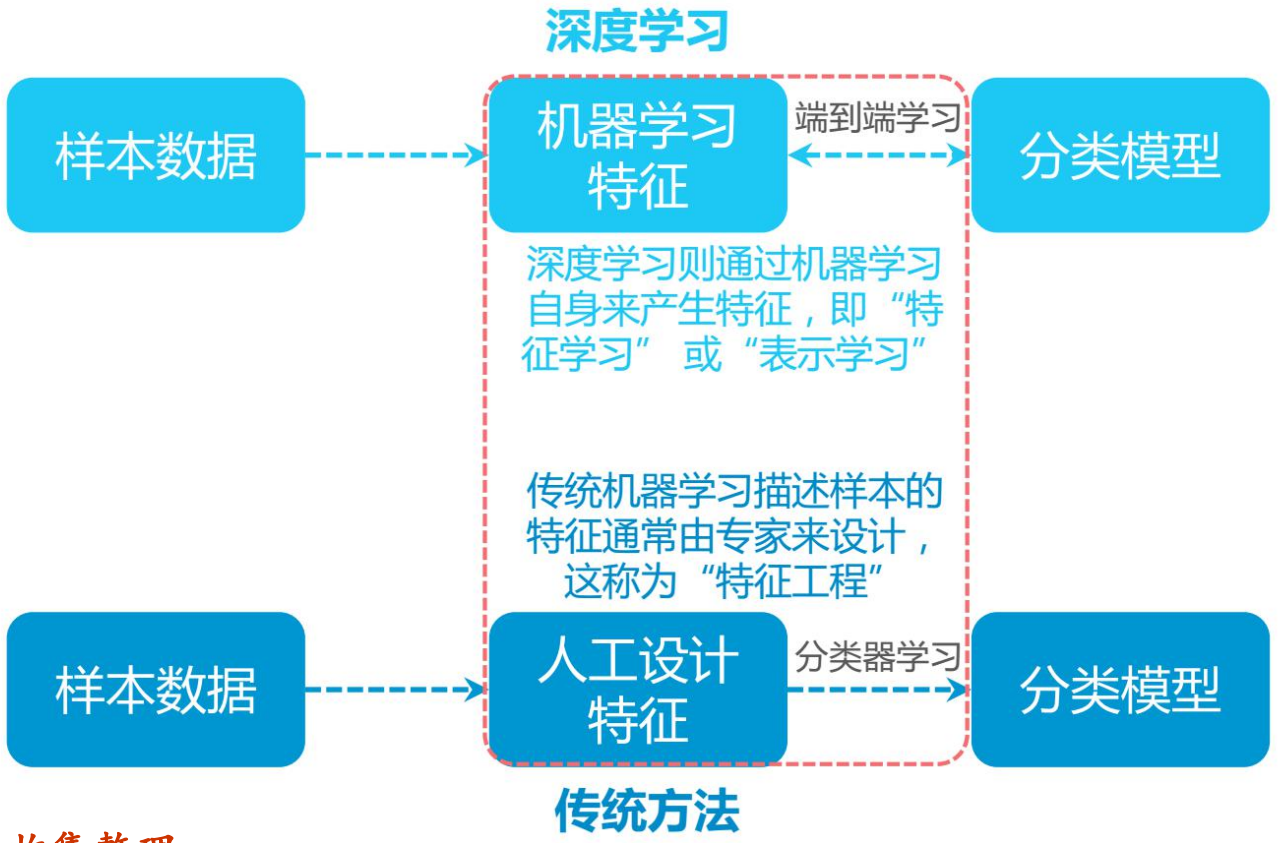
强人工智能
- 感知智能：受限的环境
- 自动驾驶、人脸识别、传感器等

弱人工智能
- 计算智能：人工定义的严格规则
- AI剪枝优化决策，大数据存储与计算

深度学习vs机器学习vs人工智能

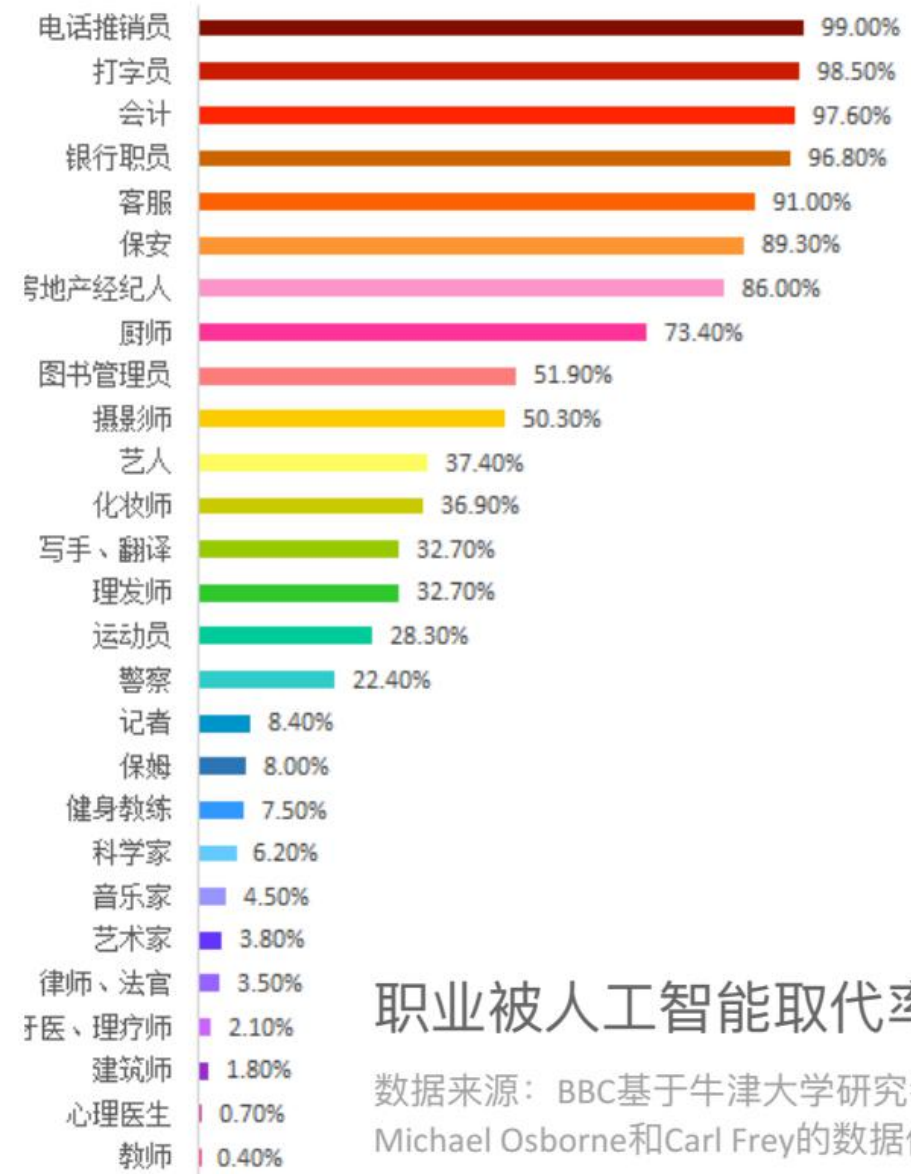
深度学习 < 机器学习 < 人工智能

深度学习与传统方法的区别



注：图来自于艾瑞收集整理

AI将取代大量人力工作



职业被人工智能取代率

数据来源：BBC基于牛津大学研究者Michael Osborne和Carl Frey的数据体系分析


翻译


记者


司机


客服


保姆


助理


保安

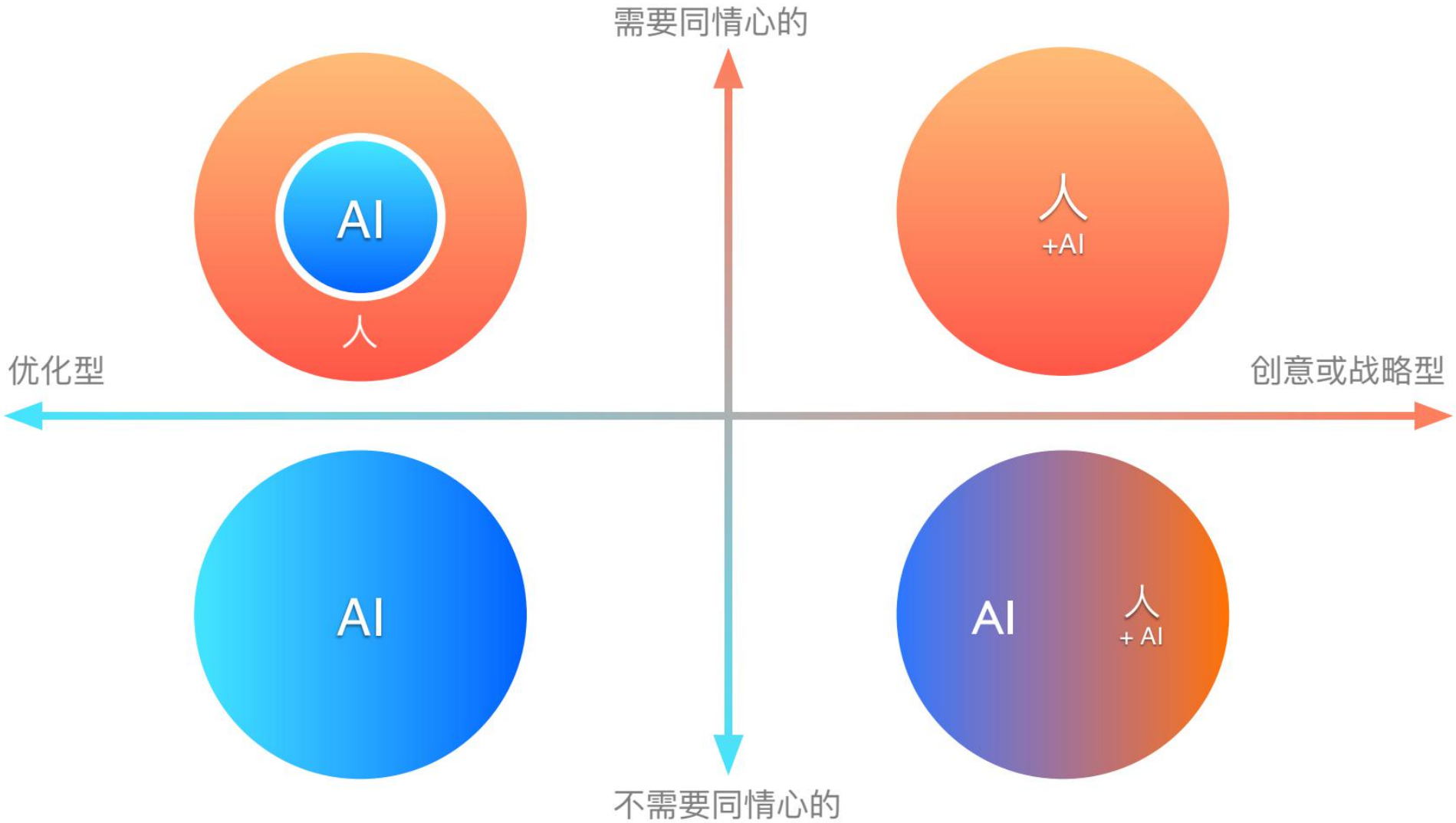

交易员

甚至还有…


法律咨询师


放射科医生

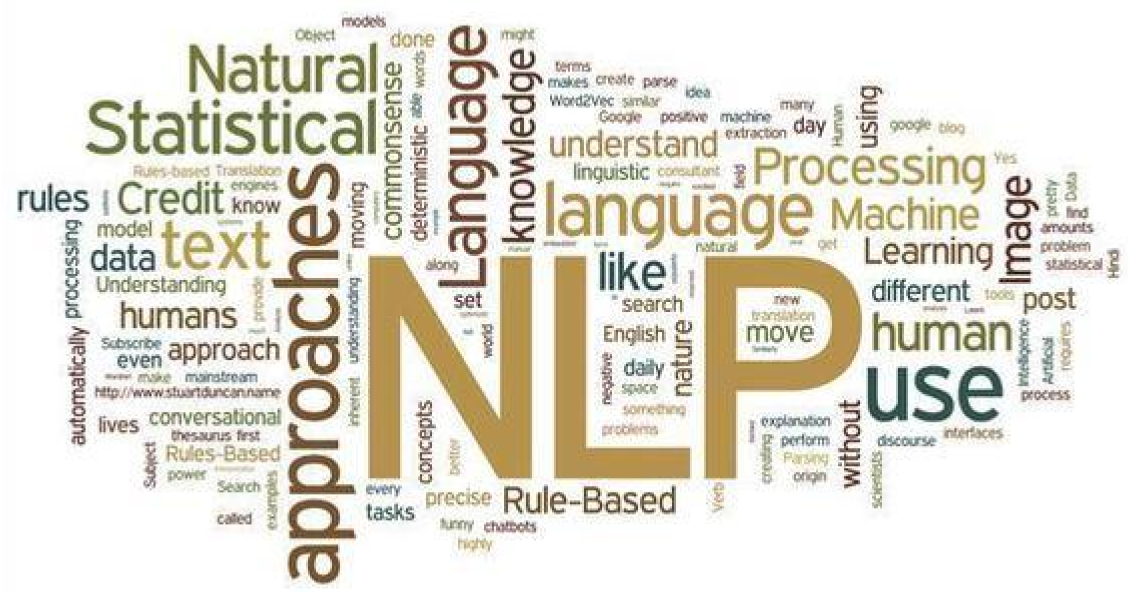
人工智能替代性分析



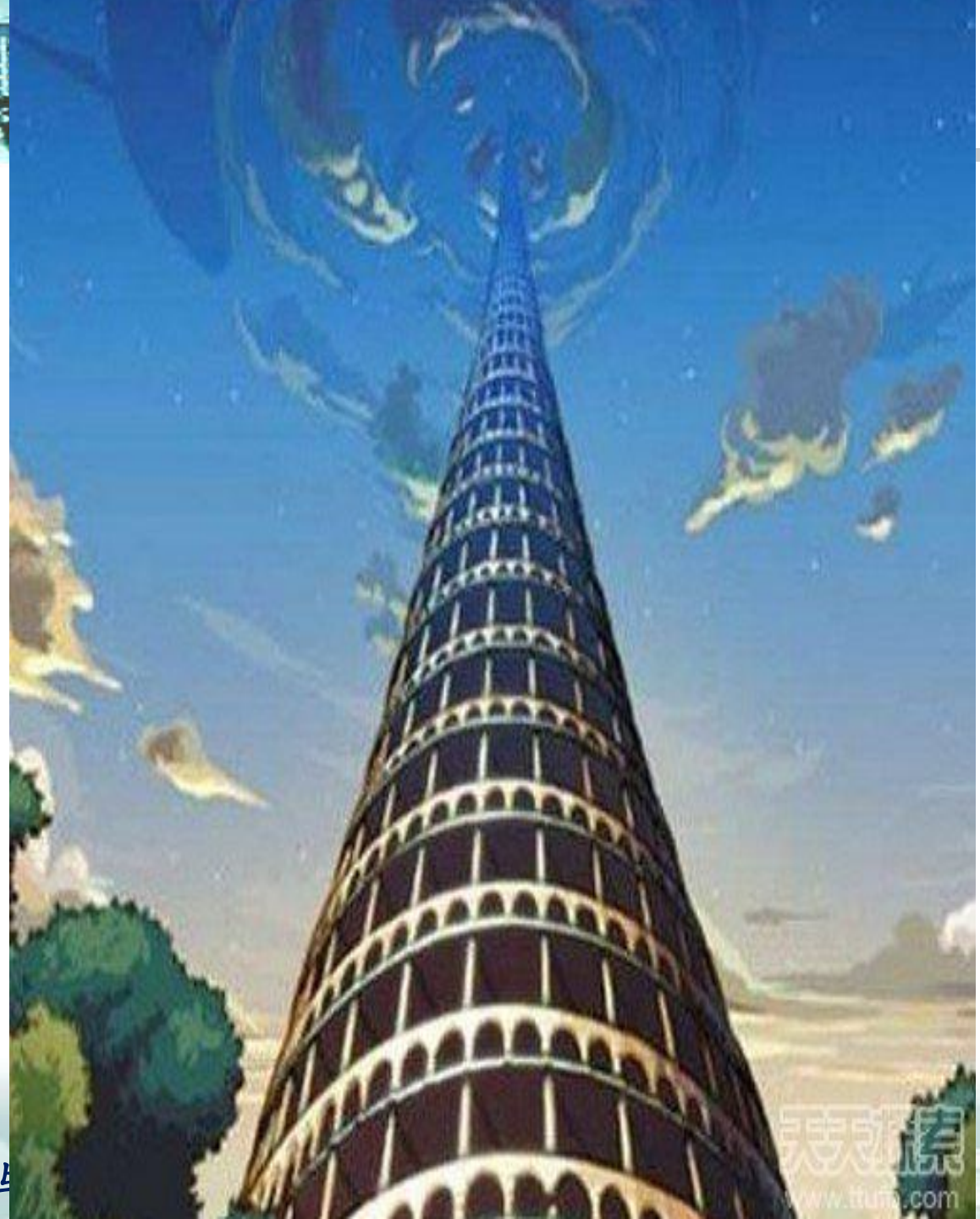
注：图来自于李开复报告

人工智能不能干什么？

- 跨领域推理
- 抽象能力
- 审美
- 知其然知其所以然
- 常识推断
- 自我意识
- 情感


NLP
approaches
Language
language
Statistical
Natural
Processing
Machine
Learning
understand
knowledge
commonsense
deterministic
Credit
text
data
rules
humans
approach
Understanding
conversational
Rules-Based
Rule-Based
like
use
human
different
Image
post
move
search
English
daily
nature
without
concepts
precise
tasks
linguistic
Word2Vec
http://www.stuartduncan.name

