大数据智能之道法术

Big Data Intelligence: Tao, Principle and Tactics

张华平 博士

NLPIR 大数据搜索与挖掘实验室

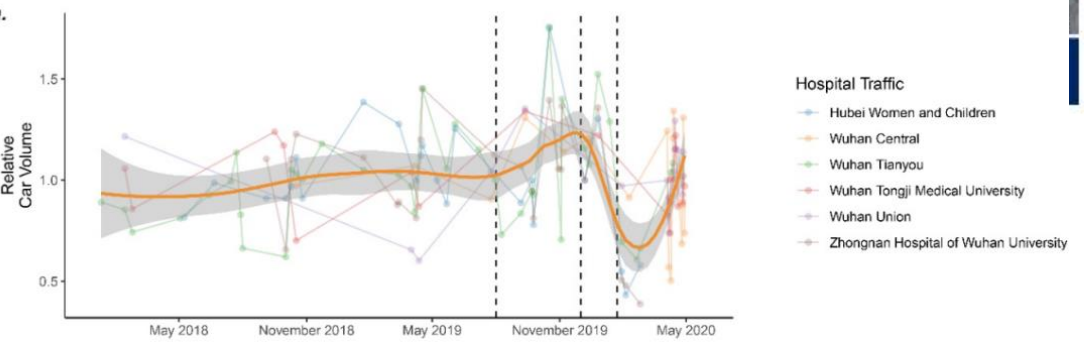
kevinzhang@bit.edu.cn

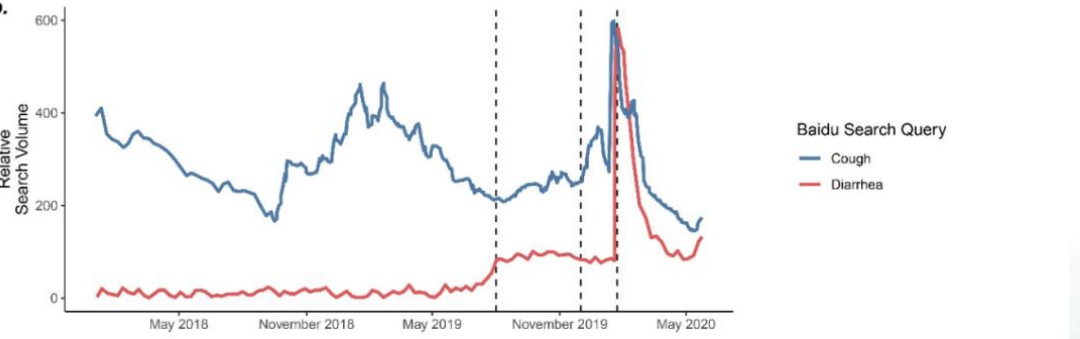
www.nlpir.org

2020.10

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...

大数据智能：一切信息都将成为呈堂证供

三峡大坝固若金汤，可以抵挡万年一遇洪水

http://www.sina.com.cn 2003年06月01日08:54 金羊网-新快报

三峡大坝可抵御百年一遇特大洪水

2006年05月22日09:02 新华网 我要评论(35) 字号：T | T

新华社记者王璐、万后德、余国庆、姚力刚三峡报道

凤凰网资讯 > 大陆 > 南方多省遭暴雨袭击 > 正文

三峡集团董事长：三峡大坝防汛标准为千年一遇

2010年07月19日 21:02 新华网 【大 中 小】【打印】 共有评论4条

遇万年一遇洪峰 三峡大坝仍安全

2011/06/20 00:43 来源：YNET.com 北青网 北京青年报

工程院院士：三峡大坝抵御百年一遇洪水是肯定的

2011-05-26 13:10:40 来源：中国网 有0人参与 手机看新闻 转发到微博(0)