认识自然语言处理



巴别通天塔之惑

NLP两大挑战：

歧义消解
人类知识

人工智能？

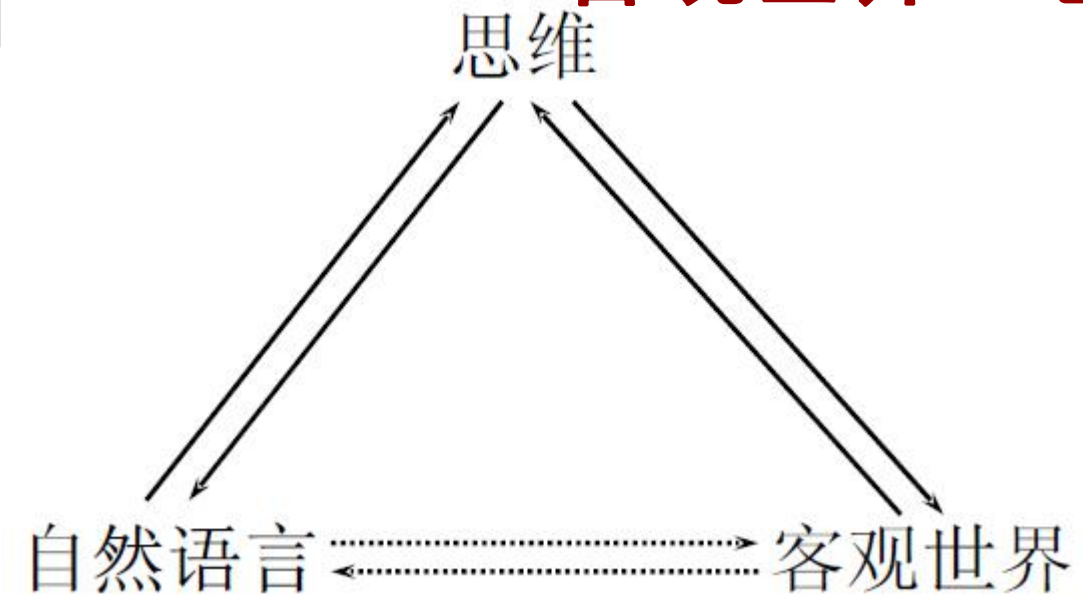
- 足球队和乒乓球队：一个谁也打不过，一个谁也打不过。

右面的鸡才是最好的

[illegible]

我们一定要把鸡打成右派!!!

客观世界->思维->自然语言



- 衰减效应：
  - 思维最多只能反映80%的客观世界；
  - 自然语言只能反映80%的思维：词不达意，答非所问；
  - 听众最多只能听懂80%；
  - 听懂的部分只有80%能反映到思维中；
  - 分析客观世界的最多只能利用80%。

相关概念

- 自然语言处理：技术概念
  - is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora.
- 计算语言学：学科概念
  - Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective.
- 文本挖掘：应用概念
  - is the process of deriving high-quality information from text.

计算语言学定义

计算语言学是一门以计算为手段对自然语言进行研究和处理的科学。

Computational Linguistics

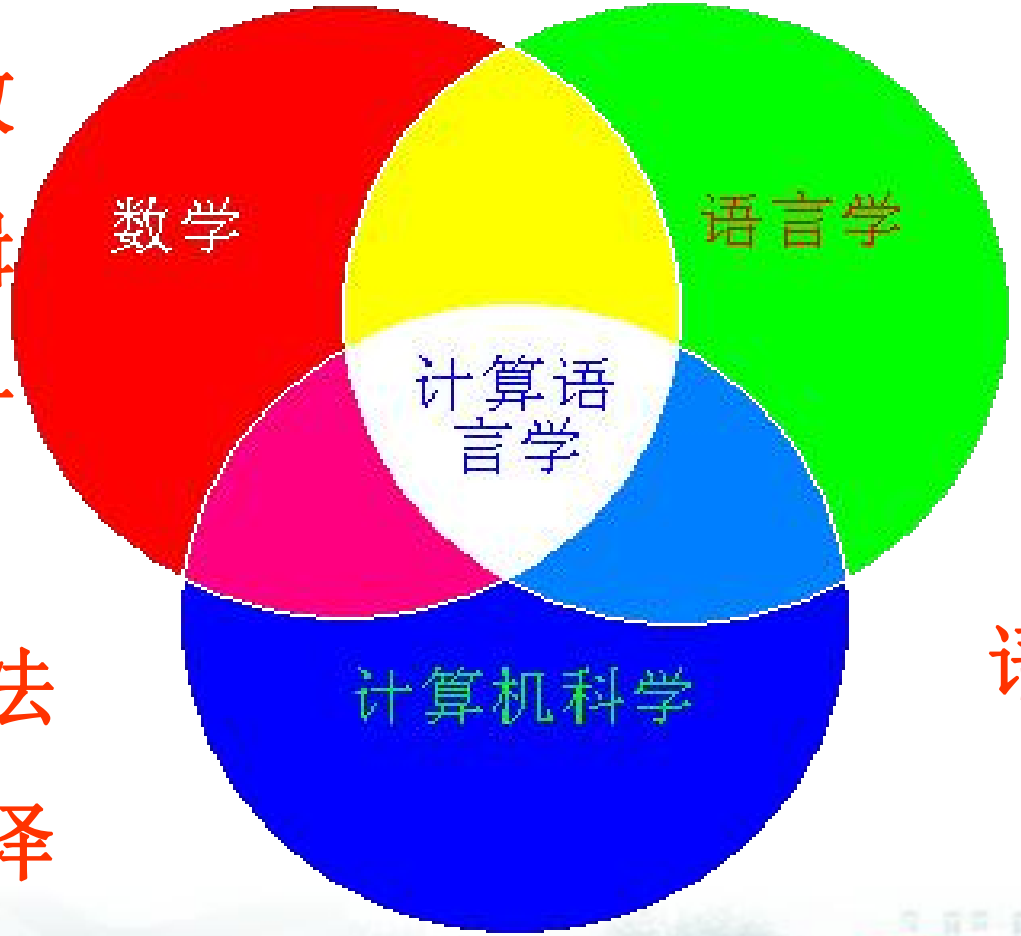
Natural language processing

"Wherever there is Artificial Intelligence, there is Artificial Stupidity."

“哪里有人工智能，哪里就有人工愚蠢”。

(从一到万，seed->see+ed)

计算语言学与其他学科的关系



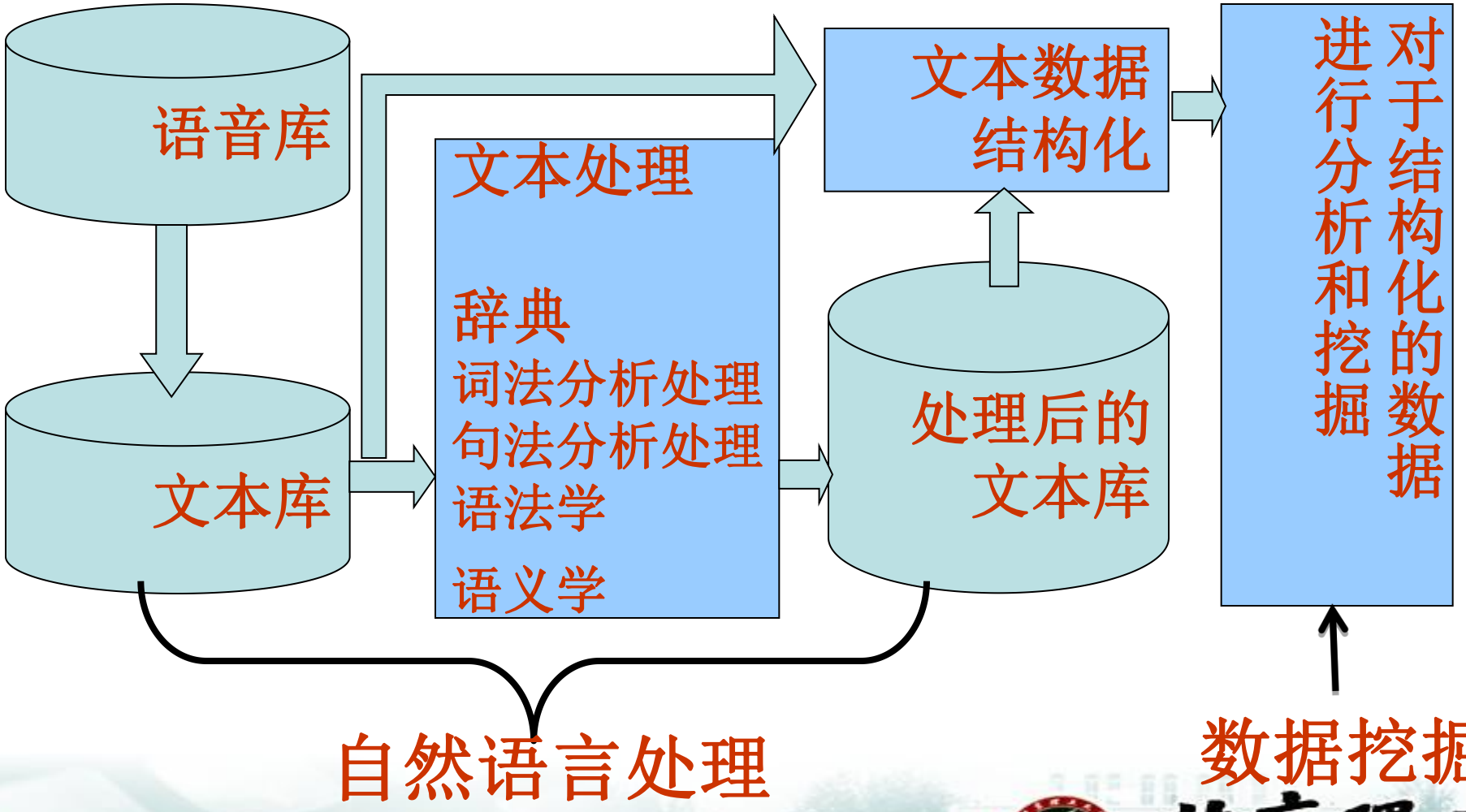
代数
逻辑
统计

词法
句法
语义
词汇
语料库

算法
编译

人工智能

文本挖掘的框架

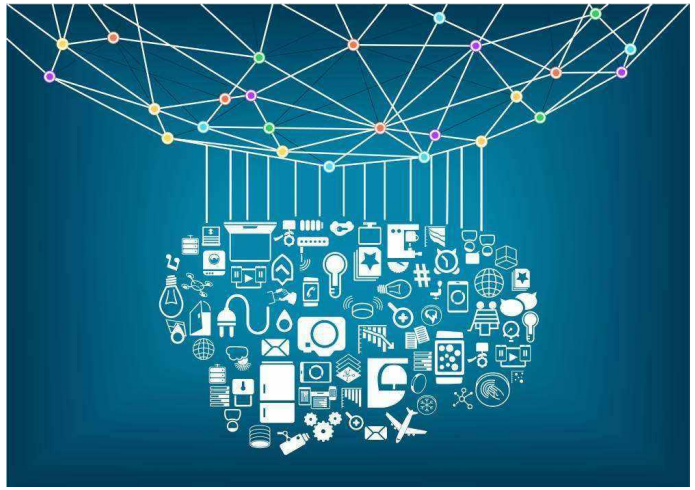


自然语言处理框架体系

- 基础理论
  - 自动机 形式逻辑 统计机器学习，汉语语言学 形式语法理论
- 语言资源
  - 语料库 词典
- 关键技术
  - 汉字编码 词法分析 句法分析 语义分析 文本生成 语音识别
- 应用系统
  - 人机对话，机器翻译，社交网络分析， 分类，聚类，检索，过滤，信息抽取，音字转换

NLP主要算法体系-编码识别为例

- 理性主义（规则方法）：Rule-Based
  - 自动机 形式逻辑 统计机器学习，汉语语言学 形式语法理论
- 经验主义（统计方法） ：Statistics-Based Language Model
  - 线性：N-Gram; Bayes；Maximum Entropy；Hidden Markov Model；Conditional Radom Fields; Support Vector Machine;
- 脑认知：神经网络，深度神经网络; CNN; RNN；LSTM;



大数据智能应用

NLPIR大数据搜索与挖掘实验室

定位： 多语种· 多模态· 语义智能一流研究中心

20年NLP技术积累

- ✓ **中文信息处理最高奖**钱伟长一等奖（国内唯一分词方向）
- ✓ **算法：** 20+自主可控全链路NLP模块
  20M 缓存边缘计算
  数据： 10GB语料库，20亿语义知识库
  知识： 10+行业先验知识积累

人、事、物、时空增强分析

- ✓ **多模态融合：** NLP+OCR+语音+图文比对的语义增强分析平台
- ✓ **启动：** <100份 小样本冷启动
  分析： KGB知识图谱关联分析
  生成： 报告智能生成

巡场订阅模式

- ✓ **国内：** 400+ 本地部署标杆客户
  军工、中央网信办、公安部、国研中心、人行、建行、中电科、航天科工、国家电网、华为等
- ✓ **全球：** 40万用户验证，新闻集团、韩国RSN、意大利大使馆、新加坡南洋理工、日立等

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

多语种智能信息处理专委会

- 多语种智能信息处理的国际化产学研用权威公益性组织
- 未来主要为多语种情报服务，服务“一带一路” 战略。

NLPIR实验室主任

张华平 博士

- ICTCLAS汉语分词创立者

  创建并运营NLPIR大数据语义增强分析平台

- 北京理工大学副教授，大数据搜索与挖掘实验室主任
- 中国人工智能学会多语种智能信息处理专委会秘书长

✓ 中文信息处理领域最高奖：钱伟长中文信息处理一等

（全国唯一“分词”方向）

✓ 新疆自治区科技进步

✓ 第一届ACL-SIGHAN国际汉语分词大赛

✓ 国家973汉语评测

✓ 中央网信办、中宣部、公安部等部委特聘技术顾问

✓ 国办电子政务总体组评审专家


证 书

钱伟长中文信息处理科学技术奖

为表彰在推动我国中文信息处理科学技术进步工作中做出贡献者，特颁发此证书，以资鼓励。

项目名称：ICTCLAS汉语词法分析系统
获 奖 者：张华平
奖励等级：一 等
获奖日期：2010.12.18
证 书 号：2010-1-2-2

中国中文信息学会
2010年12月18日

钱伟长一等奖证书


新疆维吾尔自治区科学技术进步奖

奖 励 证 书

为表彰在科学技术进步中做出重大贡献的个人，特颁发自治区科学技术进步奖，以资鼓励。

获奖成果：维吾尔文目标站点内容监测关键技术研究与系统开发
主要完成人：张华平（名列第三）
奖励等级：二等奖
奖励年度：2016年度

新疆维吾尔自治区人民政府
2017年09月21日

证书编号：620160805

新疆科技进步二等奖

张华平教授受CCTV采访解读苹果FBI揭秘大战。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

研究方向布局

- 基础算法：深度学习、小样本学习、终身学习、提示学习；
- 语义增强分析
  - 自然语言处理：中英文、小语种；
  - 多模态处理：OCR、语音识别
  - 自然语言生成：自动摘要、特定领域GIS推理与生成
- 知识图谱
  - 知识图谱构建与推理应用；汽车营销
- 社交网络分析
  - 特定群体跟踪；
- 应用布局
  - 智能情报：舆情、军事情报、市长热线、汽车营销
  - 智能文书：档案处理、文档核查、辅助生成