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

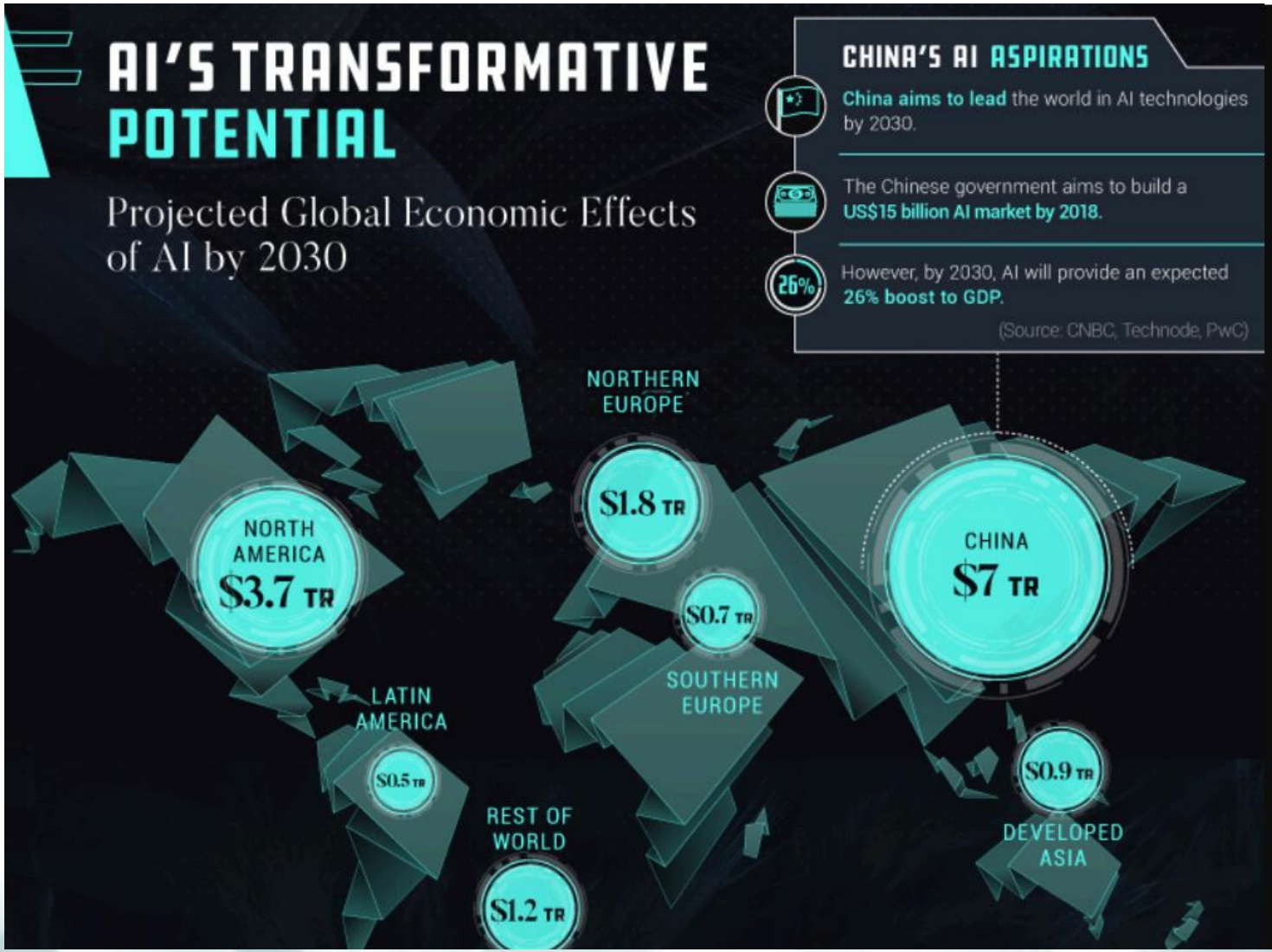
我们的见解：

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

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。


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

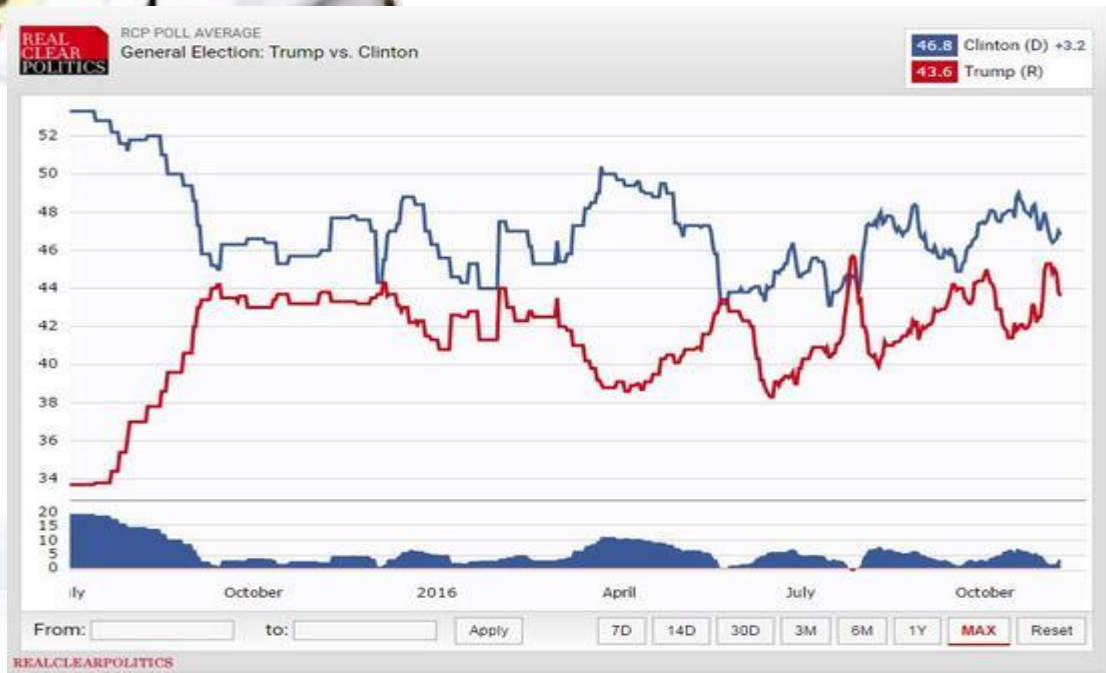
大数据时代

21世纪10
年代

决策方式的革命

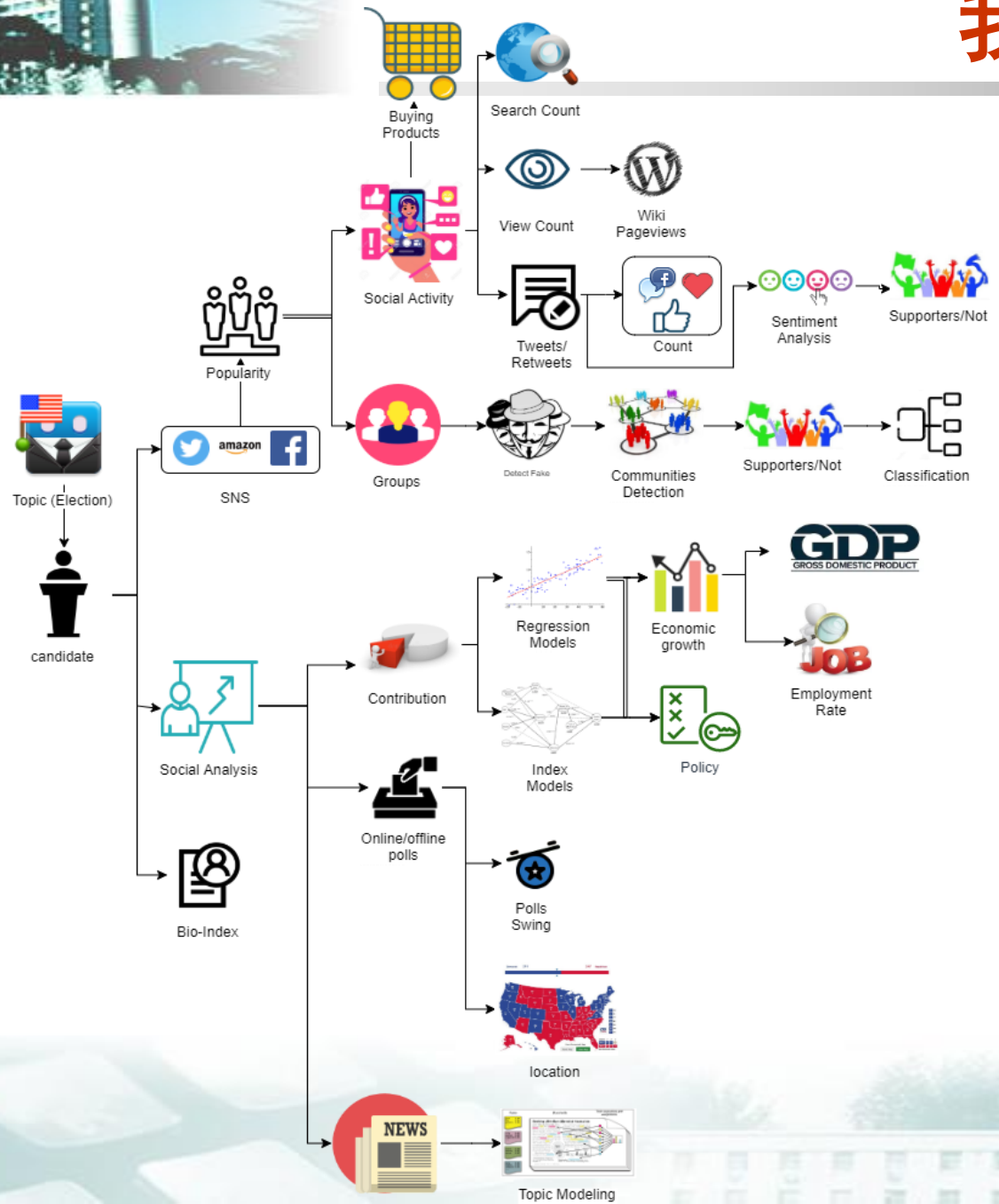
大数据颠覆决策模式

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%；
- 目前在预测美国大选，当前特朗普胜算大于拜登10%。

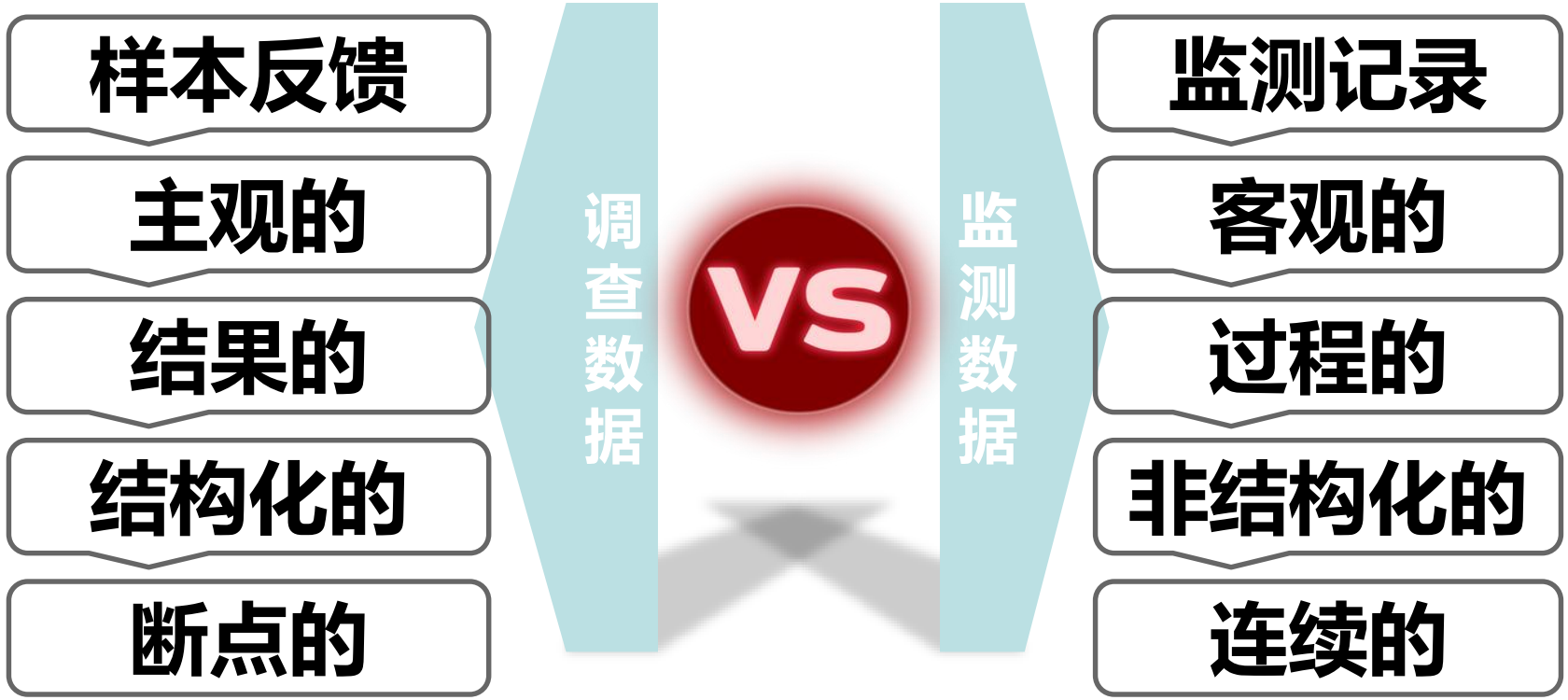
幸存者偏差：你不了解的中国

- 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 &
Batch
Structured
Streaming Data
Zettabytes
Terabytes
省
好
多
快

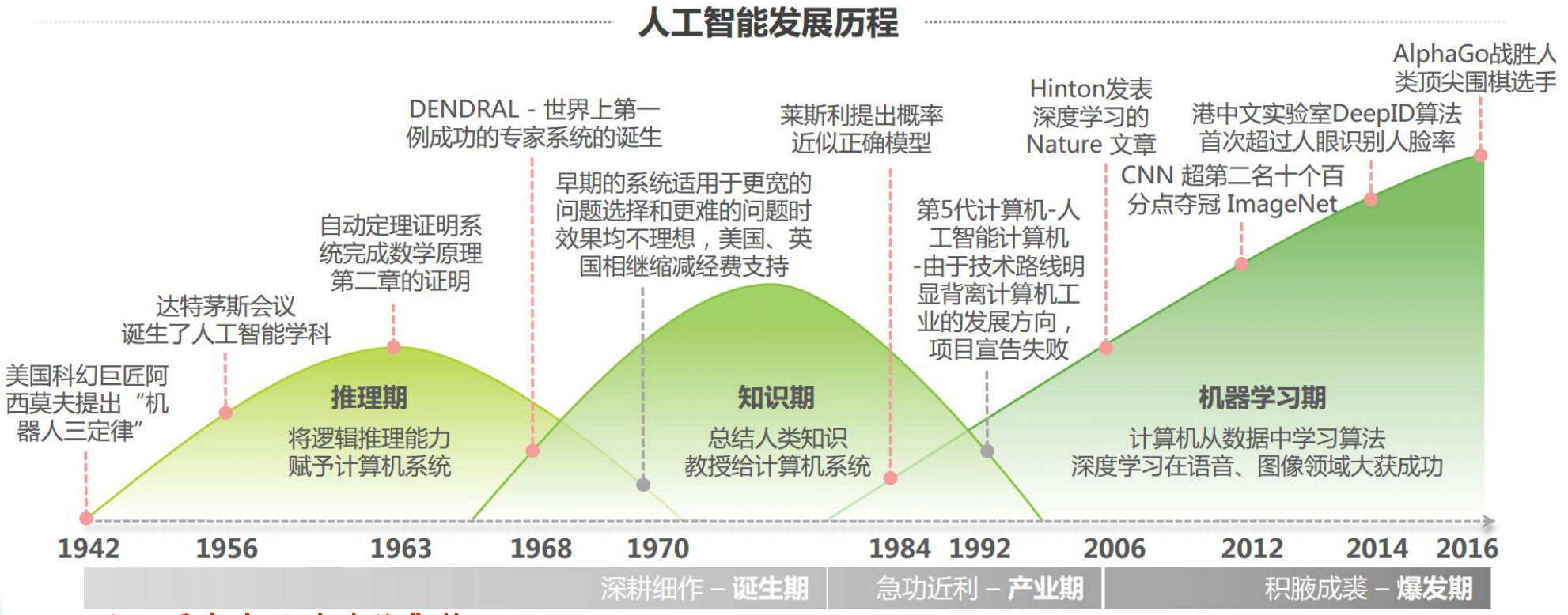
小数据和大数据的区别



认识人工智能

人工智能定义与发展历程

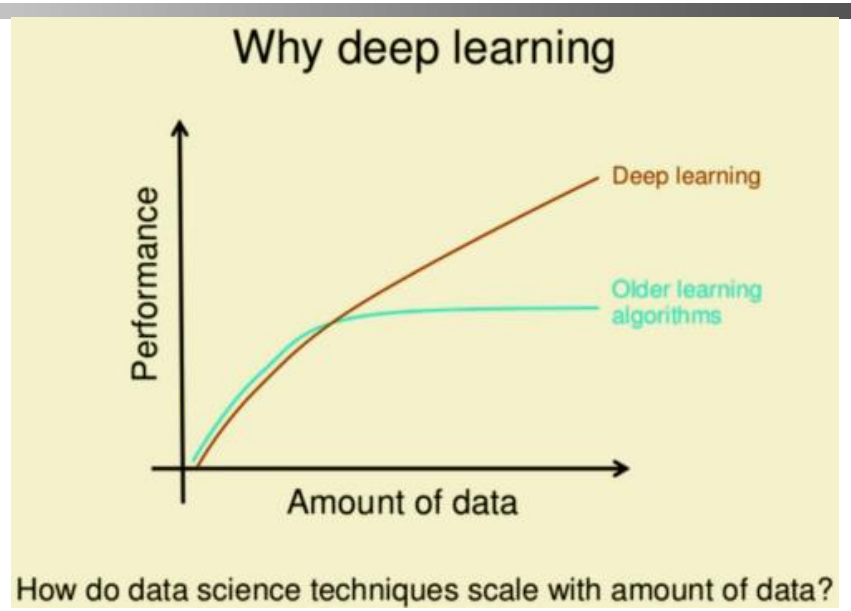
- 1956年，约翰麦卡锡(John McCathy)在达特茅斯会议上首次提出人工智能（Artificial Intelligence: AI）的定义：使一部机器的反应方式像一个人在行动时所依据的智能。