承担的科研课题

- 面向互联网信息的司法舆情监测与分级预警技术研究及系统研发 2018YFC0832304 “公共安全风险防控与应急技术装备”重点专项（司法专题任务）
- 国家自然科学基金面上项目2项，语义主题与社交关系融合的特定群体发现关键技术研究 61772075
- 国家242信息安全计划十余项 ；BB文本校对
- ZF文本生成某课题
- GF创新特区:终身自主学习、配置安全、内容安全
- GF战略：人工智能、未来网络、未来电子、生物交叉

基于NLPIR引擎的语义增强分析平台

场景层

公安：警情分析 | WW预测 | 重复报警

军工：情报平台 | 智能参谋 | 文书核查 | 会议纪要 | 辅助定密

政府/企业：投诉分析 | 会议纪要 | 公文核查 | 舆情监测

金融：合规风险 | 票据比对 | 信息抽取

科教：论文核查 | 专题平台

国际化：中文处理 | 汉语教学

…：… | …

产品层

公共安全语义智能分析 | 智能文本核查 | 智能会议纪要 | 知识图谱加工厂

NLPIR语义增强分析平台（NLPIR语义分析引擎+JZSearch搜索+九眼过滤+KGB知识图谱工具等）

技术层

数据导入：精准采集 | OCR | 语音识别 | DB导入 | 文档解析

多语分析：分词标注 | 依存文法 | 情感分析 | 语言模型 | 文本校对

信息抽取：新词发现 | 摘要提取 | 实体抽取 | 关键词提取 | 过滤与脱敏

文本挖掘：文档去重 | 相似比对 | 文本分类 | 文本聚类 | 话题分析

知识关联：全文检索 | 地图可视化 | 词向量扩展 | 知识生成 | 知识推理

基础层

算源：语料库 | 知识库 | 外部库

算法：专家启发 | 机器学习 | 终身学习

算力：国产芯片 | 国产OS | 可控云

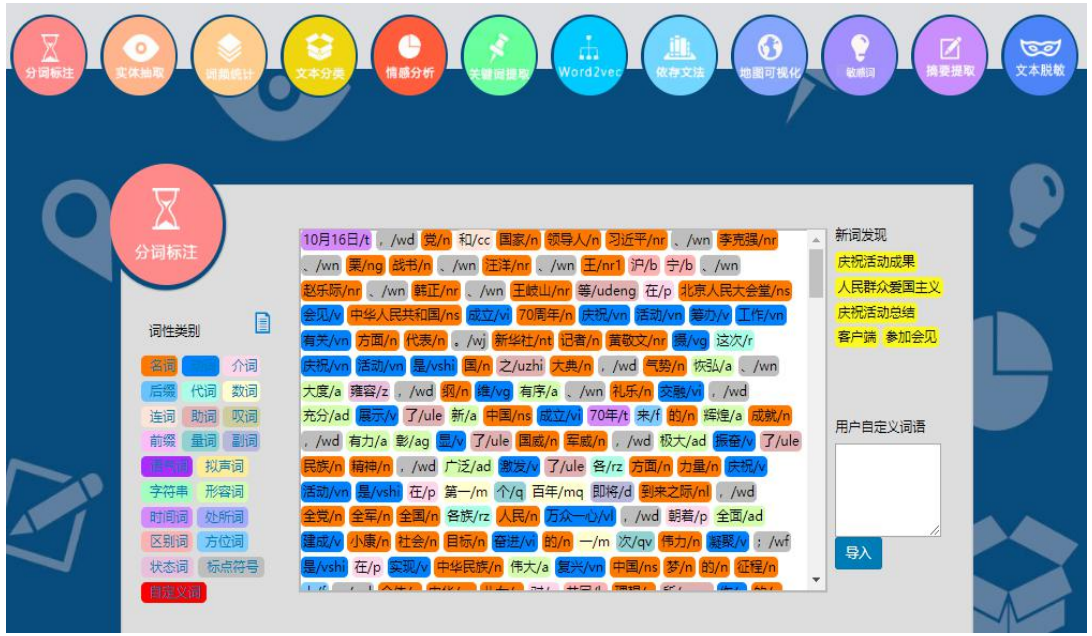
NLPIR inside

语义增强分析平台

- ✓ 具备70-80%应用功能
- ✓ 一周实现目标测试
- ✓ <100份样本冷启动
- ✓ 降低20-30%开发成本
- ✓ 提升系统智能化水平

核心技术：NLPIR语义分析引擎

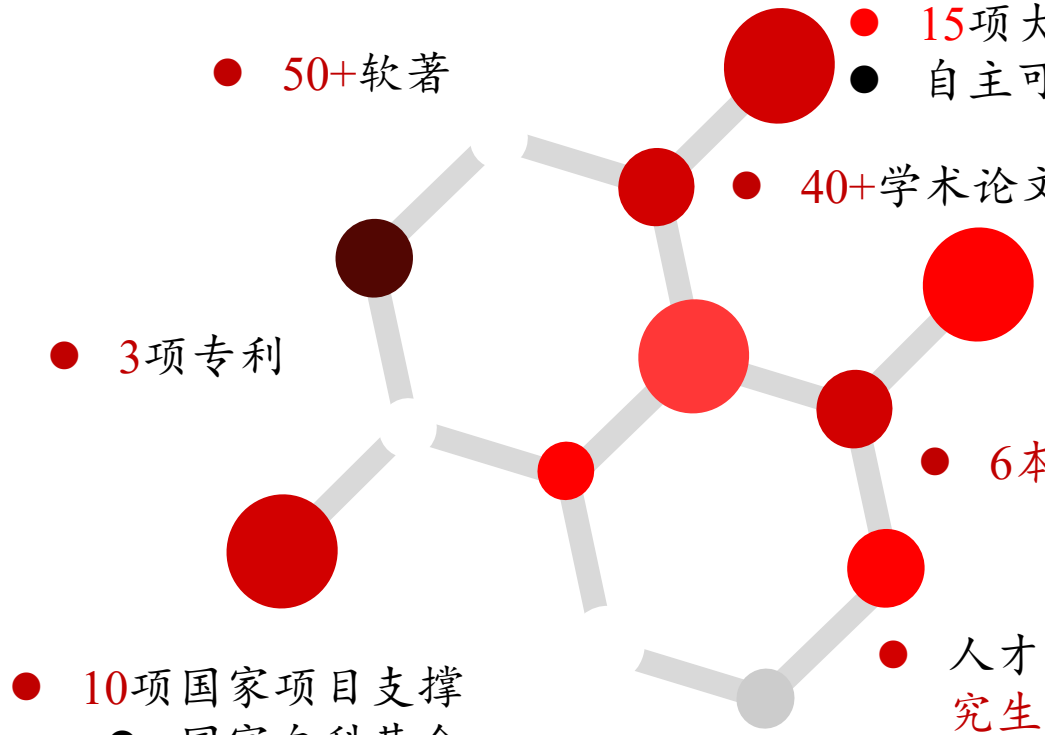
“20年技术迭代
帮助全球40万用户实时理解中文”



- 多语言智能分析：分词准确率接近99%，切分粒度可调整，融合20余部行业专有词典；新词发现、文本关键词提取：可处理海量网络文本数据（平均每小时处理至少50万篇文档）、关键词按照影响权重排序；一带一路小语种分析：维吾尔语、藏语、粤语、Hindi、缅甸、印尼
- 文本生成：自动摘要、报告生成
- 语义边缘计算：20M缓存即可部署
- Demo地址：http://ictclas.nlpir.org/nlpir/

主要成果

- 1个NLPIR大数据语义智能分析平台
- 15项大数据语义分析算法组件
- 自主可控，全面支持国产CPU/操作系统



- 50+软著
- 40+学术论文，SCI/EI 30+
- 3项专利
- 6本专著
- 人才培养：70+博士硕士研究生，网信十佳讲师1人
- 10项国家项目支撑
  - 国家自科基金
  - 国家重点研发
  - 国家242
  - ZF预研/GF创新特区
- 1000+论文引用
- 服务在线用户11.9亿人
- 近3年42万+技术开发用户
- **直接经济效益12.4亿元**

社会效益

- **国家层面：**NLPIR已经应用于政府机构等国家部门，极大程度提高了其文本处理效率，提升了国家智能化水平。

- **社会层面：**NLPIR也出口到美国NCR、新闻集团、意大利ExpertSystem公司、韩国RSN公司、日本NEC与日立、新加坡南洋理工大学等国际知名机构。与华为、人民网、长安汽车也建立了合作；

- **科研层面：**NLPIR无偿提供在线分析平台和免费授权供广大自然语义处理的研究人员使用，就百度学术和知网对NLPIR/ICTCLAS的引用量已经均超过1000，Github的start量和Fork量也均超过1000。从2018起开发用户数超过42万。依据人民网2020年年报，人民网 PC 端、手机网、客户端、新媒体账号、代运营平台、聚合分发平台在全球覆盖直接用户数超过 11.9 亿。

奖励

科技

钱伟长中文信息技术处理科学技术一等奖

政府

新疆维吾尔自治区科学技术进步奖二等奖

、河北技术成果资质

行业

新闻出版广电总局知识服务重点实验室-核心技术全国第二名

首席数据官联盟-中国大数据自然语言处理方向全国第一名

网络舆情挖掘系统、中文分词标注等20+项软件著作权

完美双数组管理和检索方法、互联网信息整合与更新方法等多项专利

学术认证

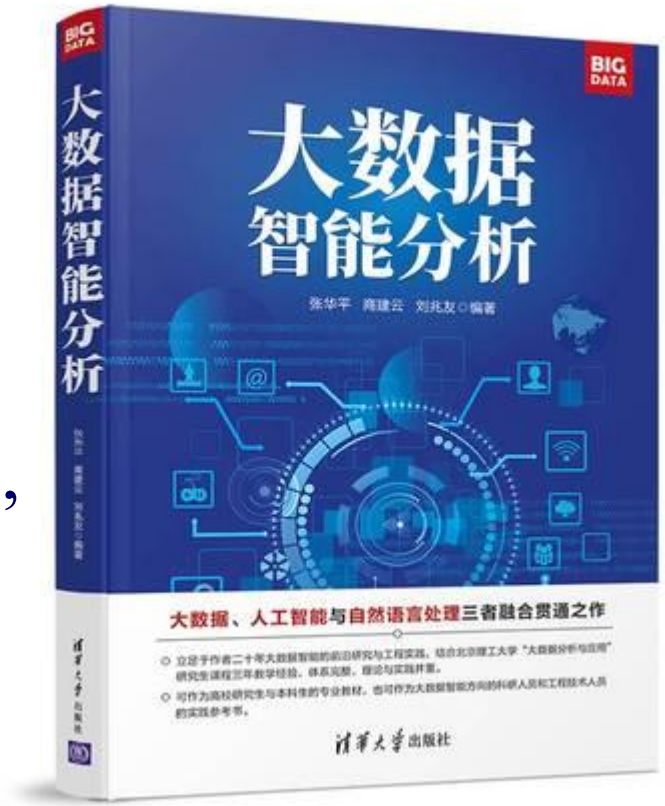
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

出版的专著

Huaping Zhang et, Big Data Intelligence (Book), Tsinghua Univ. Press, 2019

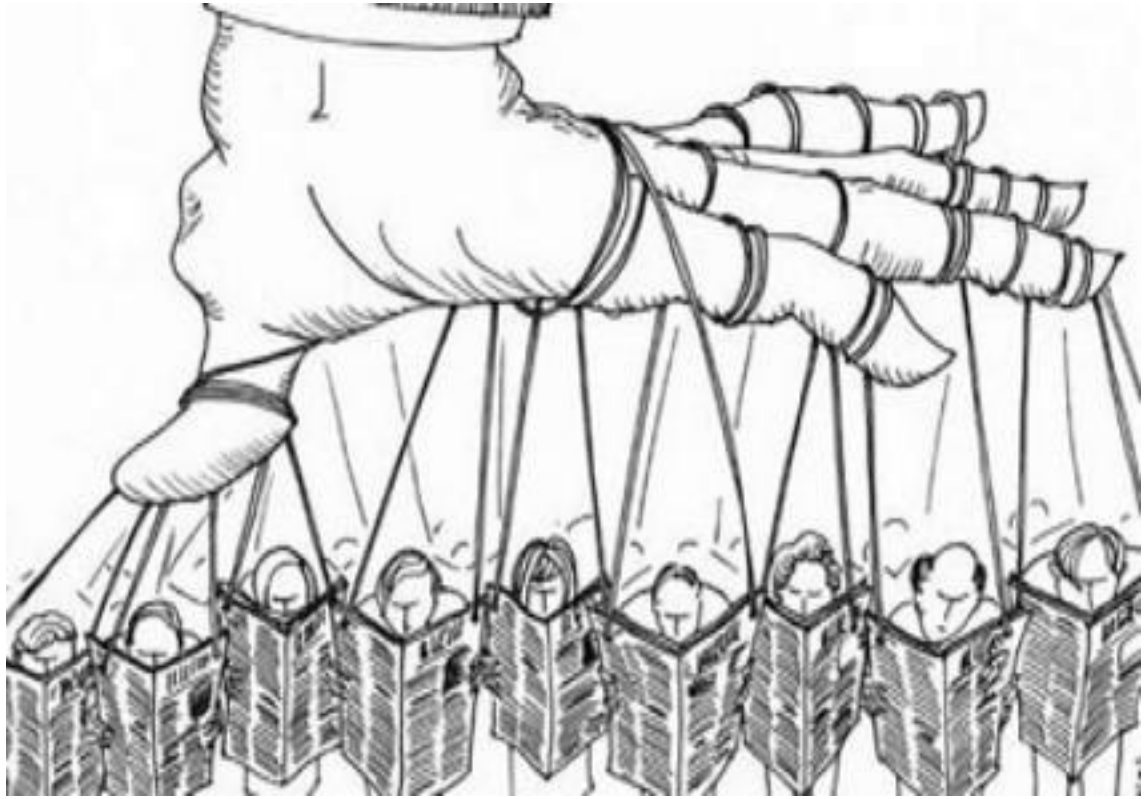
Huaping Zhang et, Big Data Big Talk (Book), Publishing House of Electronics Industry, 2019
ISBN:978-7-121-30181-0

Huaping Zhang et, Big Data Search and Mining (Book), Science Press, 2014.5（ISBN:978-7-03-040318-6）



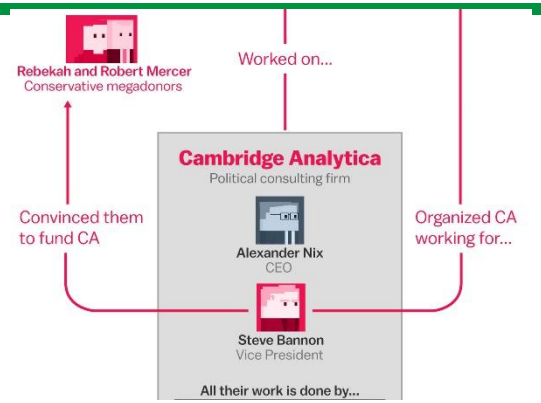
北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

最恐怖黑客：黑掉你的脑子



社交舆情操控

- 剑桥分析控制62国选举



- 美国2016大选社交网络干预

- ISIS全球吸引恐怖分子

社交舆情操控

目标群体

Cambridge Analytica

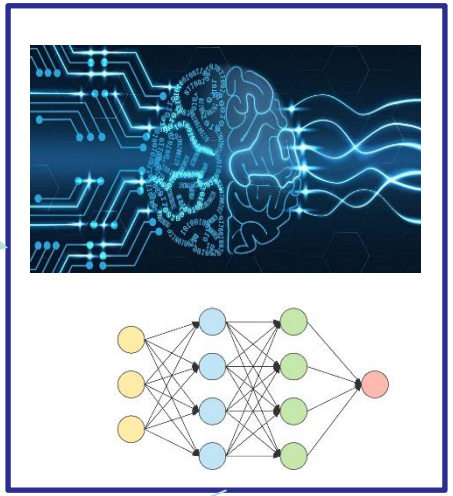
RUSSIAN INSTITUTE FOR STRATEGIC STUDIES

ISIS

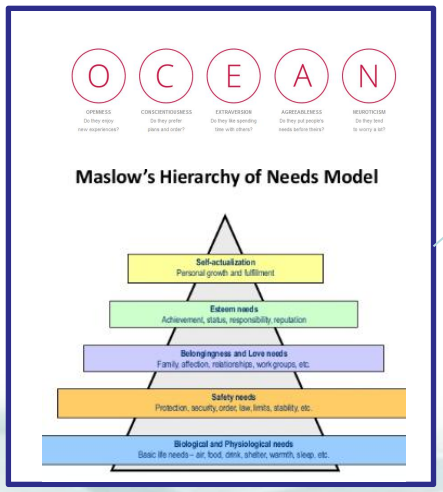
舆论操控者

社交网络

计算模型



心理学模型



舆论操控

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

剑桥分析:方法论

5,000 data points per person

We collect up to 5,000 data points on over 220 million Americans, and use more than 100 data variables to model target audience groups and predict the behavior of like-minded people.