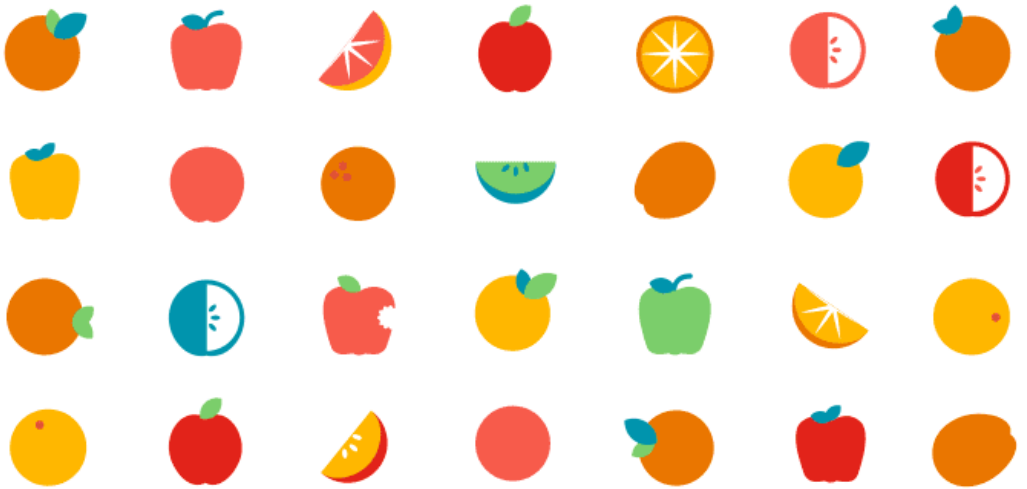


注：图来自于艾瑞收集整理

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

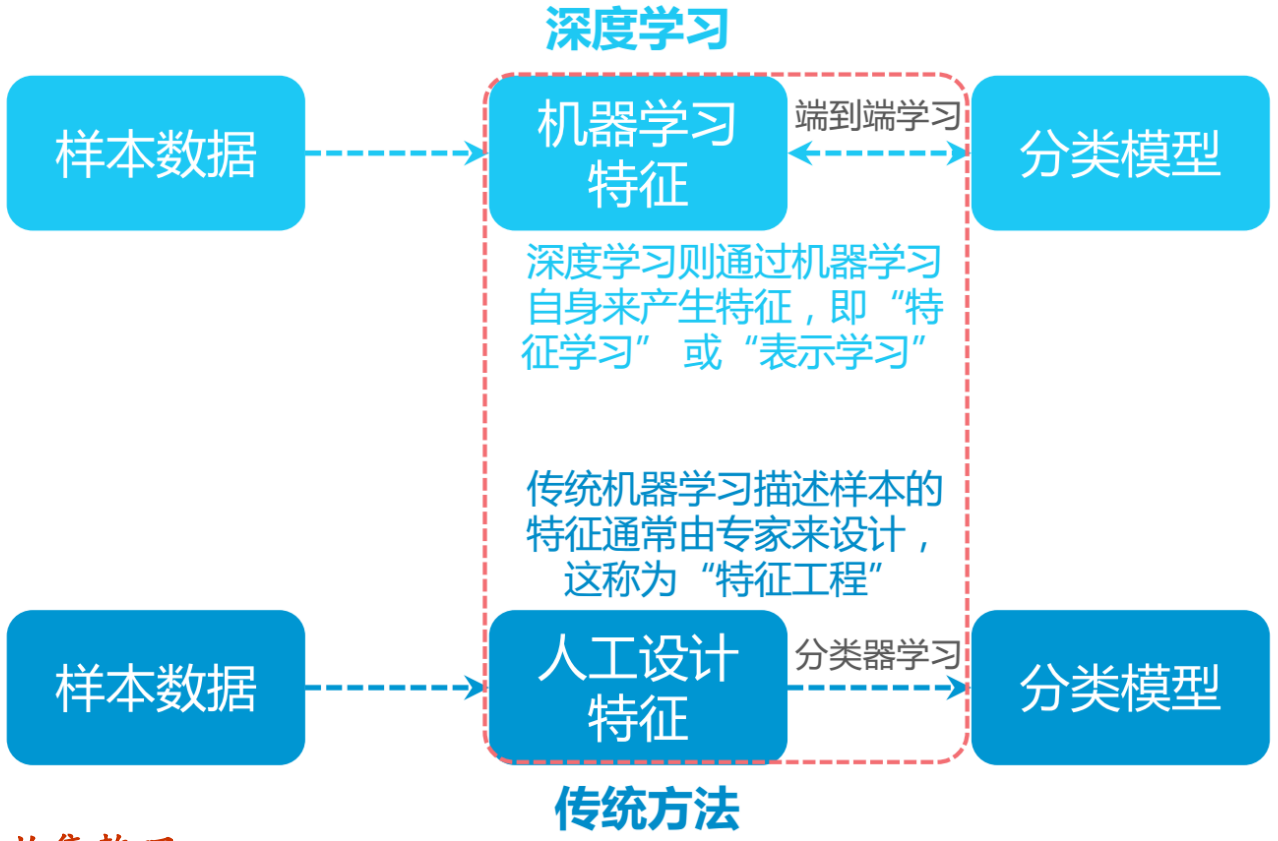
强人工智能
- 感知智能：受限的环境
- 自动驾驶、人脸识别、传感器等

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

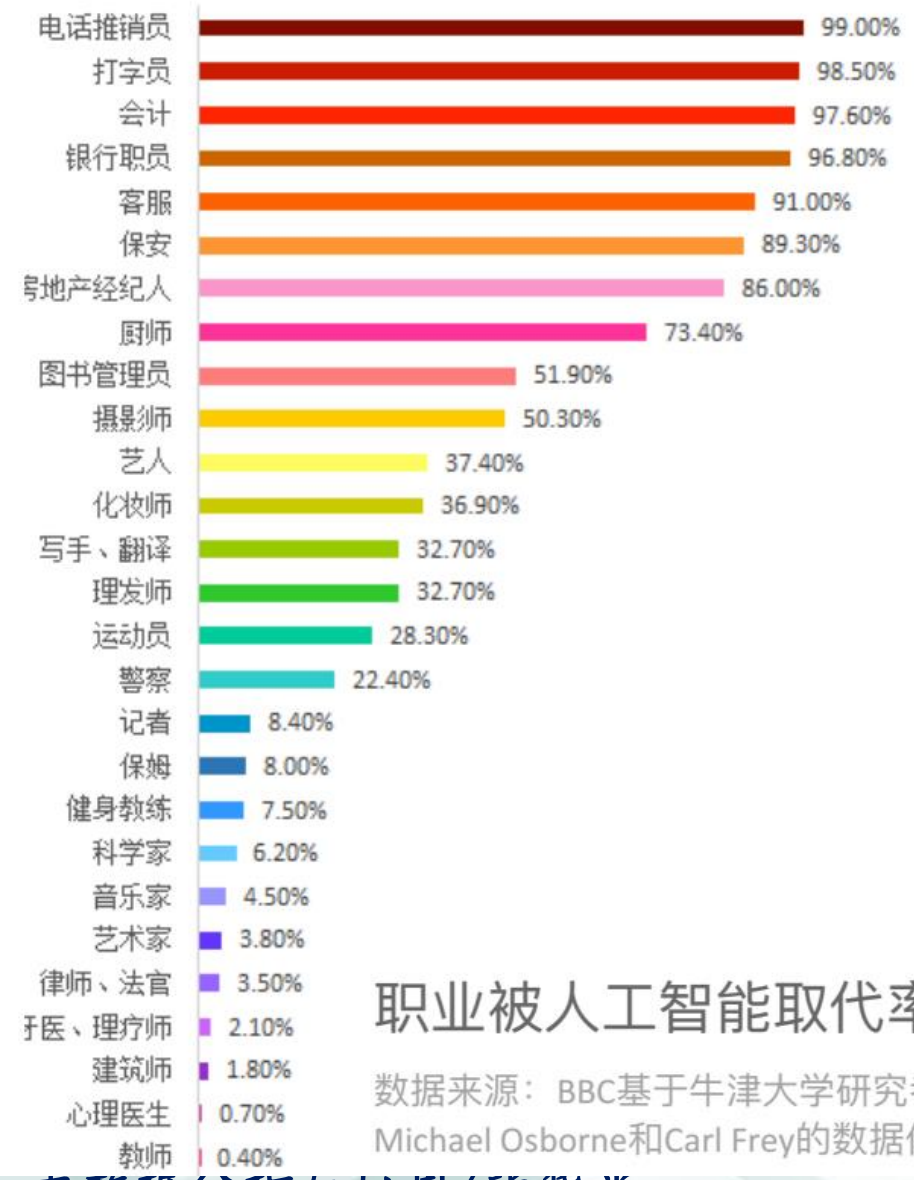
深度学习 < 机器学习 < 人工智能

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

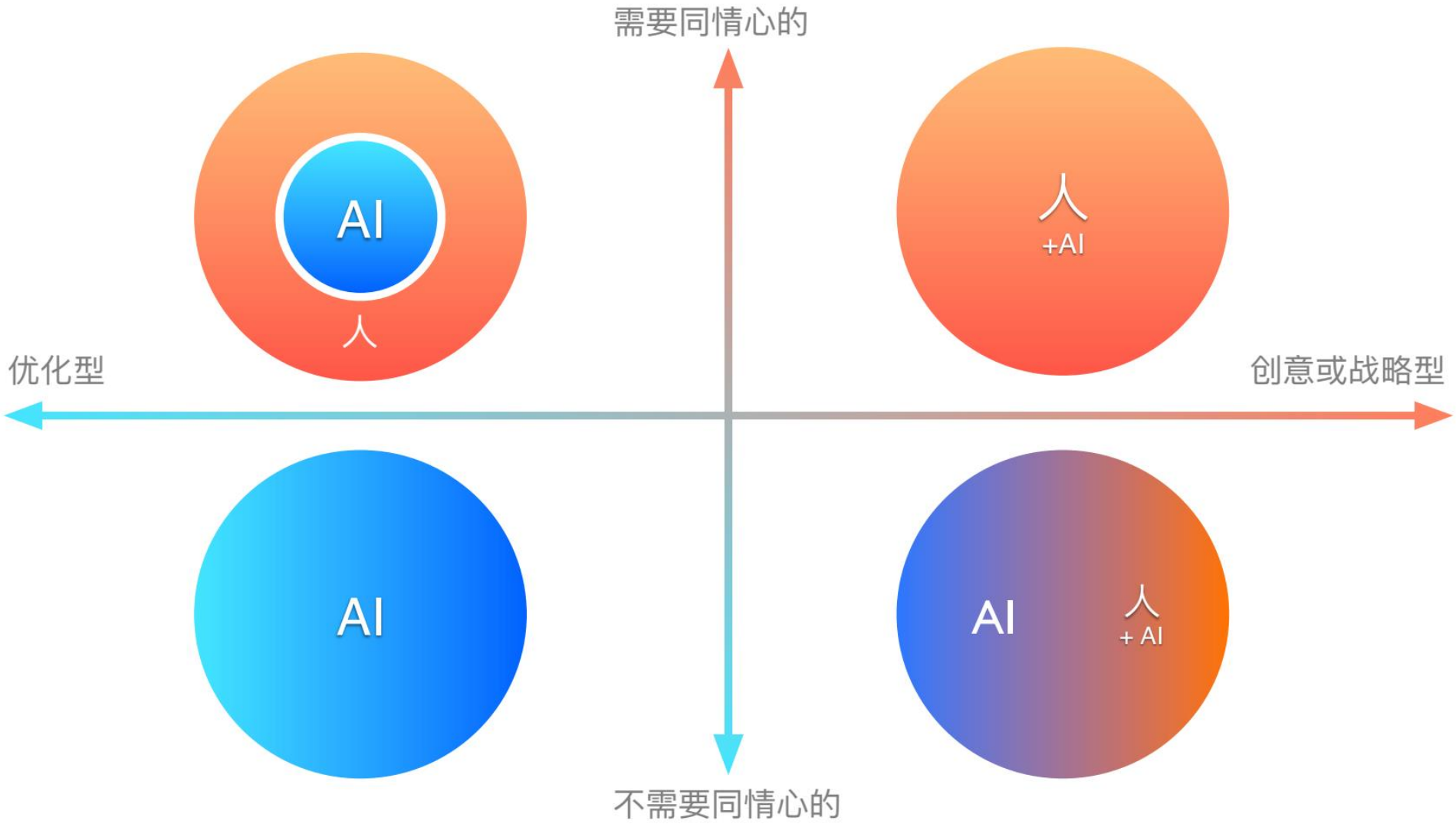

交易员

甚至还有…


法律咨询师


放射科医生

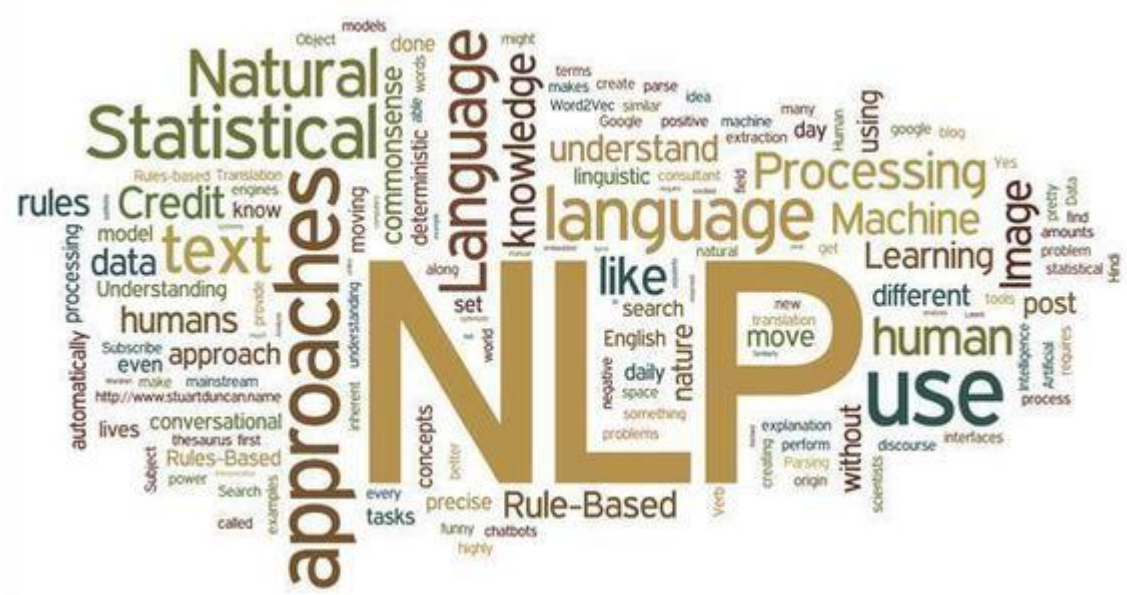
人工智能替代性分析



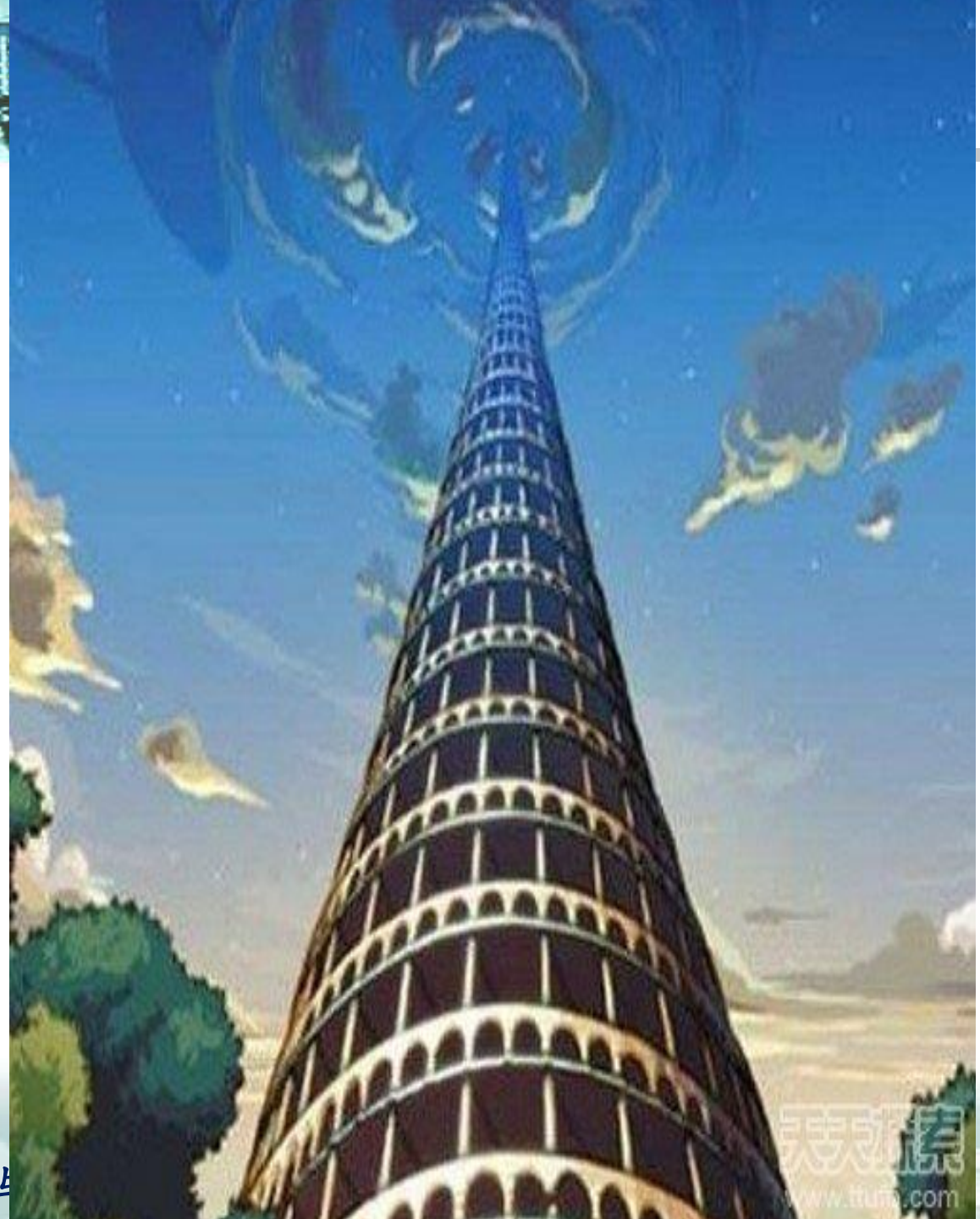
注：图来自于李开复报告

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
Natural
Statistical
approaches
Language
knowledge
understand
Processing
language
Machine
Learning
like
use
human
Rule-Based
rules
Credit
text
data
Understanding
humans
approach
conversational
Rules-Based
commonsense
deterministic
concepts
precise
tasks
without
different
post