Constant testing & improving

Our data scientists and psychologists are constantly testing new modeling and research techniques to ensure all our data sets and audience segments are the most advanced in the market.

OCEAN and the Big Five

We use the established scientific OCEAN scale of personality traits to understand what people care about, why they behave the way they do, and what really drives their decision making.

O	C	E	A	N
OPENNESS	CONSCIENTIOUSNESS	EXTRAVERSION	AGREEABLENESS	NEUROTICISM
Do they enjoy new experiences?	Do they prefer plans and order?	Do they like spending time with others?	Do they put people's needs before theirs?	Do they tend to worry a lot?

Spot the differences

Understanding the complex web of OCEAN personality traits behind behavior lets us see why people who look similar on the surface often want and respond to completely different things.

Conversations that move people

When you go beneath the surface and learn what people really care about you can create fully integrated engagement strategies that connect with every person at the individual level.

Same demographics, different personalities

Female
25-35 Years old
AMEX User

Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism

People with high openness and extraversion love new experiences they can share with lots of people.

Female
25-35 Years old
AMEX User

Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism

People with low openness and extraversion really value down time spent with their closest friends.

We Call This Behavioral Microtargeting

Discover. Understand. Engage. Repeat.

Combine our full suite of data-driven audience insight and engagement techniques with our unique and powerful Behavioral Microtargeting service that constantly learns, improves and delivers.

With Behavioral Microtargeting you'll be able to anticipate the needs of your customers and predict how their behavior will change over time, so you can build services, products and campaigns they really love.

Geographic View

Demographic View

Psychographic View

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

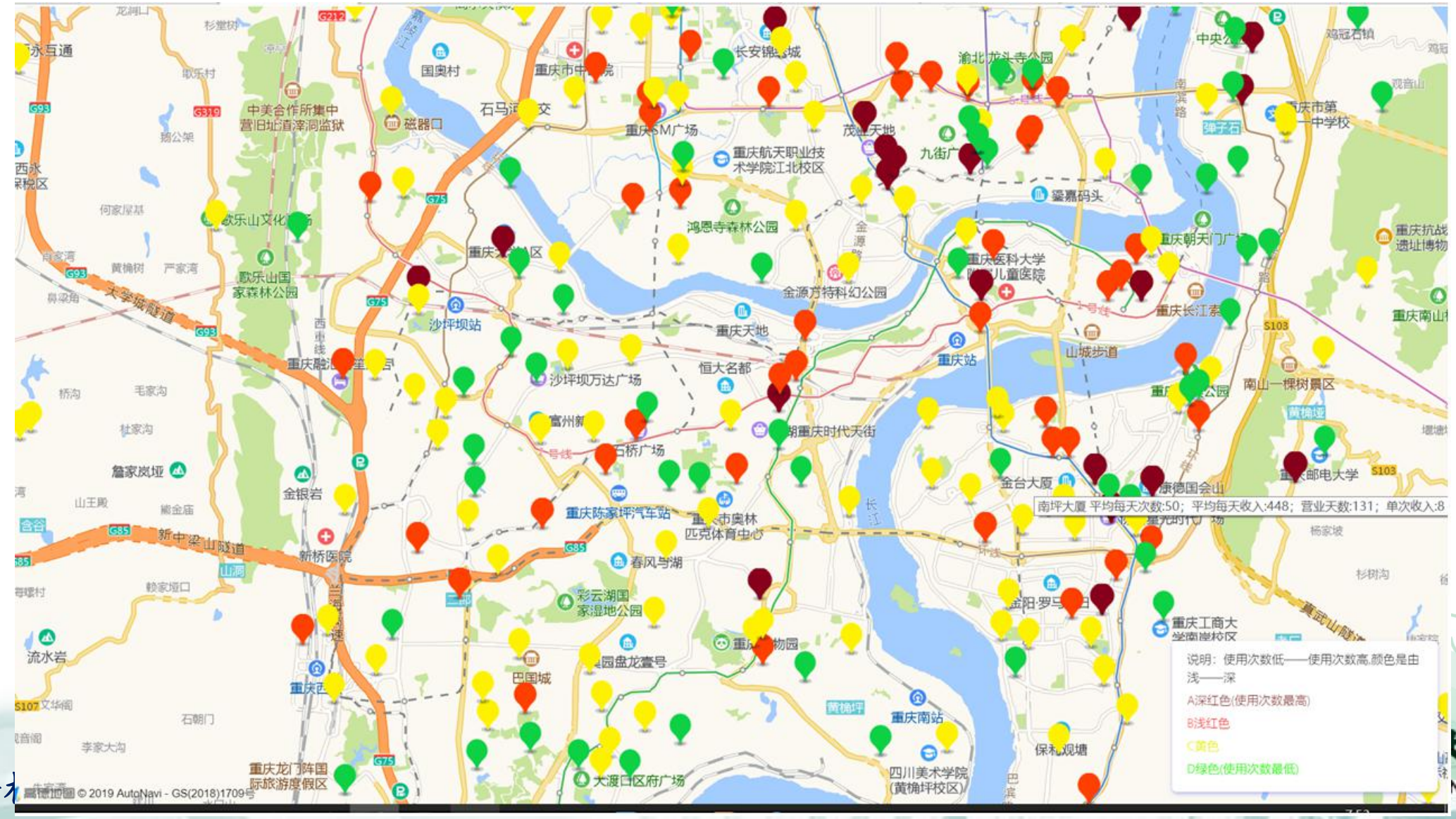
ISIS: 网络极端势力如何吸引恐怖分子

Table 1 Offender behaviour characteristics

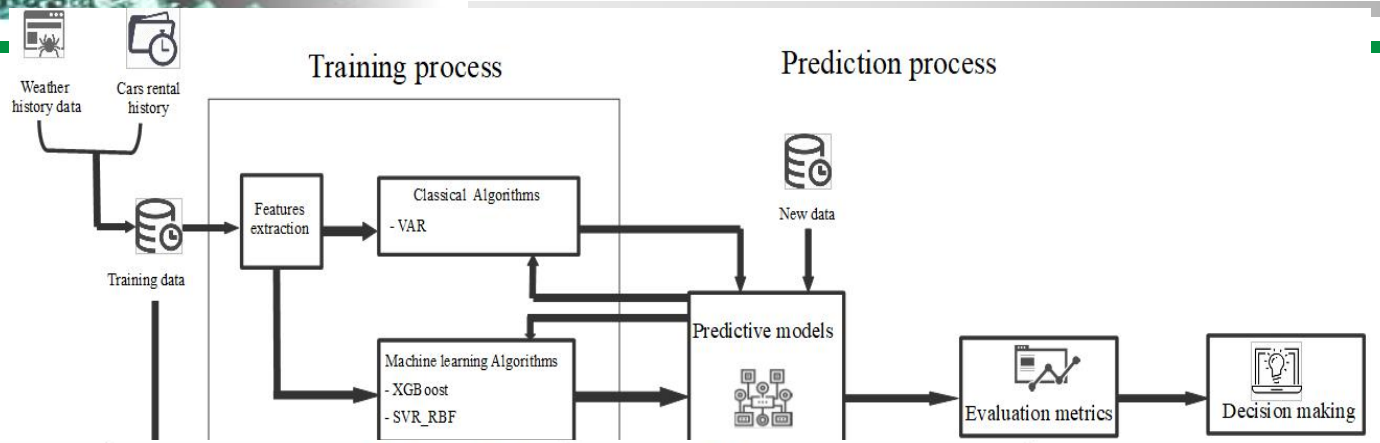
Type	Characteristics	Cases on Twitter	Cases on Facebook	Total No of Cases
Cyber Mobs	Using social media platforms to create a mob mentality and urging others to fight for the Isis goal. This is done through group posts, videos and comments of hate directing groups of Muslim's to fight. Often personified through retweets, likes and views of specific Isis propaganda materials.	78	55	133
Loners	Often done through individual posts and comments. This individual is someone who is attracted to the Isis campaign but clearly is exposed to individual grievances and has a lone mentality.	51	65	116
Fantasists	Someone using social media platforms to fantasise over the Isis movement. In particular, these individuals have blurred the lines between reality and fiction and are making direct plea's to fight for Isis.	45	94	139
Thrill Seekers	People who are promoting Isis propaganda through videos and posts and forums. Indeed, some of these individuals claim to be directly using the Internet for online extremist purposes. These individuals are describing the sense of adrenaline rush they are receiving by watching and partaking in fighting on the battlefield whether online or offline.	85	98	183
Moral Crusaders	These individuals are talking about the moral duty to fight. Many of these individuals are also constructing arguments based on ideology and theology as a means to promise people external rewards.	140	95	235
Narcissists	These people are using political, foreign policy and individual grievances as a means to whip up a climate of revenge seeking and wanting to fight for the Isis mission and goals.	166	104	270
Identity Seekers	Mostly this is users who appear to be seeking some form of identity. Primarily people searching for some form of masculinity and therefore the Isis recruitment drive appeals to them. This applies to males and females.	87	101	188

Awan I. Cyber-extremism: Isis and the power of social media. Society. 2017 Apr 1;54(2):138-49

某汽车出行宏观画像


南坪大厦 平均每天次数:50；平均每天收入:448；营业天数:131；单次收入:8
说明：使用次数低——使用次数高,颜色是由浅——深
A深红色(使用次数最高)
B浅红色
C黄色
D绿色(使用次数最低)
© 2019 AutoNavi - GS(2018)1709号

某汽车出行租车预测模型



	History + weather + weekend/weekday			History + weather			History + weekend/weekday			History		
	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
Class A (#16)	**0.29858**	**0.35530**	**0.59607**	0.40409	0.51835	0.71996	0.44704	0.61301	0.78295	0.68891	0.68891	0.83000
Class B (#104)	**0.15430**	**0.17264**	**0.41550**	0.21268	0.23436	0.48411	0.30067	0.38240	0.61839	0.59425	0.59425	0.77087
Class C (p#6)	**0.25730**	**0.31151**	**0.55813**	0.32569	0.49750	0.70534	0.42327	0.72102	0.84913	0.78899	0.78899	0.88825
Class D (#28)	**0.00083**	**0.00083**	**0.02888**	0.00250	0.00250	0.05002	0.02335	0.02335	0.15282	0.06839	0.06839	0.26152
Class E (# 25)	**0.01168**	**0.01251**	**0.11185**	0.01418	0.01501	0.12253	0.01585	0.02419	0.15552	0.02419	0.02419	0.15552

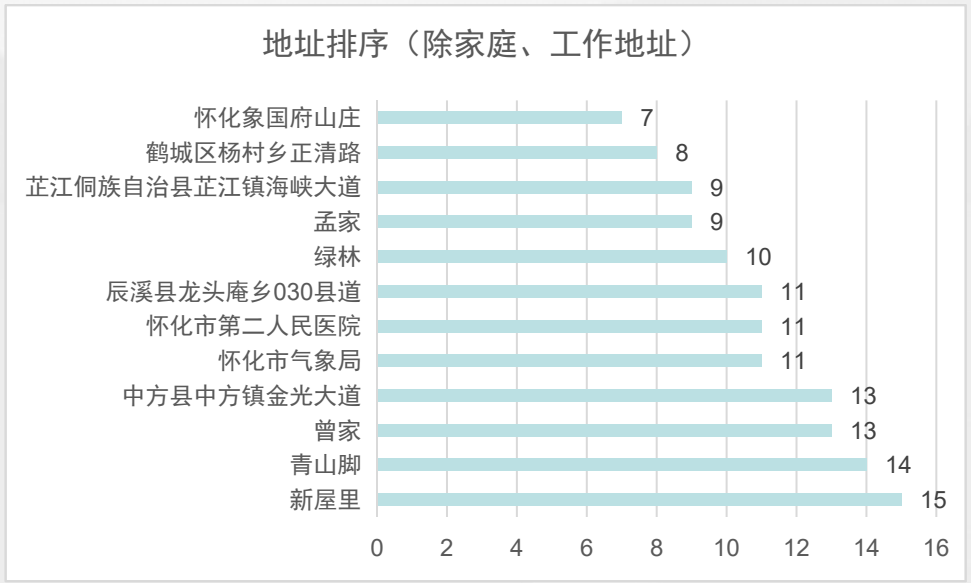
Table 7: evaluation metrics for multivariate time series forecasting using LSTM

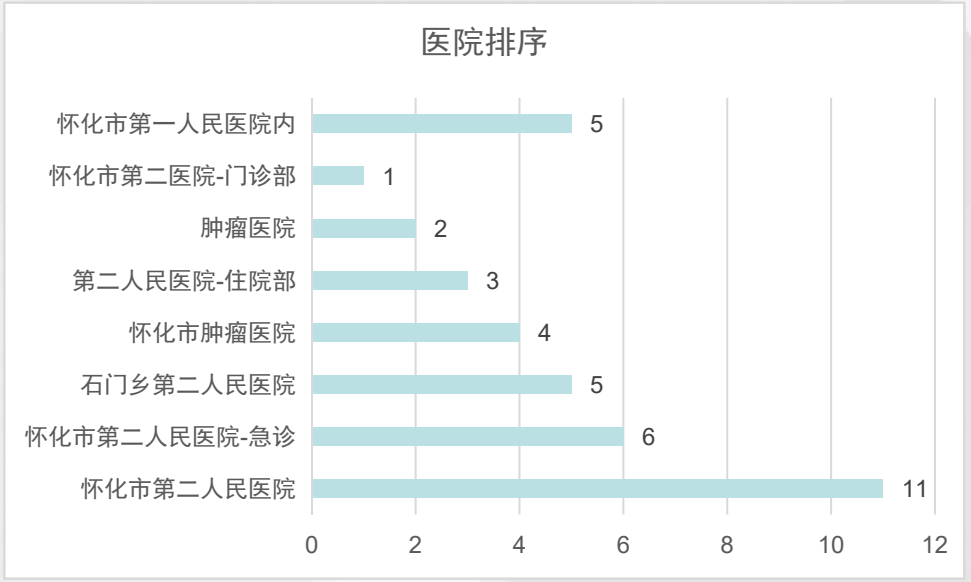
	History + weather + weekend/weekday			History + weather			History + weekend/weekday			History		
	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
Class A (# 16)	0.90898	**2.02258**	**1.42217**	**0.90889**	2.02560	1.42323	1.50311	3.15878	1.77729	1.50419	3.16149	1.77805
Class B (#104)	0.97605	**2.11016**	**1.45264**	**0.97571**	2.11519	1.45436	0.97568	2.11486	1.45436	0.97568	2.11487	1.45259
Class C (#6)	0.90898	**2.02258**	**1.42217**	**0.90889**	2.02560	1.42323	0.90901	1.42194	2.02192	0.90884	2.02483	1.42297
Class D (#28)	**0.60189**	**0.66971**	**0.81836**	0.60221	0.66972	0.81836	0.60192	0.81836	0.66971	0.60225	0.66972	0.81836
Class E (#25)	**0.16379**	**0.10897**	**0.33011**	0.16437	0.10903	0.33013	0.163867	0.10898	0.33013	0.16449	0.10905	0.33023