image
move
English
search
nature
daily
set
automatically processing
lives
Subscribe
even
mainstream
http://www.stuartduncanname
thesaurus first
Search
called
every
chatbots
funny
highly
discourse
interfaces
process
Artificial
Intelligence
requires
amounts
problem
statistical
pretty
Data
find
Hmd
post
using
Word2Vec
Google
positive
machine
extraction
day
linguistic
consultant
Yes
natural
new
translation
explanation
perform
Parsing
origin
scientists
Verb
problems
space
something
negative
inherent
understanding
moving
along
set
world
Object
models
done
might
terms
makes
create
parse
idea
similar
many
Human
google
blog
Rules-based
Translation
engines
know
model
provide
Subject
power
examples
better
inherent
able
words

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

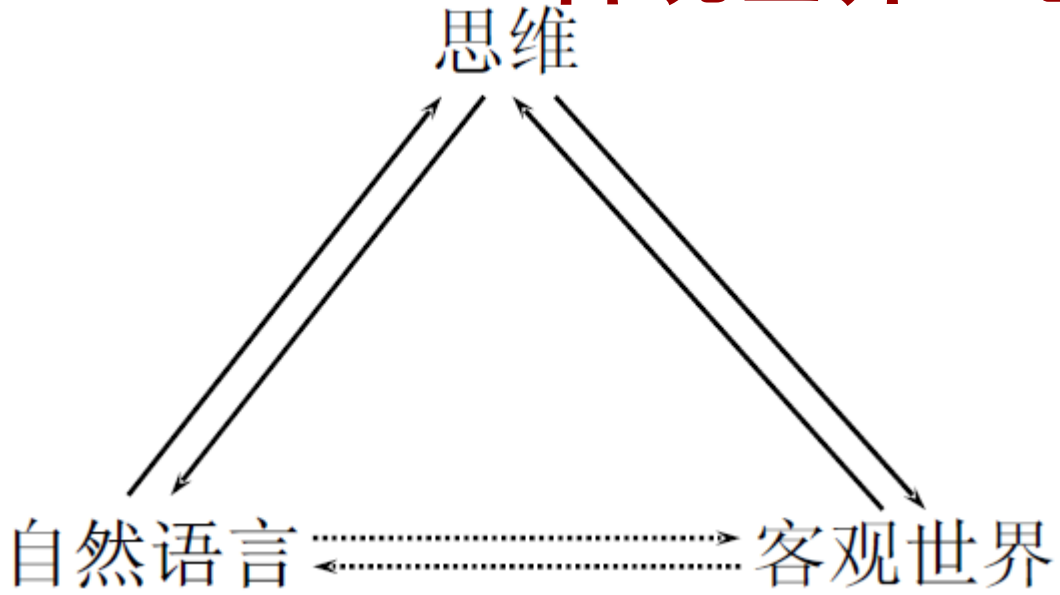
我们是做右撇子鸡(要吃左撇子鸡请去麦当劳)

我们做鸡的时候是很认真的.

我们行……侧.