GPS人物画像：POI

- 除去可能是家庭、工作的地址，选其它高频地址分析4人除家和公司常去的地方。

湖南怀化气象工作人员



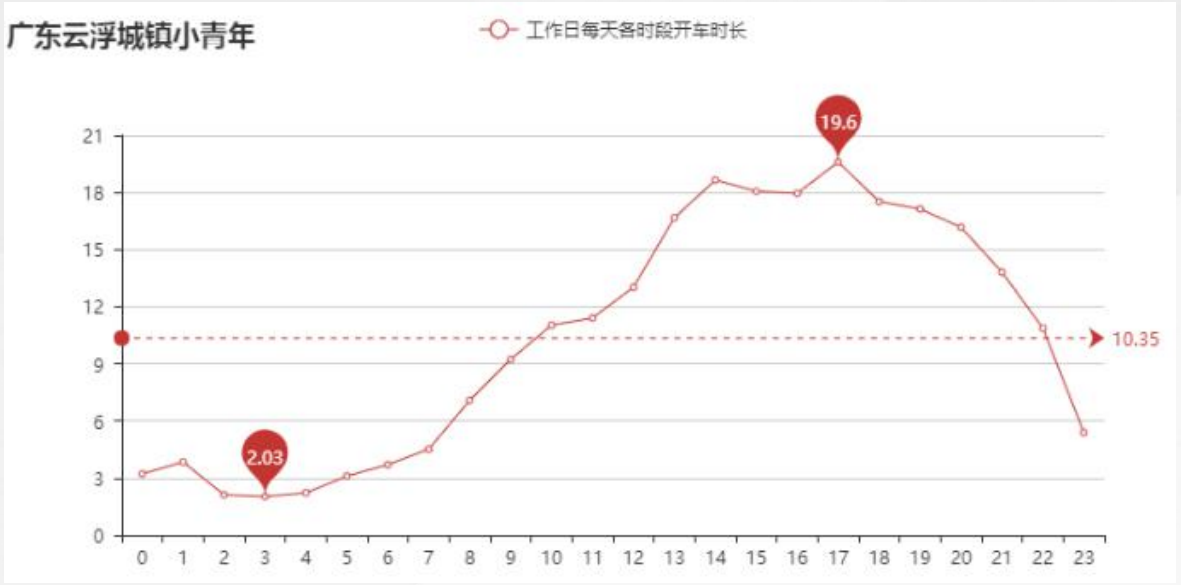


医疗：怀化市第二人民医院共出现32次，此人经常开车去医院，属于紧急情况用车较多的类型。其他高频地址比如：新屋里、青山脚、曾家、绿林、孟家都是村庄，距离家庭住址盛世嘉园车程约一小时左右，不属于休闲娱乐场所，可能是探亲访友用车。

GPS人物画像：工作性质

工作日每天各时段开车时长=工作日总开车时长/工作日总天数

高峰时段：7点和17点，典型早8晚5型用车
工作性质：政府单位



午后用车型：用车高峰时段14-22点
工作性质：非正式单位，自由职业者

智能机器阅读

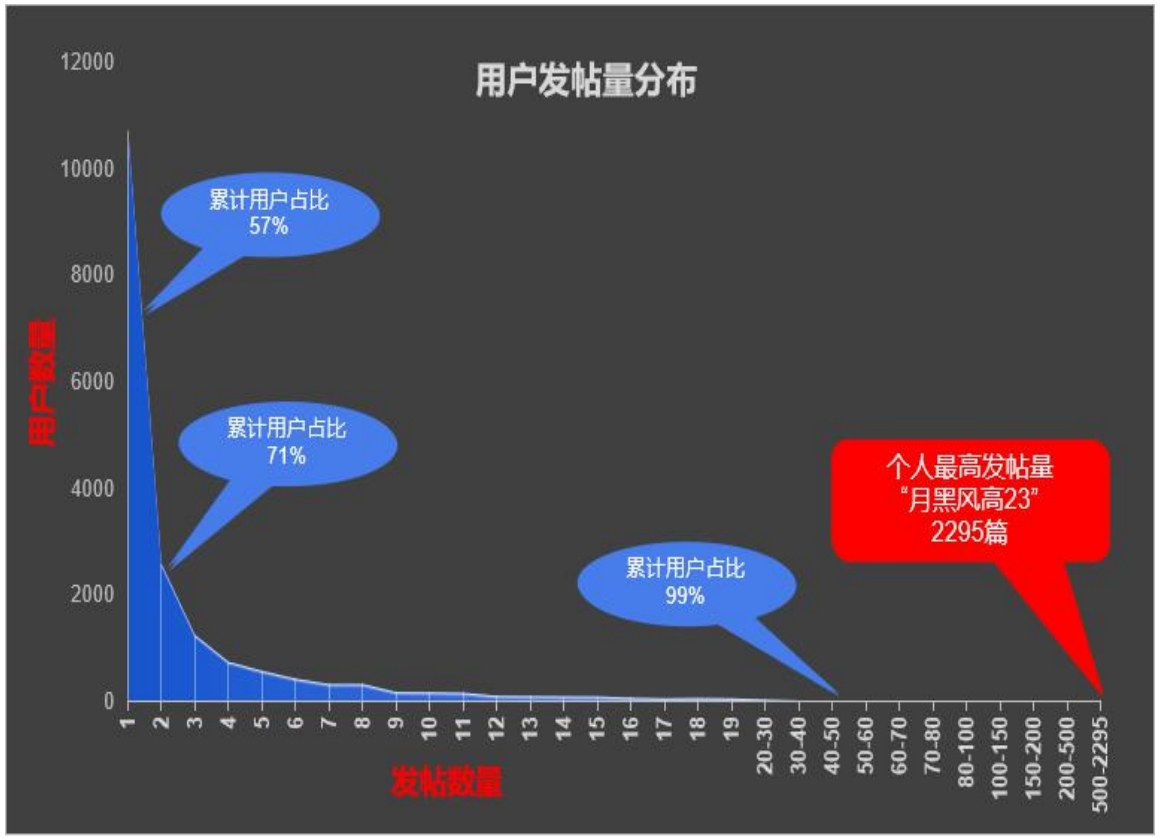
摘要提取

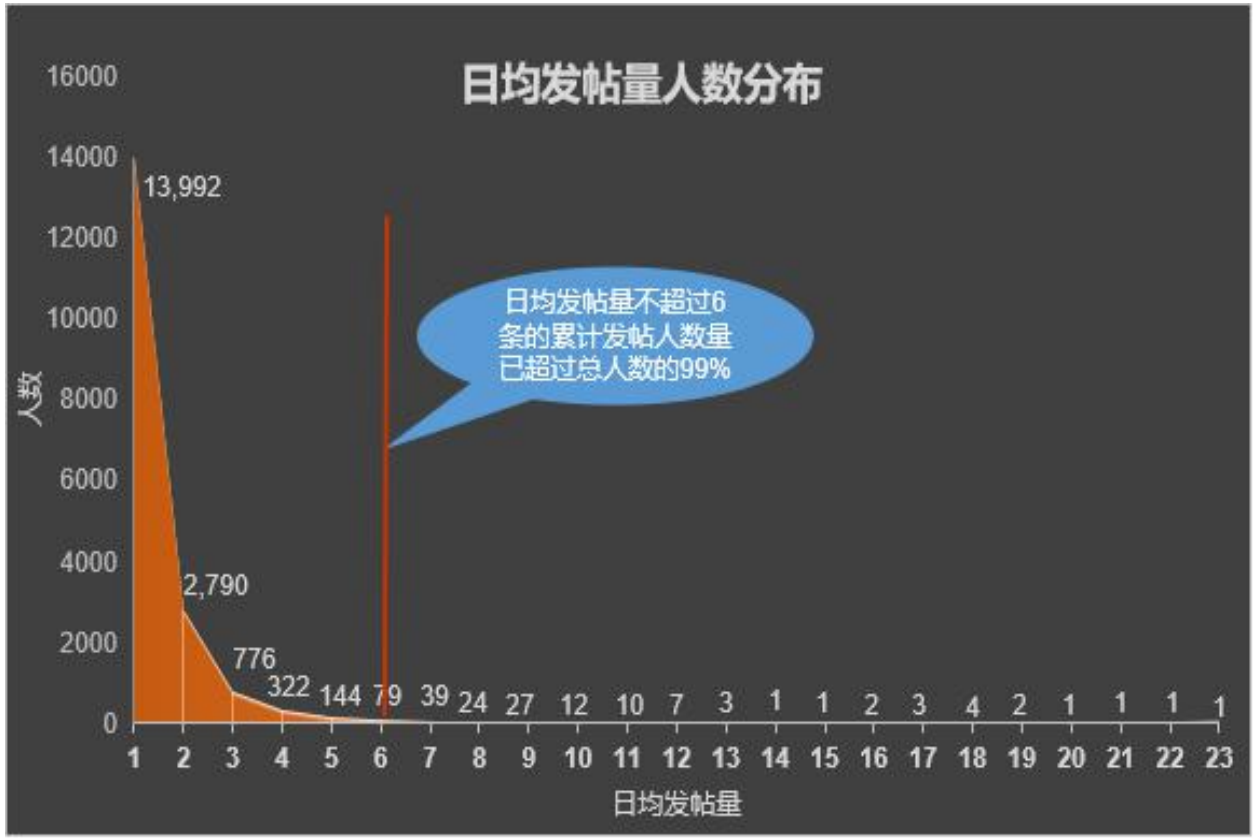
东盟加快发展数字经济（国际视点）--国际--人民网本报驻泰国记者刘慧2021年08月11日05:21来源：人民网 - 人民日
盟地区多国近期持续出台数字化发展战略，促进数字经济加速发展。数字支付业务也在东盟国家呈上升态势。中国和东
家抓住数字化转型机遇，在数字基础设施、5G、人工智能、大数据等领域打造更多合作亮点。

文档支持

某品牌舆论场画像

发帖人总量：18645个； 发帖总量：87484条； 人均发帖量：4.7条/人





北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

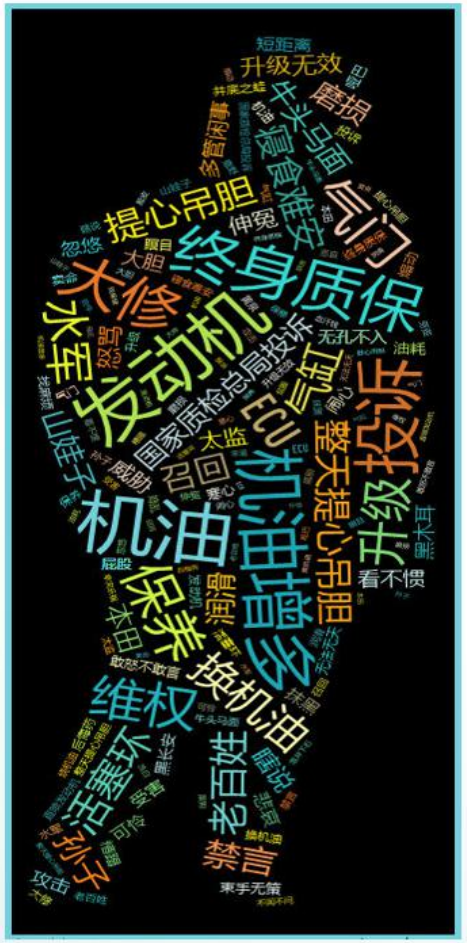
某品牌舆论场画像：从2万到200

发帖总量与日均发帖量都异常

发帖总量异常账号　　日均发帖量异常账号

某品牌舆论场画像：极端黑公关

发帖人“追逐人生 LXB”语义画像



敌对攻击号

用户 id：追逐人生 LXB

发帖数量：659 篇

平均每天发帖：30 条

情感得分：-205

涉及发帖标题：77 个

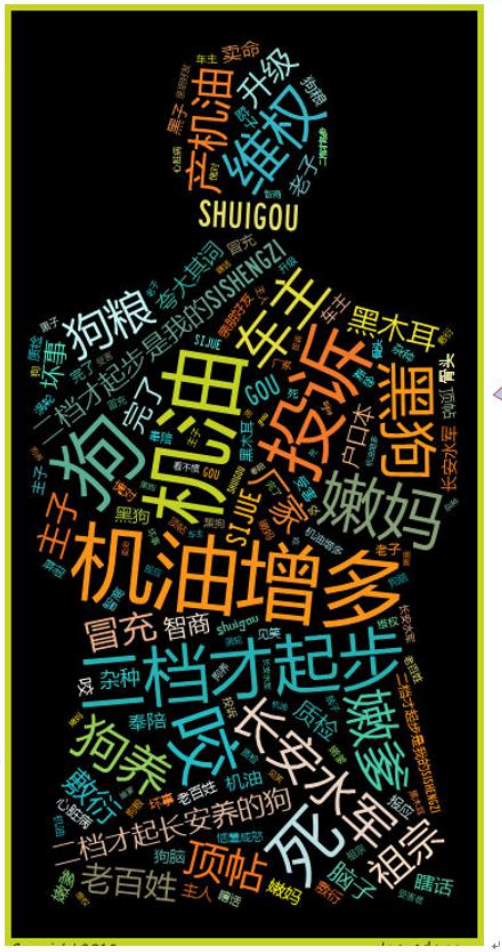
发布平台：

汽车之家（仅 1 个）

污言秽语极多！

狗、咬、死、嫩妈、嫩爹、黑狗、狗养、长安水军、狗脑、黑子、祖宗等

发帖人“长安非常不安”语义画像



发帖频率极高！

情感极其负面！

发帖总量：601 条

发帖频率：每天 50 条

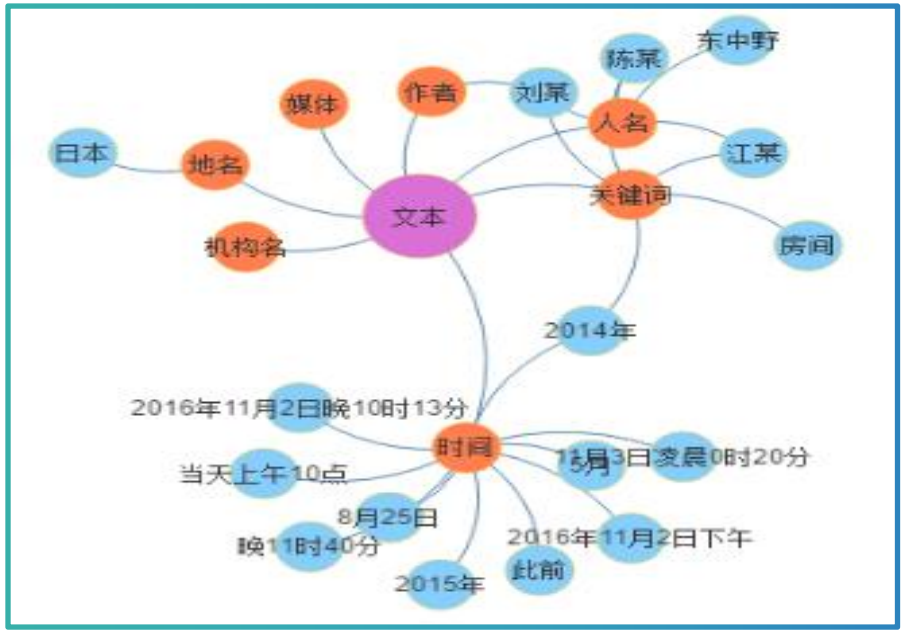
发帖情感：-710

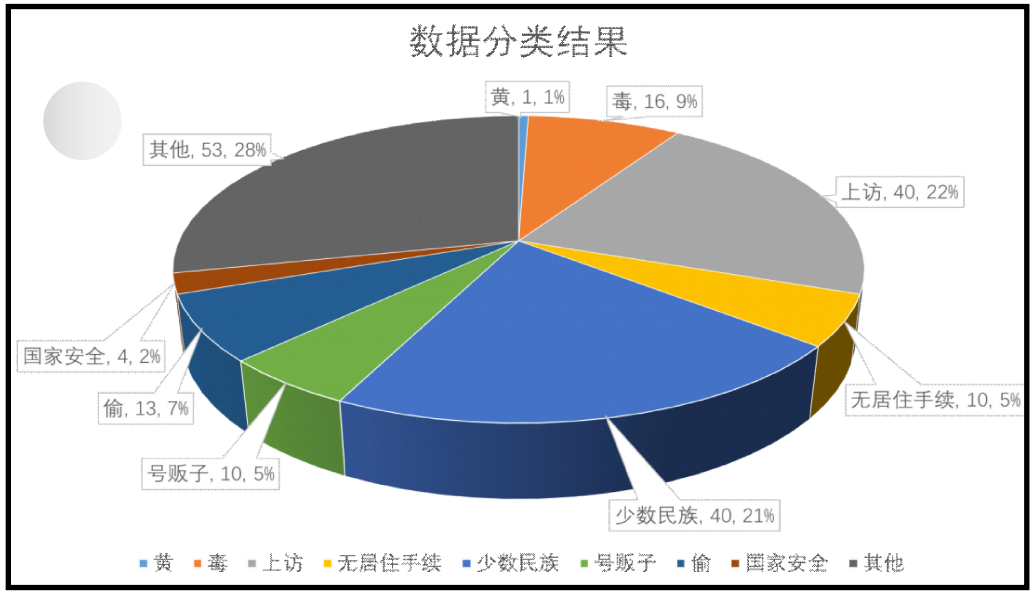
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