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

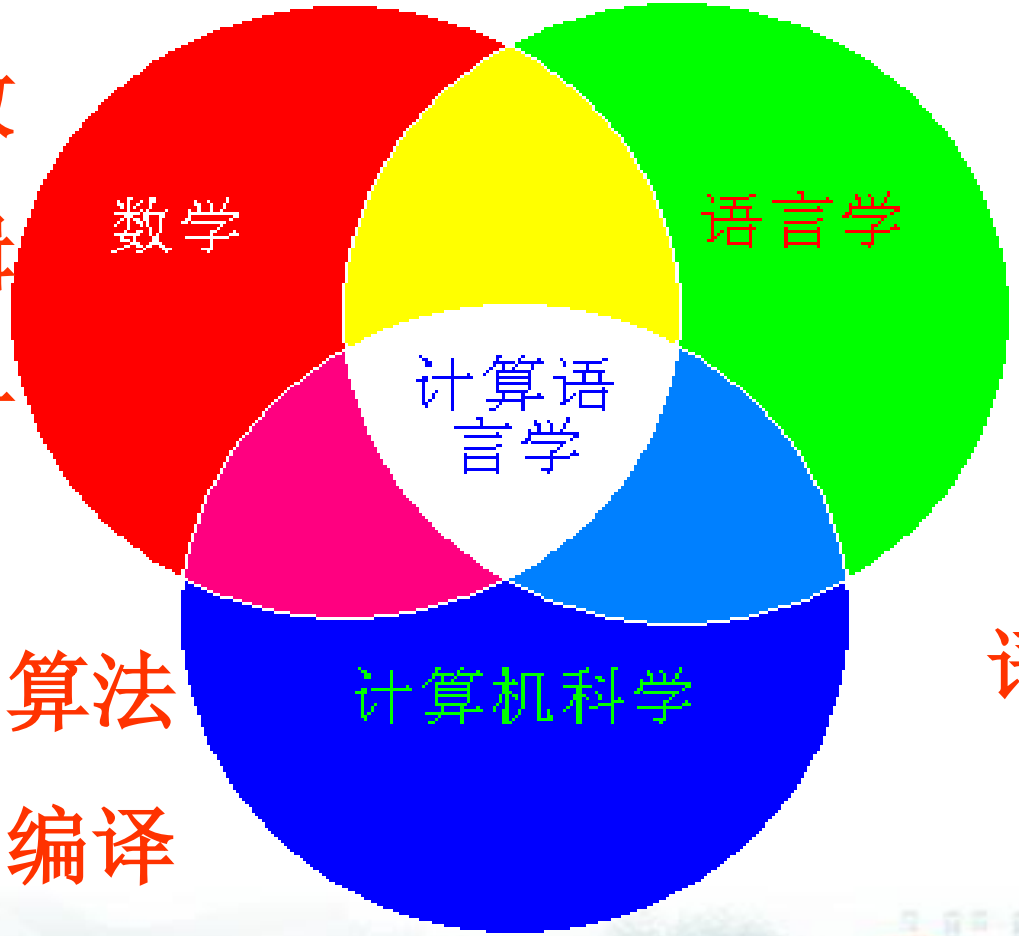
Natural language processing

"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



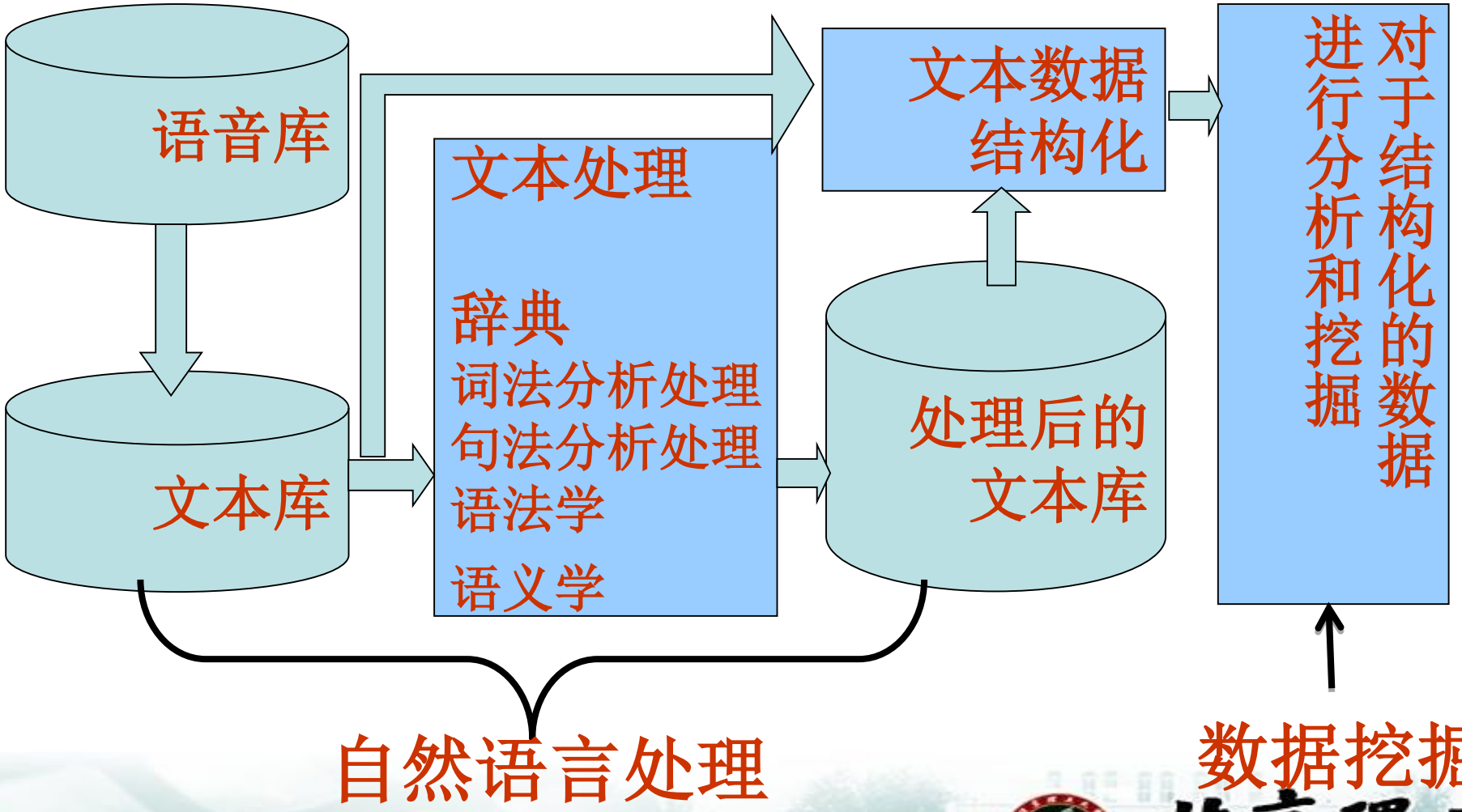
代数
逻辑
统计

词法
句法
语义
词汇
语料库

算法
编译

人工智

文本挖掘的框架



自然语言处理框架体系

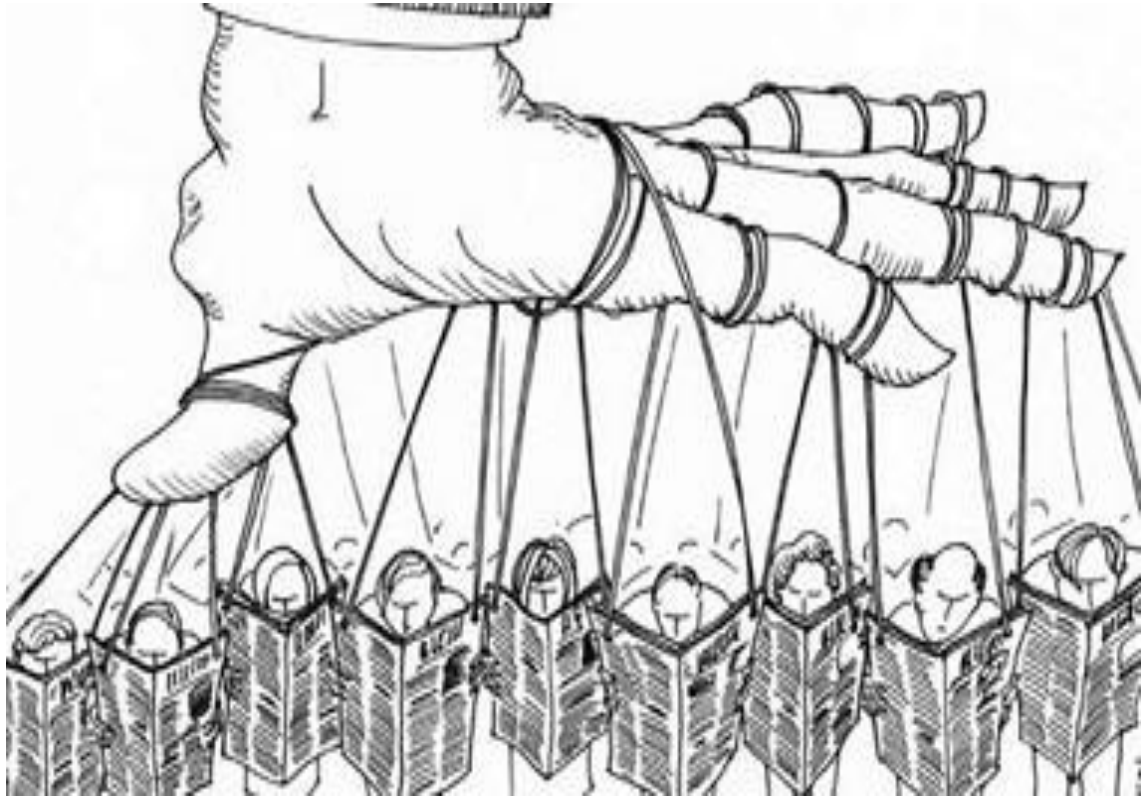
- 基础理论
  - 自动机 形式逻辑 统计机器学习，汉语语言学 形式语法理论
- 语言资源
  - 语料库 词典
- 关键技术
  - 汉字编码 词法分析 句法分析 语义分析 文本生成 语音识别
- 应用系统
  - 人机对话，机器翻译，社交网络分析，分类，聚类，检索，过滤，信息抽取，音字转换

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;

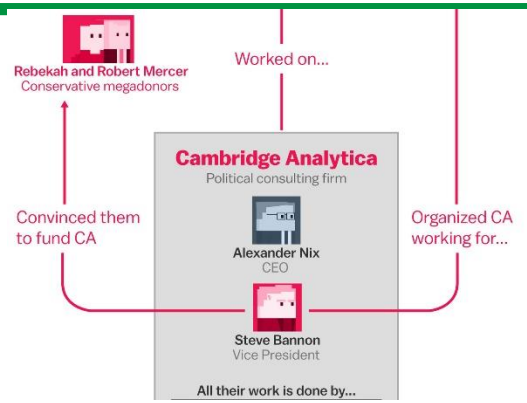
大数据智能应用

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

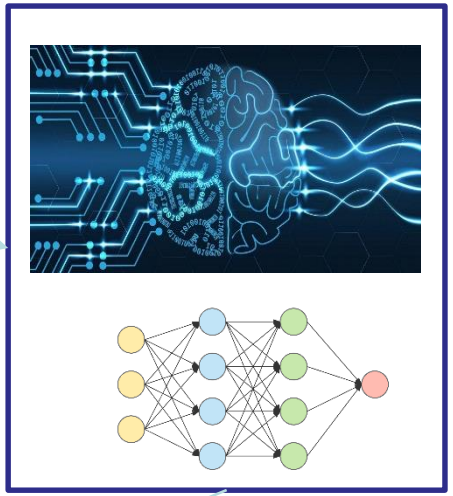
RUSSIAN INSTITUTE FOR STRATEGIC STUDIES

ISIS

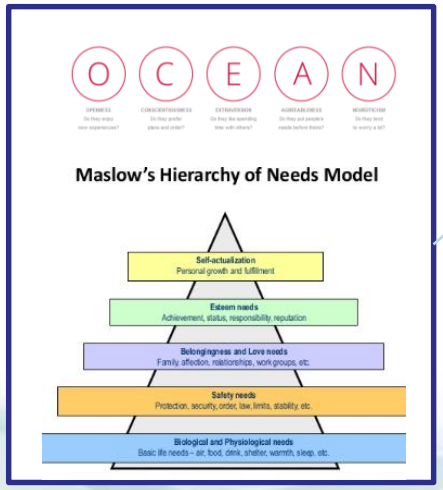
舆论操控者

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

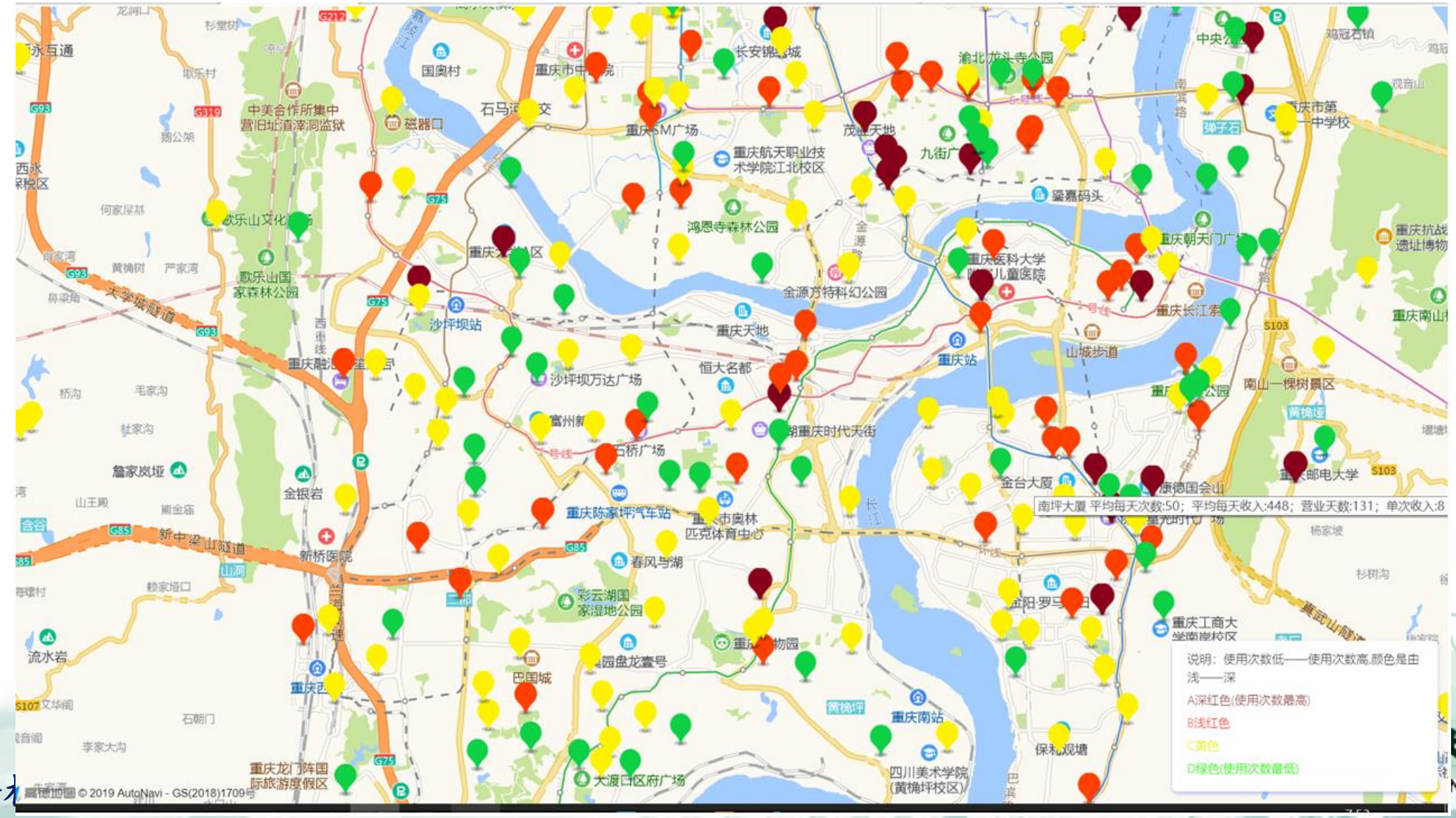
ISIS: 网络极端势力如何吸引恐怖分子

Table 1 Offender behaviour characteristics

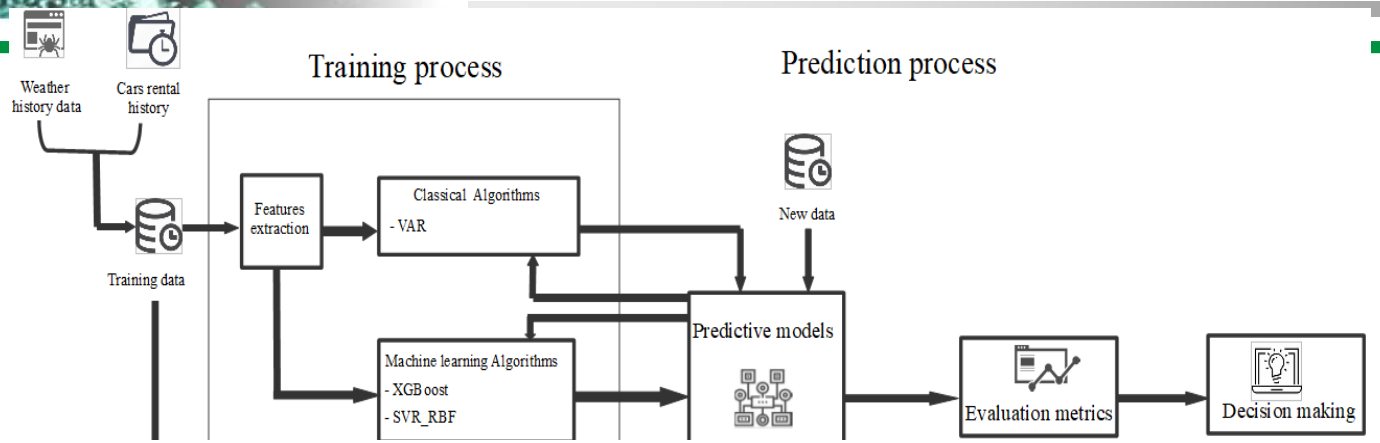
Type	Characteristics	Cases on Twitter	Cases on Facebook	Total No of Cases
Cyber Mobs	Using social media platforms to create a mob mentality and urging others to fight for the Isis goal. This is done through group posts, videos and comments of hate directing groups of Muslim's to fight. Often personified 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

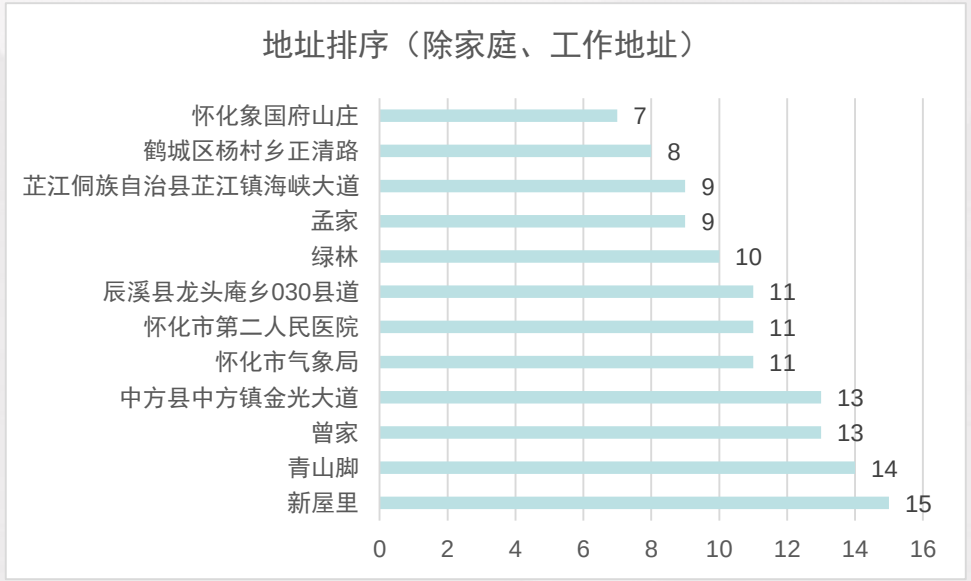
Table 7: evaluation metrics for multivariate time series forecasting using LSTM

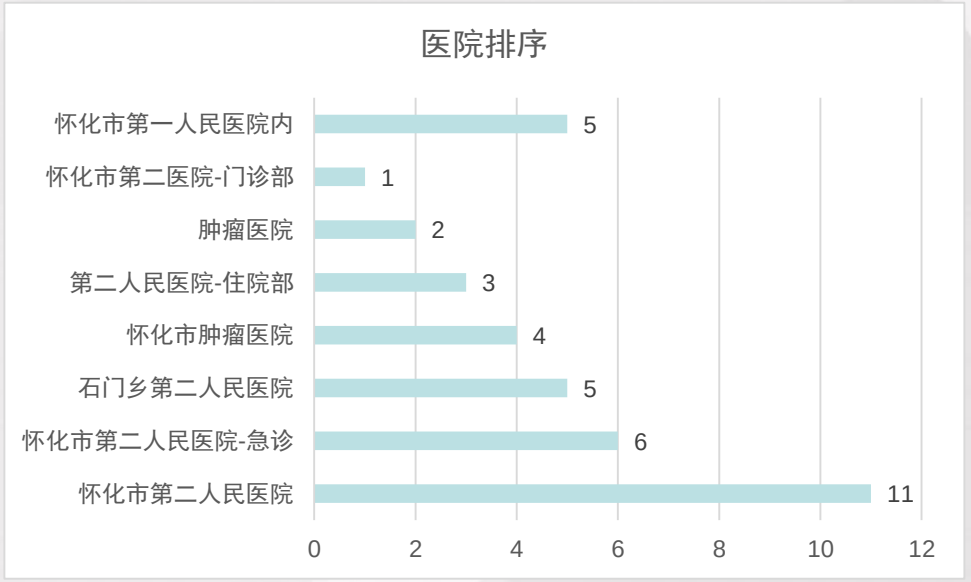
	History + weather + weekend/weekday			History + weather			History + weekend/weekday			History		
	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
Class A (# 16)	0.90898	**2.02258**	**1.42217**	**0.90889**	2.02560	1.42323	1.50311	3.15878	1.77729	1.50419	3.16149	1.77805
Class B (#104)	0.97605	**2.11016**	**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