公安语义增强应用

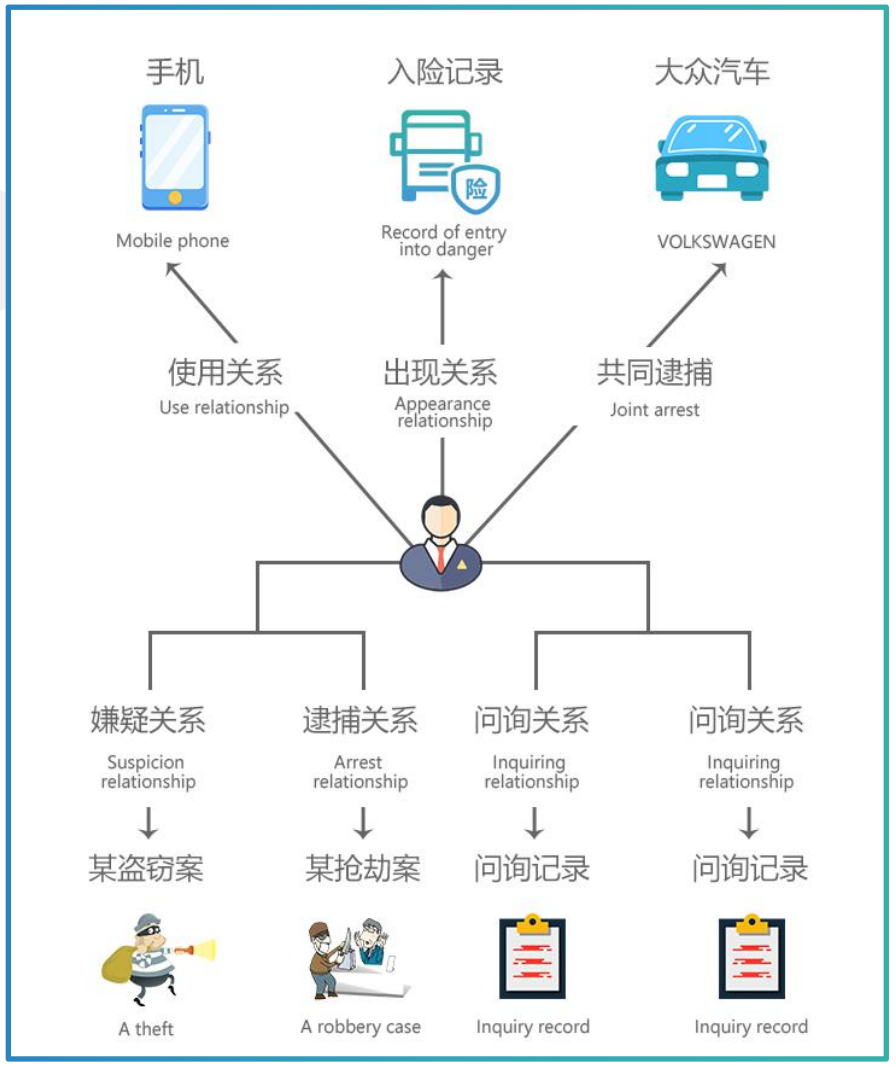
繁 拼

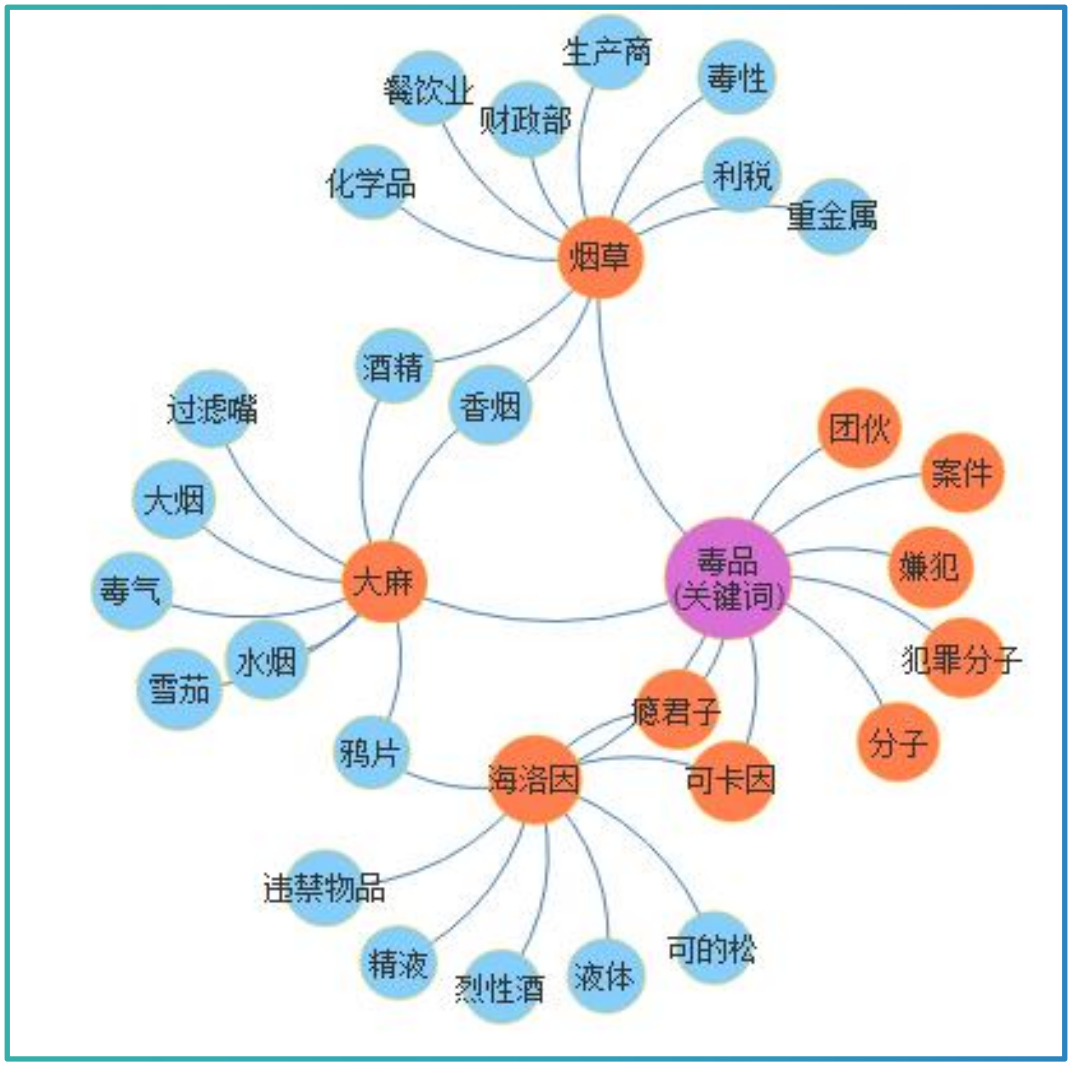
2014年，刘某到日本留学。
2015年底，刘某因与舍友相处不愉快，要求学校调整宿舍，被安排到姜某宿舍。
两人相识，并发现是老乡，两家相聚仅10公里。
5月起，刘某与结识不过一个月的陈某同居。
8月25日，双方发生矛盾，刘某被赶出家门，到打工店长家借住。
2016年11月2日下午，陈某找到江某住处，试图与刘某角色感情问题。
此时，刘某独自一人在房间里。应刘某的请求，善良勇敢的江某专程从外面赶回
2016年11月2日晚10时13分，江某到达东中野站。原本仅需10分钟即刻到家的江某
晚11时40分，陈某尾随刘某江某到达二人住所，约20分钟后开始敲门。刘某躲进
11月3日凌晨0时20分左右，在刘某两次报警间隙，孤立无援的江某被陈某用利刃
江某身上没有水果刀（凶器）的刀鞘部分。
案发当天上午10点多，陈某借过学校一个实验室的钥匙，检方称行凶所用的刀此
刀的塑料包装壳在实验室里被发现，刀刃套在案发现场二楼到三楼楼梯上找到。
被告人辩护律师称，凶器是一把水果刀





公安语义增强应用


手机
入险记录
大众汽车
Mobile phone
Record of entry into danger
VOLKSWAGEN
使用关系
Use relationship
出现关系
Appearance relationship
共同逮捕
Joint arrest
嫌疑关系
Suspicion relationship
逮捕关系
Arrest relationship
问询关系
Inquiring relationship
问询关系
Inquiring relationship
某盗窃案
某抢劫案
问询记录
问询记录
A theft
A robbery case
Inquiry record
Inquiry record

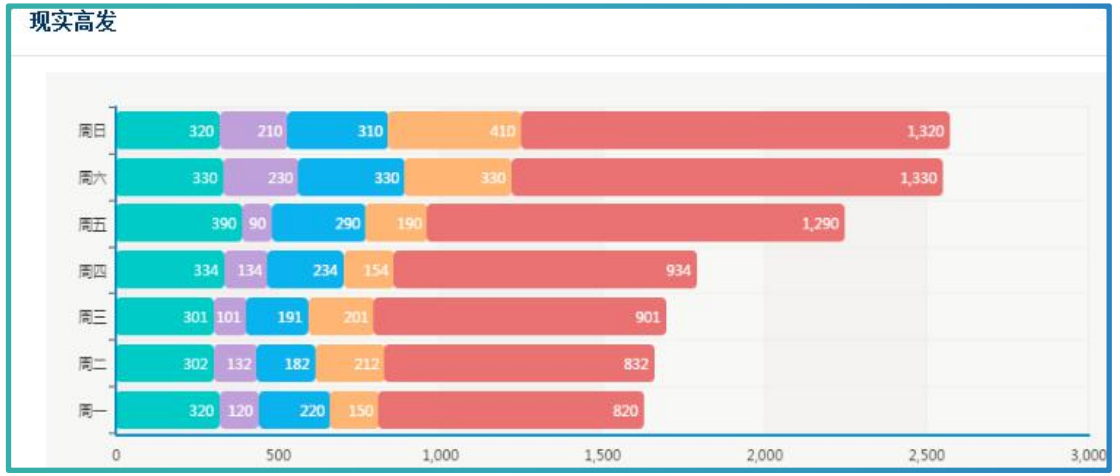

生产商
餐饮业
财政部
毒性
化学品
利税
重金属
烟草
酒精
香烟
过滤嘴
大烟
毒气
大麻
水烟
雪茄
毒品
(关键词)
团伙
案件
嫌犯
犯罪分子
分子
瘾君子
可卡因
鸦片
海洛因
违禁物品
精液
烈性酒
液体
可的松

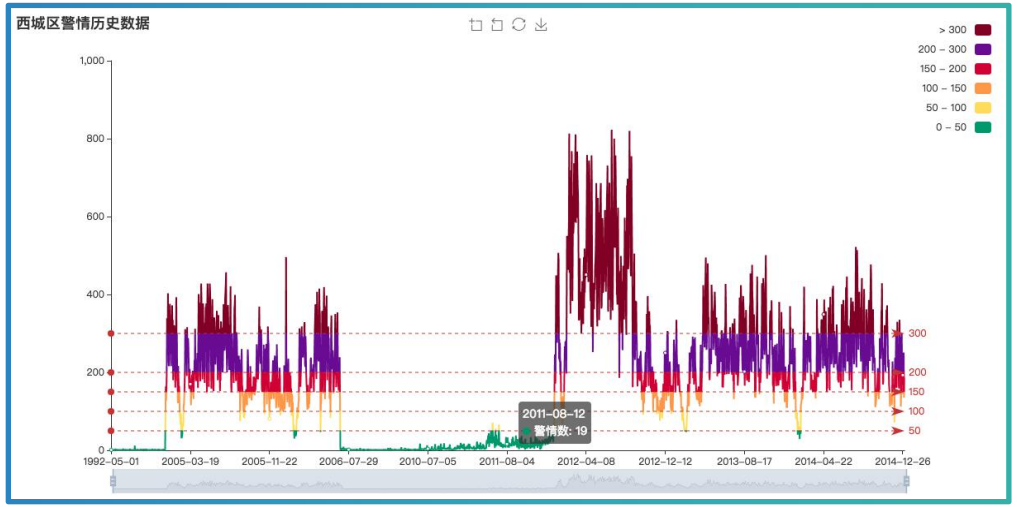
公安语义增强应用

聚类高发

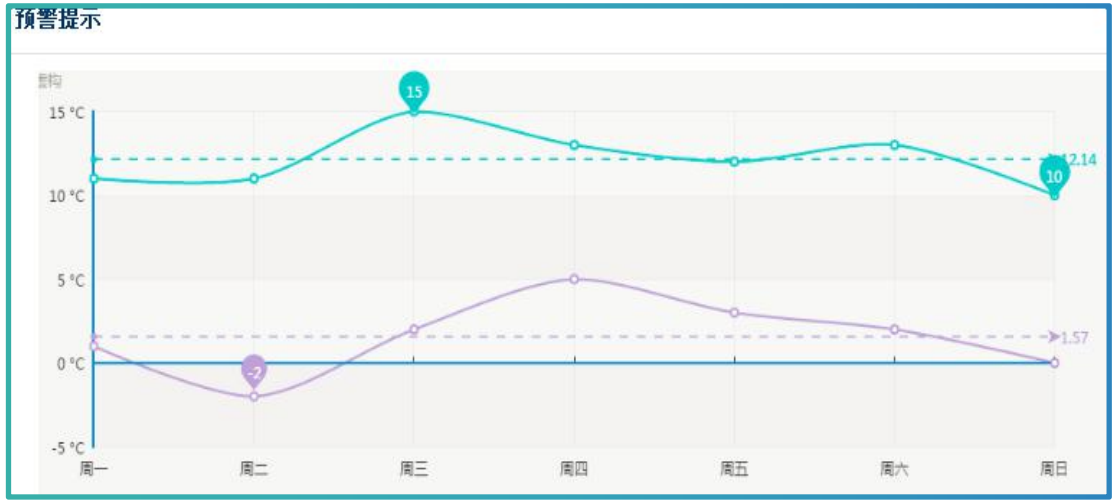
Cluster 0: 天作 使用费 非法占有
Cluster 1: 伪造 交通事故 赔偿金 保险公司
Cluster 2: 博亿 房屋买卖 虚构
Cluster 3: 鲁迅文学奖 作家协会 诗词
Cluster 4: 骗走 人民币 平房 公墓 灵泉 塔位 灵塔 地葬位
Cluster 5: 押金 天纺 柯东 冒用 丰体
Cluster 6: 为名 培训 假借 团伙 考证 帮助

现实高发





预警提示



智能情报：全球金融情报

已有相关研究1：
全球互联网金融信息挖掘服务

1、常规报告

常规信息挖掘报告
（日报、周报、月报）
专题专项报告

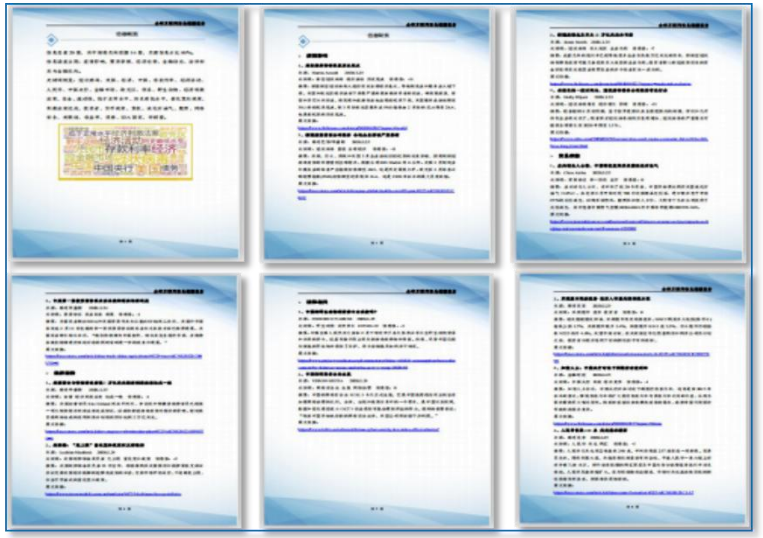
2、订制报告

背景调查报告
信息搜集报告
来源调查报告

3、风控报告

不良资产调查
投资风险分析
上市风险分析

常规报告



订制报告

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

军事知识图谱：从浅层信息挖掘深度情报

军事知识图谱：从浅层信息挖掘深度情报


观察者

首页 风闻 观察员 国际 军事 财经 产经 科技 汽车 视频

美日澳联合军演，我电侦船“抢风头”

王世纯
高武德势力一般慢慢相惜

分享到: 30 157

2019-07-09 14:57:36 字号: A- A A+ 来源: 观察者网

关键字: 815A舰赴澳大利亚军演

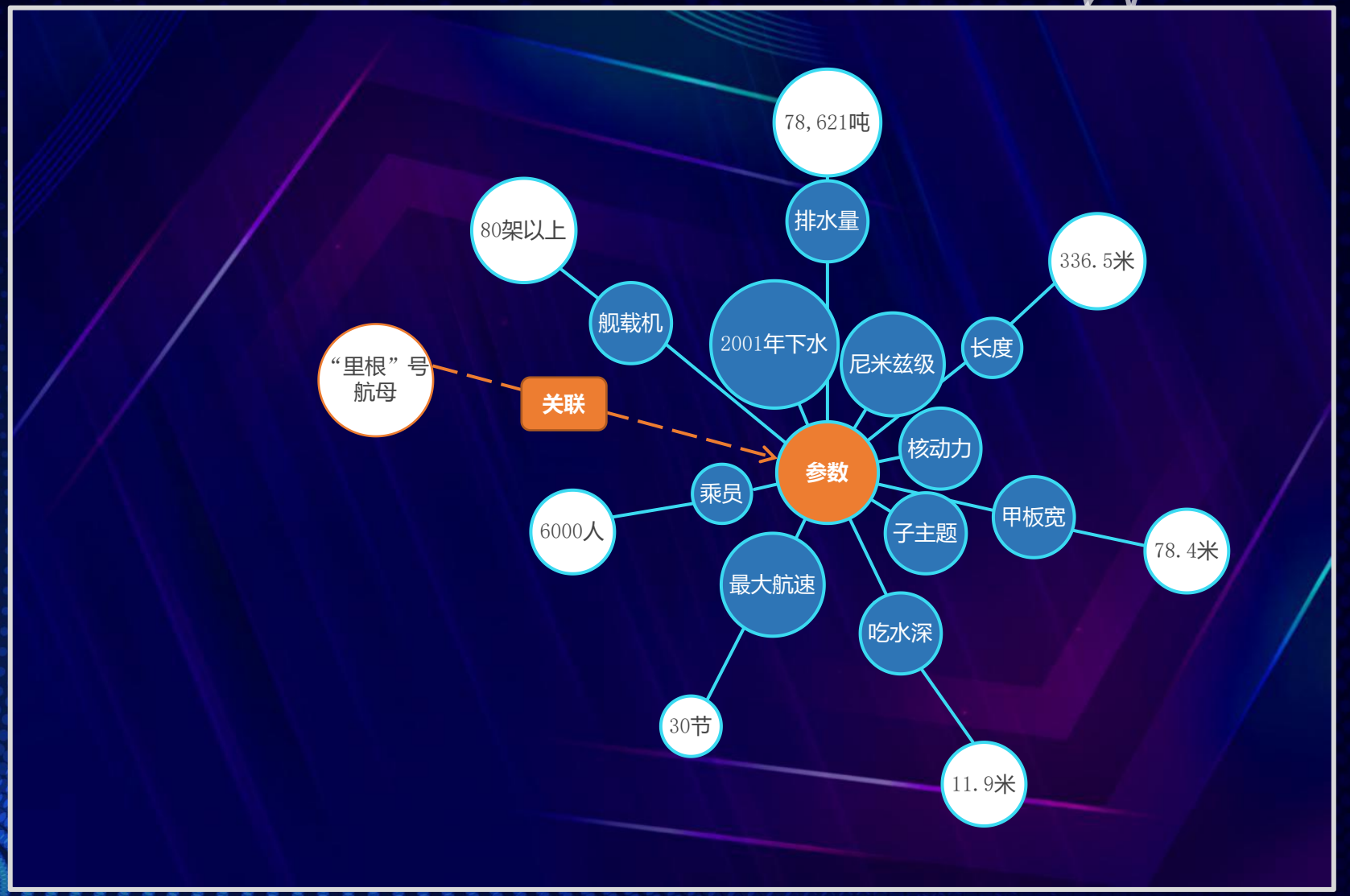
【文/观察者网 王世纯】

美日澳三国在澳大利亚举行声势浩大的海上军演，但夺去军演“风头”的，则是我军一艘815A型电子侦察船。

据澳大利亚广播公司（ABC）7日报道，美澳日“护身军刀-2019”联合军演当天在昆士兰州外海拉开序幕，此次军演声势浩大，美国海军出动印太司令部主力“里根”号航母赴昆士兰参演，而日本海上自卫队则排出“日向”级直升机航母“伊势”号和“大隅”级船坞运输舰“国东”号也赴澳参演。

美国第七舰队“里根”号航母打击大队已于7月5日抵达昆士兰 图源：美国海军

除了航母与两栖攻击舰以外，美澳还出动战机，装甲部队和两栖登陆部队进行了联演，参演人员达到2.5万人，是近年来比较大的联合军演。值得一提的是，日本自卫队去年才成立的“水陆机动团”也是首次参加军演。



军事知识图谱：从浅层信息挖掘深度情报

观察者 首页 风闻 观察员 国际 军事 财经 产经 科技 汽车 视频

美日澳联合军演，我电侦船“抢风头”

王世纯
高武德势力一般慢慢相惜

分享到: 30 157

2019-07-09 14:57:36 字号: A- A A+ 来源: 观察者网

关键字: 815A舰赴澳大利亚军演

【文/观察者网 王世纯】

美日澳三国在澳大利亚举行声势浩大的海上军演，但夺去军演“风头”的，则是我军一艘815A型电子侦察船。

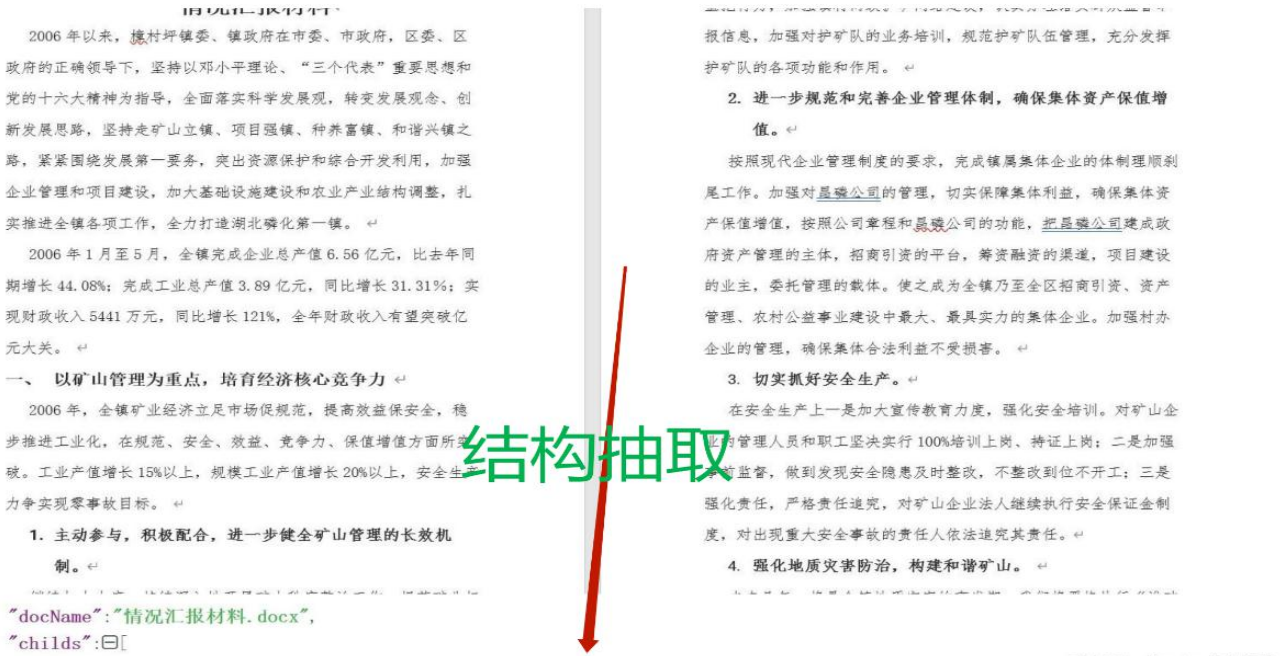
据澳大利亚广播公司（ABC）7日报道，美澳日“护身军刀-2019”联合军演当天在昆士兰州外海拉开序幕，此次军演声势浩大，美国海军出动印太司令部主力“里根”号航母赴昆士兰参演，而日本海上自卫队则排出“日向”级直升机航母“伊势”号和“大隅”级船坞运输舰“国东”号也赴澳参演。

美国第七舰队“里根”号航母打击大队已于7月5日抵达昆士兰 图源：美国海军

除了航母与两栖攻击舰以外，美澳还出动战机，装甲部队和两栖登陆部队进行了联演，参演人员达到2.5万人，是近年来比较大的联合军演。值得一提的是，日本自卫队去年才成立的“水陆机动团”也是首次参加军演。

军事知识图谱：从浅层信息挖掘深度情报

智能机器阅读：篇章结构理解



2006 年以来，樟村坪镇委、镇政府在市委、市政府，区委、区政府的正确领导下，坚持以邓小平理论、“三个代表”重要思想和党的十六大精神为指导，全面落实科学发展观，转变发展观念、创新发展思路，坚持走矿山立镇、项目强镇、种养富镇、和谐兴镇之路，紧紧围绕发展第一要务，突出资源保护和综合开发利用，加强企业管理和项目建设，加大基础设施建设和农业产业结构调整，扎实推进全镇各项工作，全力打造湖北磷化第一镇。

2006 年 1 月至 5 月，全镇完成企业总产值 6.56 亿元，比去年同期增长 44.08%；完成工业总产值 3.89 亿元，同比增长 31.31%；实现财政收入 5441 万元，同比增长 121%，全年财政收入有望突破亿元大关。

一、 以矿山管理为重点，培育经济核心竞争力

2006 年，全镇矿业经济立足市场促规范，提高效益保安全，稳步推进工业化，在规范、安全、效益、竞争力、保值增值方面所取得突破。工业产值增长 15%以上，规模工业产值增长 20%以上，安全生产力争实现零事故目标。

1. 主动参与，积极配合，进一步健全矿山管理的长效机制。

报信息，加强对护矿队的业务培训，规范护矿队伍管理，充分发挥护矿队的各项功能和作用。

2. 进一步规范和完善企业管理体制，确保集体资产保值增值。

按照现代企业管理制度的要求，完成镇属集体企业的体制理顺剥离尾工作。加强对昌磷公司的管理，切实保障集体利益，确保集体资产保值增值，按照公司章程和昌磷公司的功能，把昌磷公司建成政府资产管理的主体，招商引资的平台，筹资融资的渠道，项目建设的业主，委托管理的载体。使之成为全镇乃至全区招商引资、资产管理、农村公益事业建设中最大、最具实力的集体企业。加强村办企业的管理，确保集体合法利益不受损害。

3. 切实抓好安全生产。

在安全生产上一是加大宣传教育力度，强化安全培训。对矿山企业的管理人员和职工坚决实行 100%培训上岗、持证上岗；二是加强安全监督，做到发现安全隐患及时整改，不整改到位不开工；三是强化责任，严格责任追究，对矿山企业法人继续执行安全保证金制度，对出现重大安全事故的责任人依法追究其责任。

4. 强化地质灾害防治，构建和谐矿山。

结构抽取


```
"docName":"情况汇报材料.docx",
"childs":[
    {
        "Text":"中共樟村坪镇委樟村坪镇人民政府工作情况汇报材料\r",
        "Style":Object{...},
        "level":0,
        "childs":[
            Object{...},
            Object{...},
            {
                "Text":"以矿山管理为重点，培育经济核心竞争力  \r",
                "Style":Object{...},
                "level":1,
                "childs":[
                    Object{...},
                    {
                        "Text":"主动参与，积极配合，进一步健全矿山管理的长效机制。\r",
                        "Style":Object{...},
                        "level":2,
                        "childs":[
                            {
```


可点击key和value值进行操

智能辅助写作：文书自动核查

工 程 名 称：小学项目总承包工程

采购

一、材料名称、规格、数量及价格（见后附表）

供应

序号	材料名称	规格型号	单位	数量（暂估）	含税综合单价（元/个）	含税总价
1	潜水泵	D				
2	污水泵	2.				
3	污水泵	3.				
4	塑料管	D				
5	排水管	D				
合计						

采购方：（盖章）

[ErrorMsg: 实体前后

datetime="2018-09-19T1

中黑名单]（盖章）

采购方：中航天建设

供应方：固安轴承厂

工程名称 小学项目总

施工地点：固安县

序号	材料名称	规格型号	单位	数量（暂估）	含税综合单价（元/个）	含税总价
1	潜水泵	DN40	台	1	2900	2900
2	污水泵	2.2KW	台	2	850	~~18001700.00~~[ErrorMsg:1000:数量价格金额核算]
3	污水泵	3.0KW	台	2	950	1900
4	塑料管	DN40	米	100	8	800
5	排水管	DN65	米	60	7.5	450
合计					7750	
发票税率：3%						

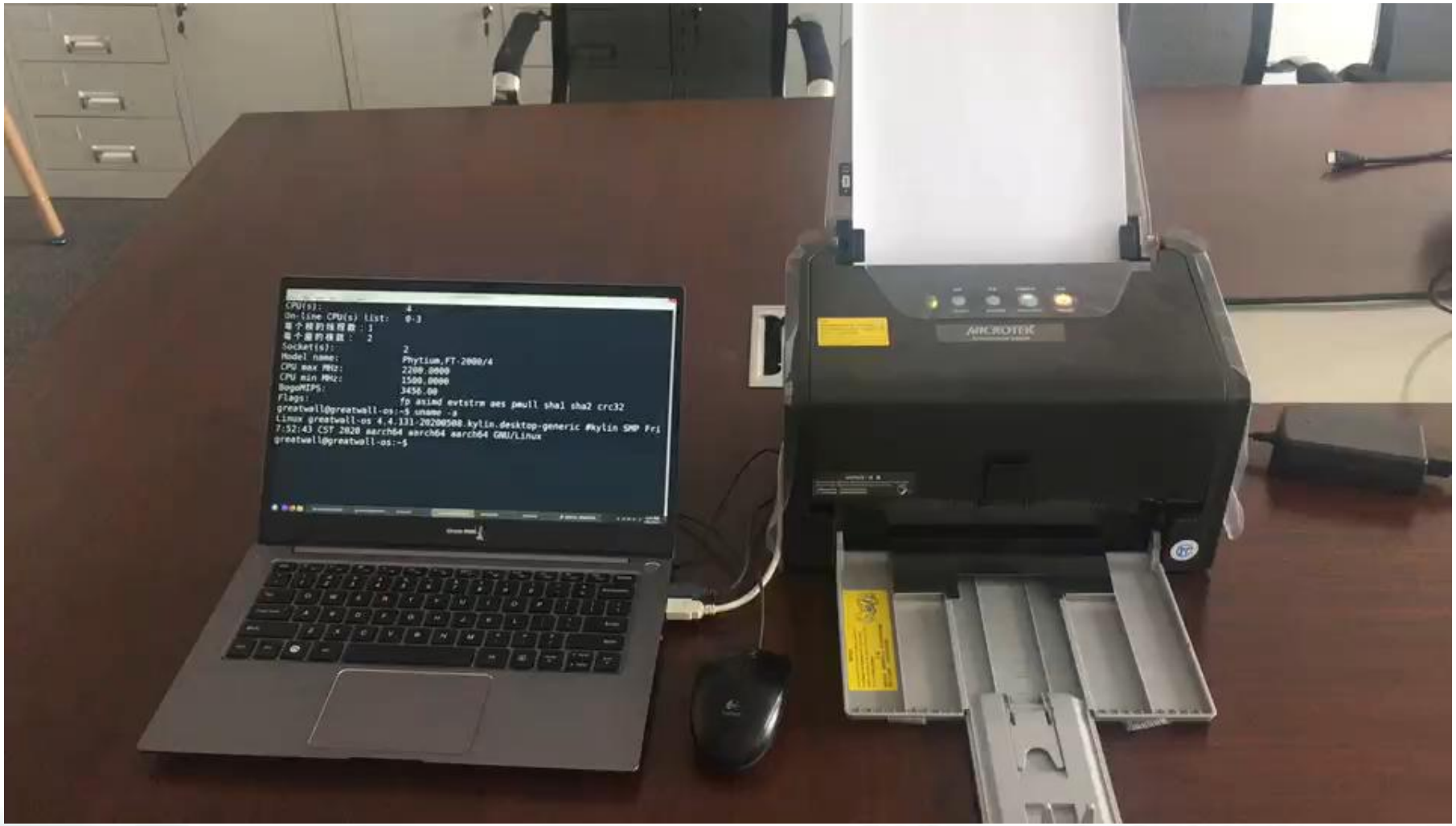
核
*数
总

1、本合同价款为：~~77517750[ErrorMsg:金额大小写不一致]~~7850.00[ErrorMsg:1000:分项总和必须等于总额] 元，（大写）：柒仟柒佰伍拾元整；

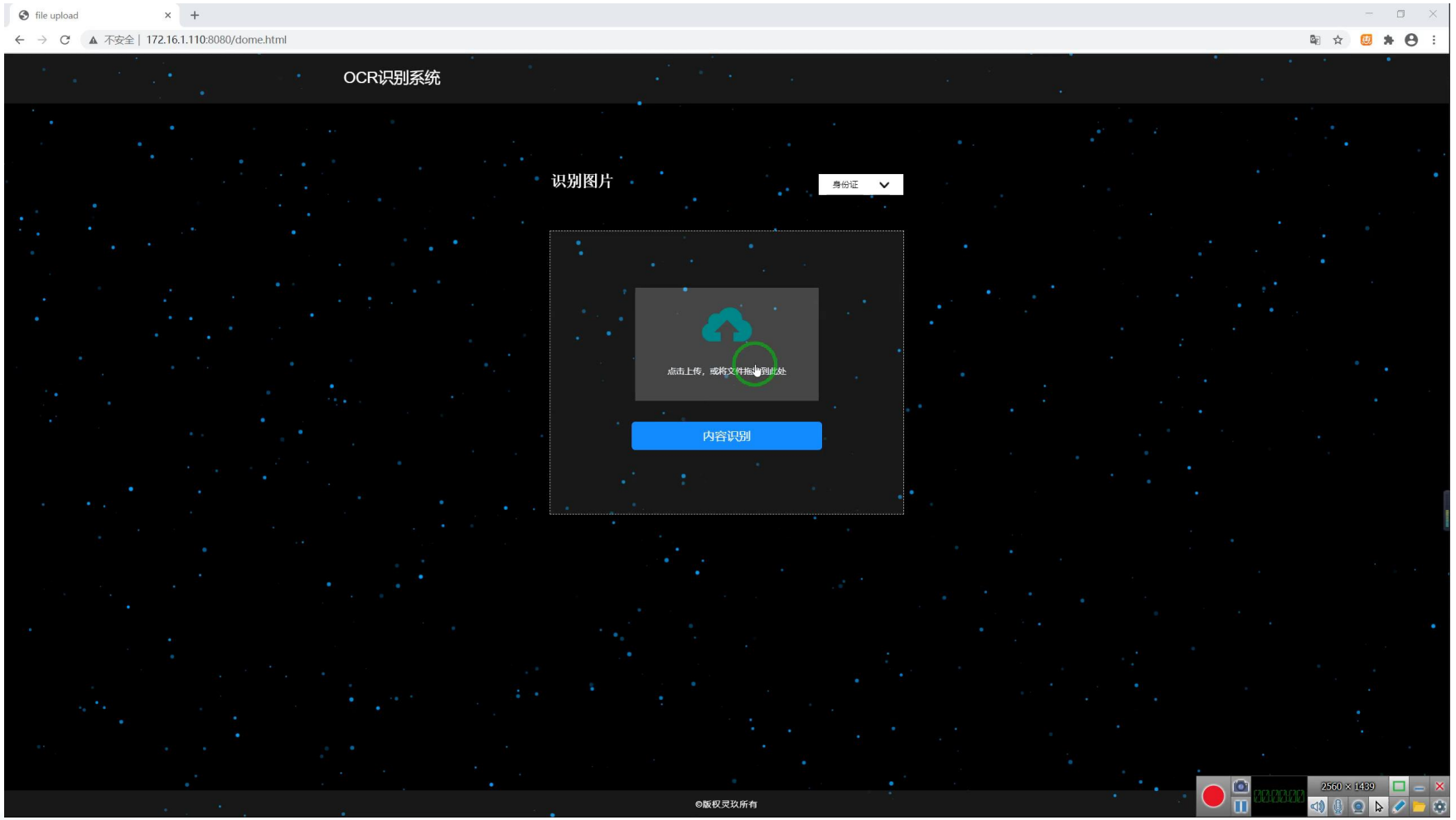
2、合同价款采用 固定单价 方式确定，合同单价包括 材料费、包装费、运输费、仓储费、利润、税金等从材料出厂运至施工现场指定位置的所有费用，结算时不再调整合同单价；

金额核算核查：金额大小写不一致；分项总价之和不等于合计总额。

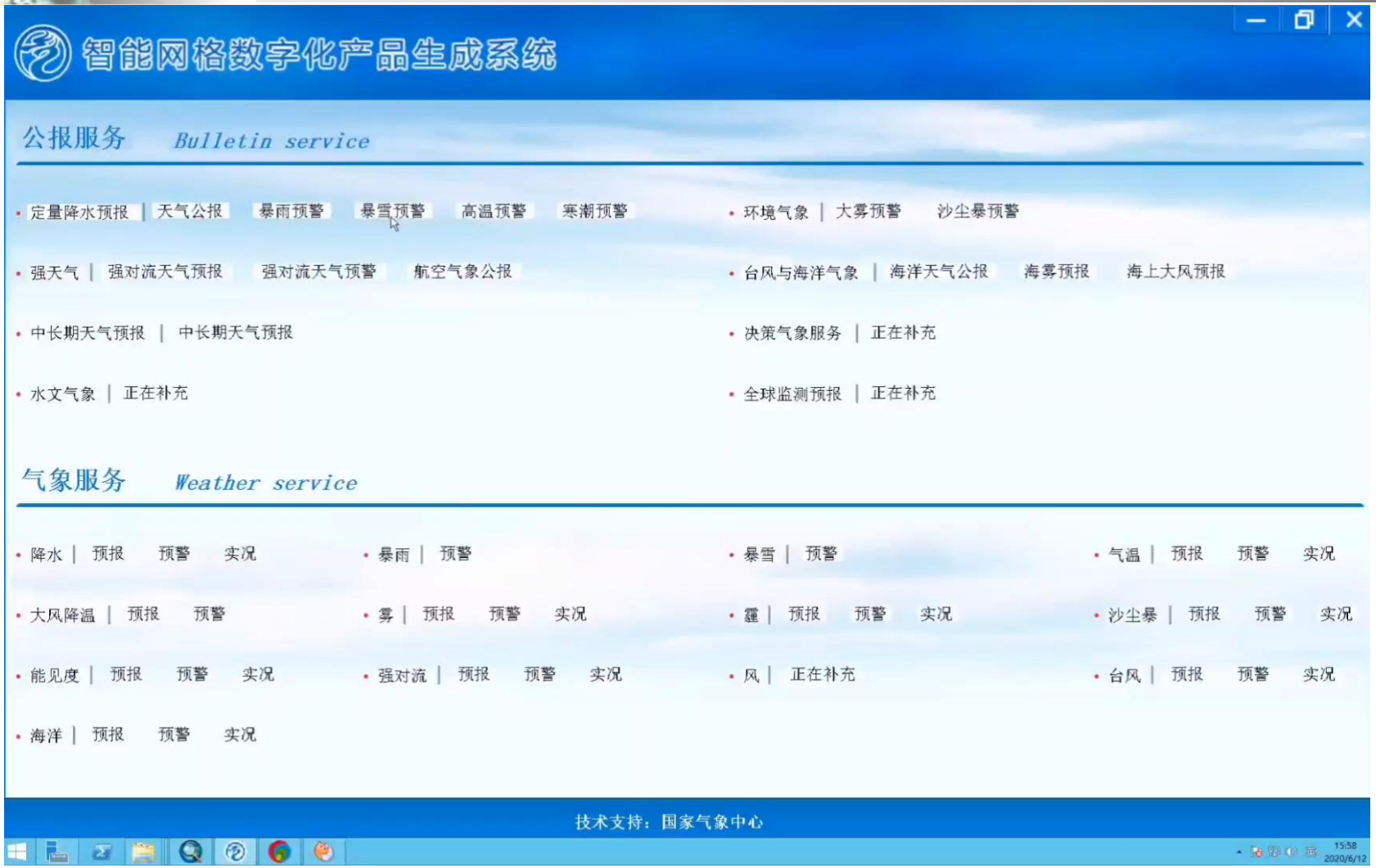
智能辅助写作：批量国产化扫描识别



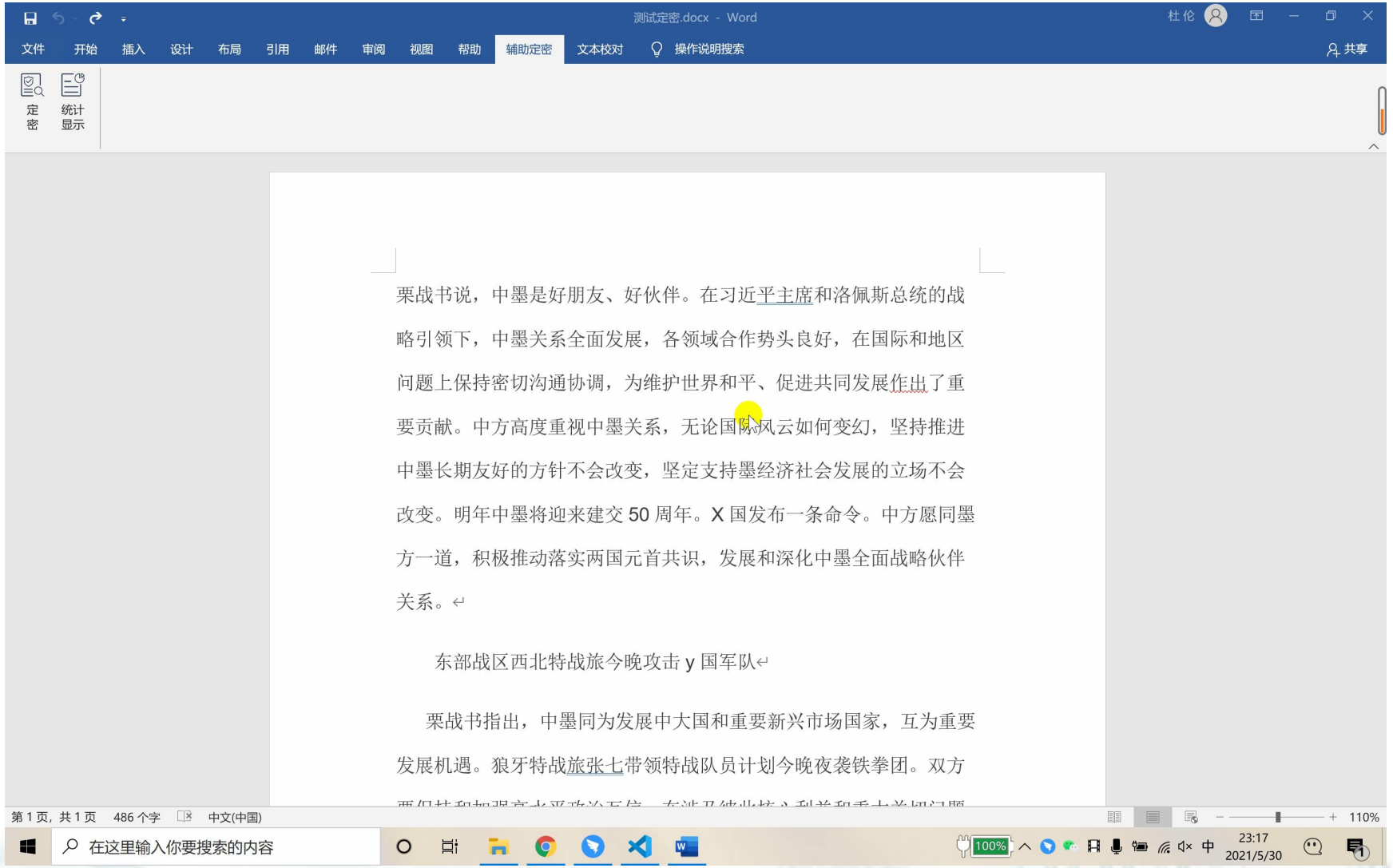
智能辅助写作：图文表识别



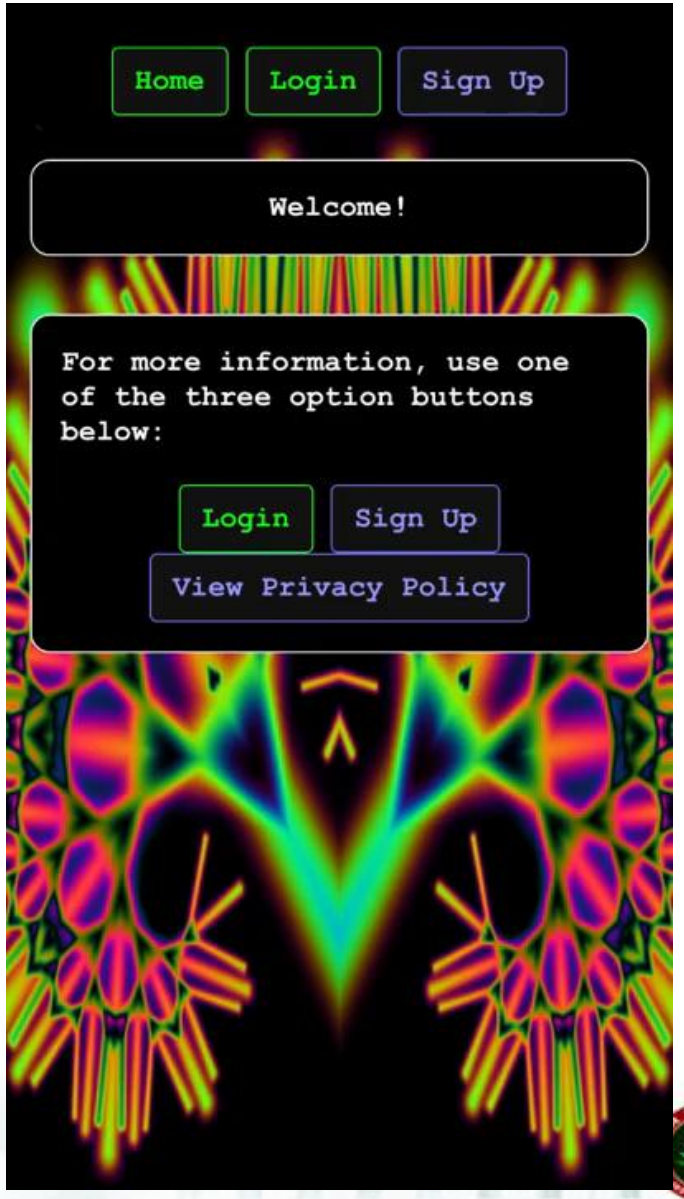
智能文书写作：气象文本生成



智能增值服务：辅助定密



多模态分析


Home
Login
Sign Up
Welcome!
For more information, use one of the three option buttons below:
Login
Sign Up
View Privacy Policy

结语：三个论断

- 人工智能：现代科学皇冠上的明珠；
- 自然语言处理：人工智能皇冠上的明珠
- 机器一思考，人类就发笑；人类一思考，上帝就发笑。

感谢关注聆听！

张华平
Email: kevinzhang@bit.edu.cn
微博：@ICTCLAS张华平博士
实验室官网：
http://www.nlpir.org

大数据千人会