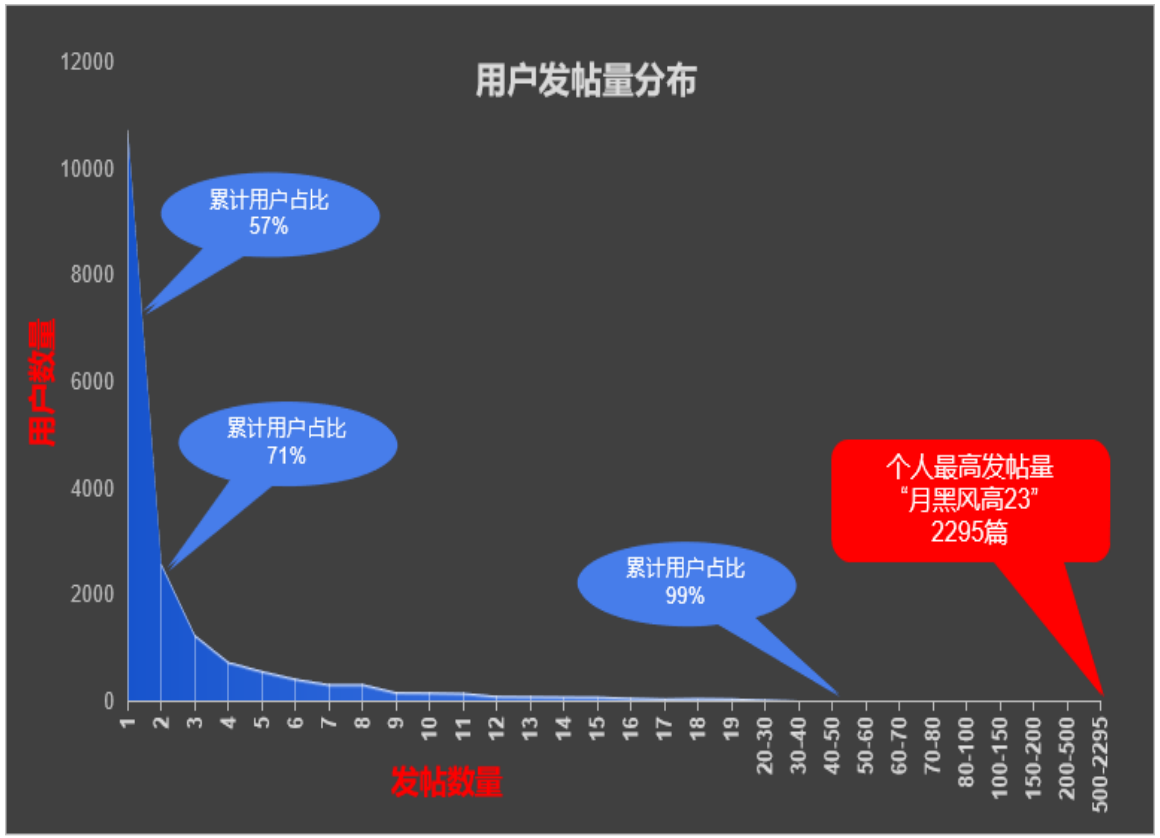
工作日每天各时段开车时长=工作日总开车时长/工作日总天数

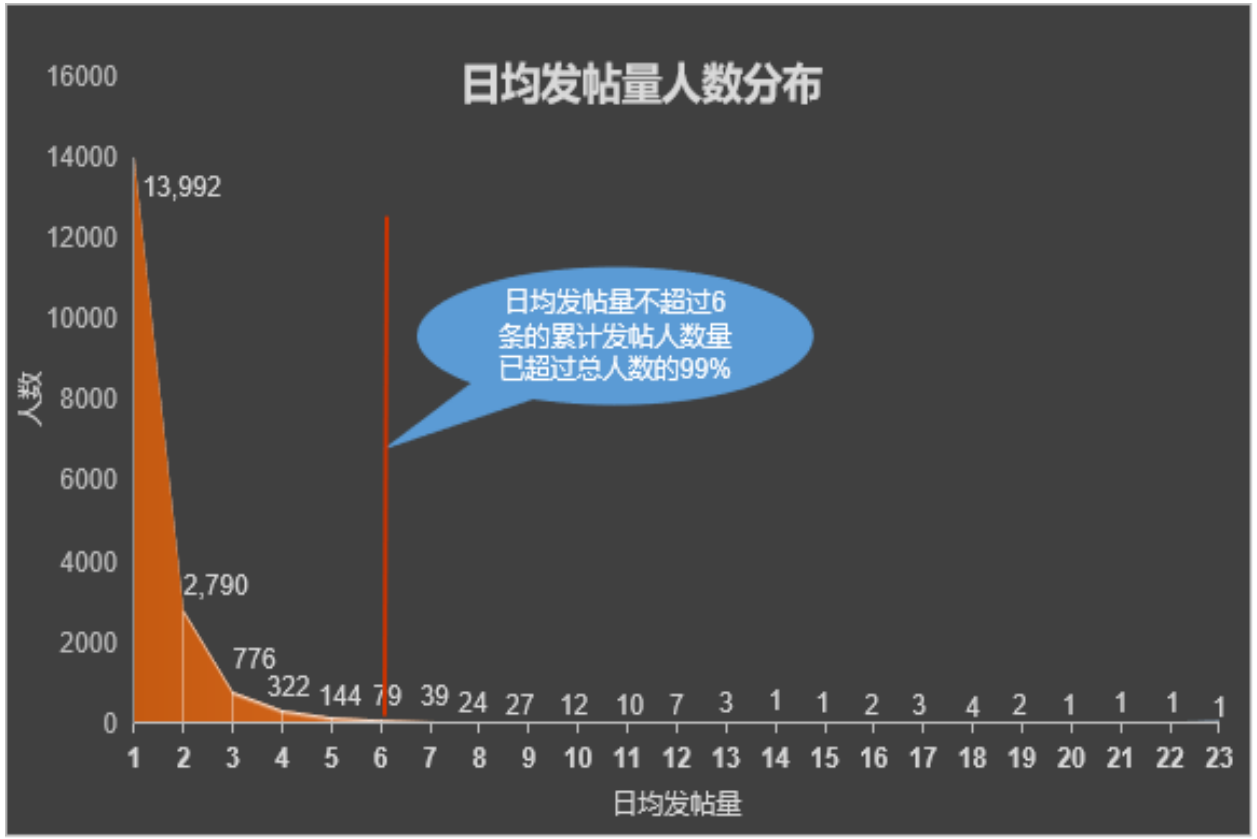
高峰时段：7点和17点，典型早8晚5型用车
工作性质：政府单位

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

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

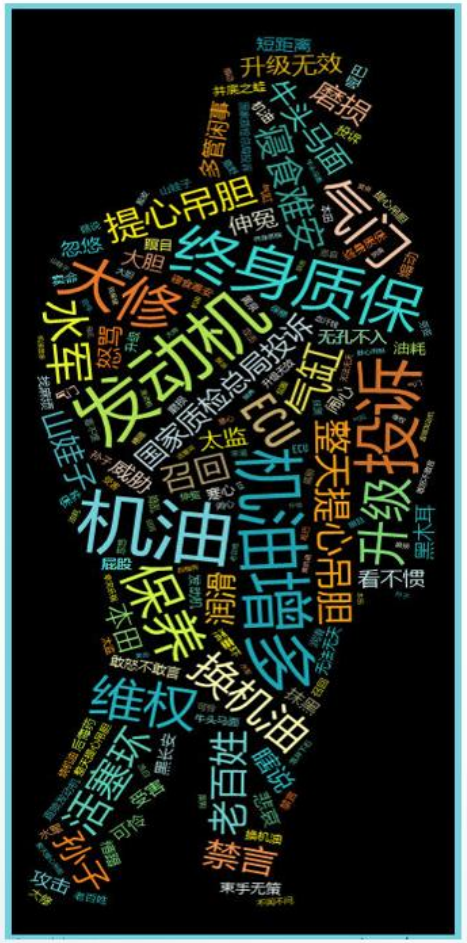
某品牌舆论场画像：从2万到200

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

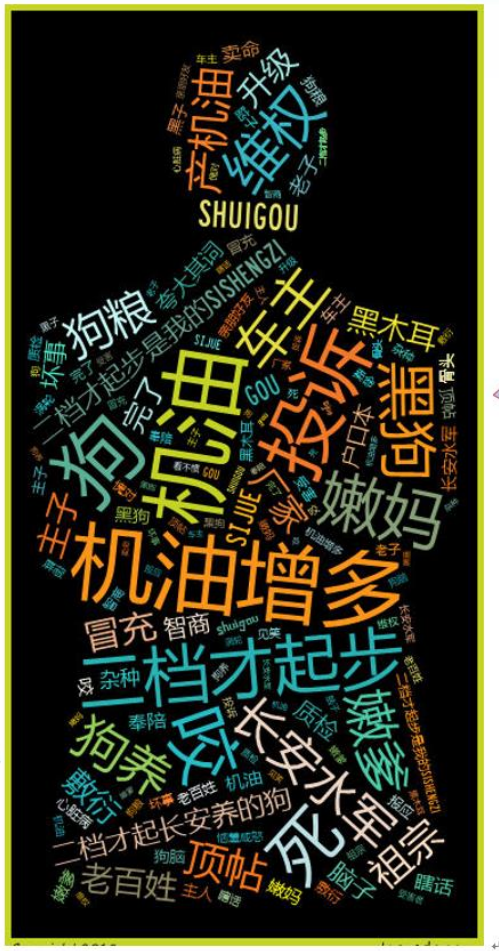
发布平台：

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

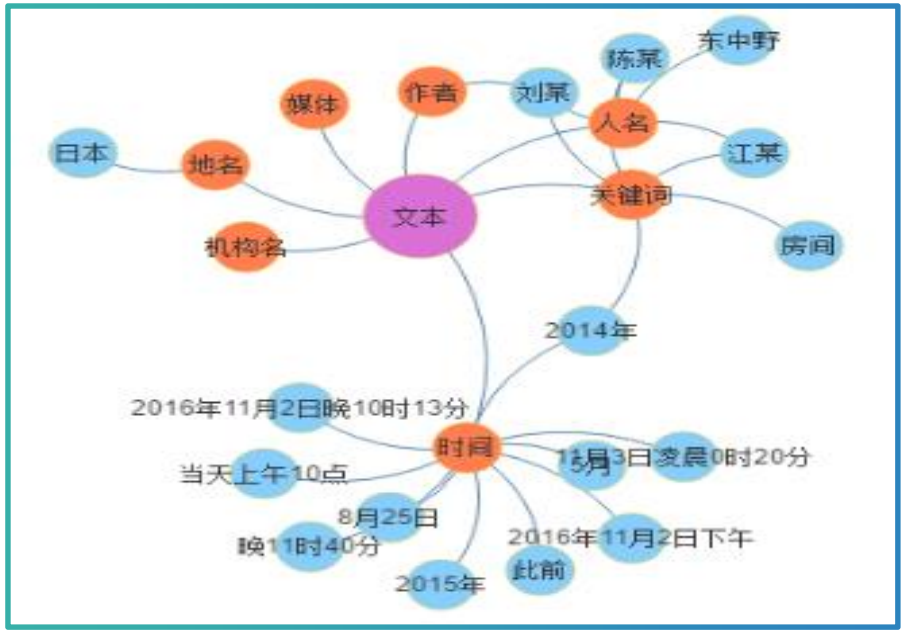
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

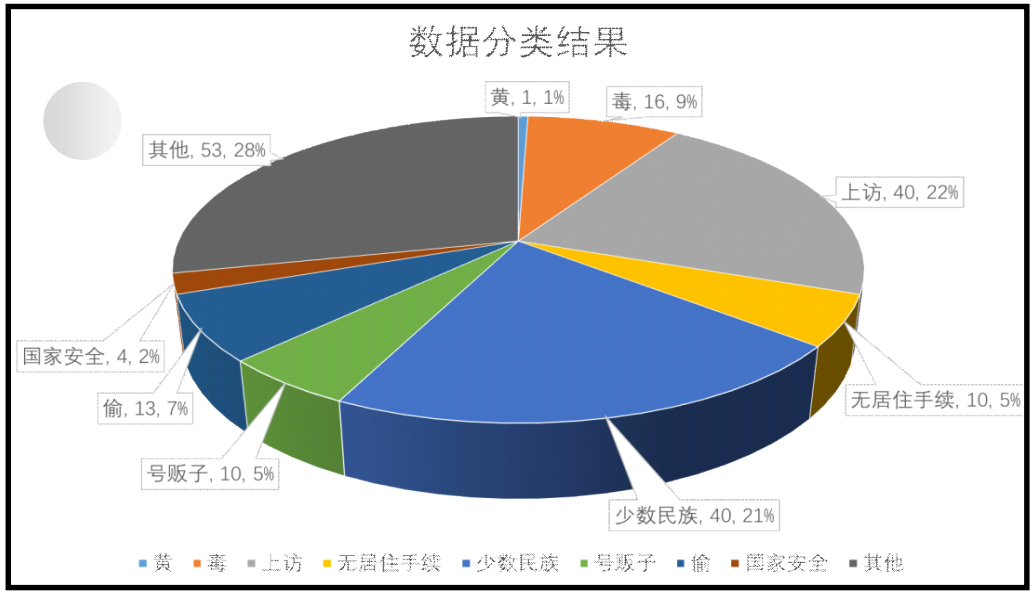
公安语义增强应用

NLPIR语义理解

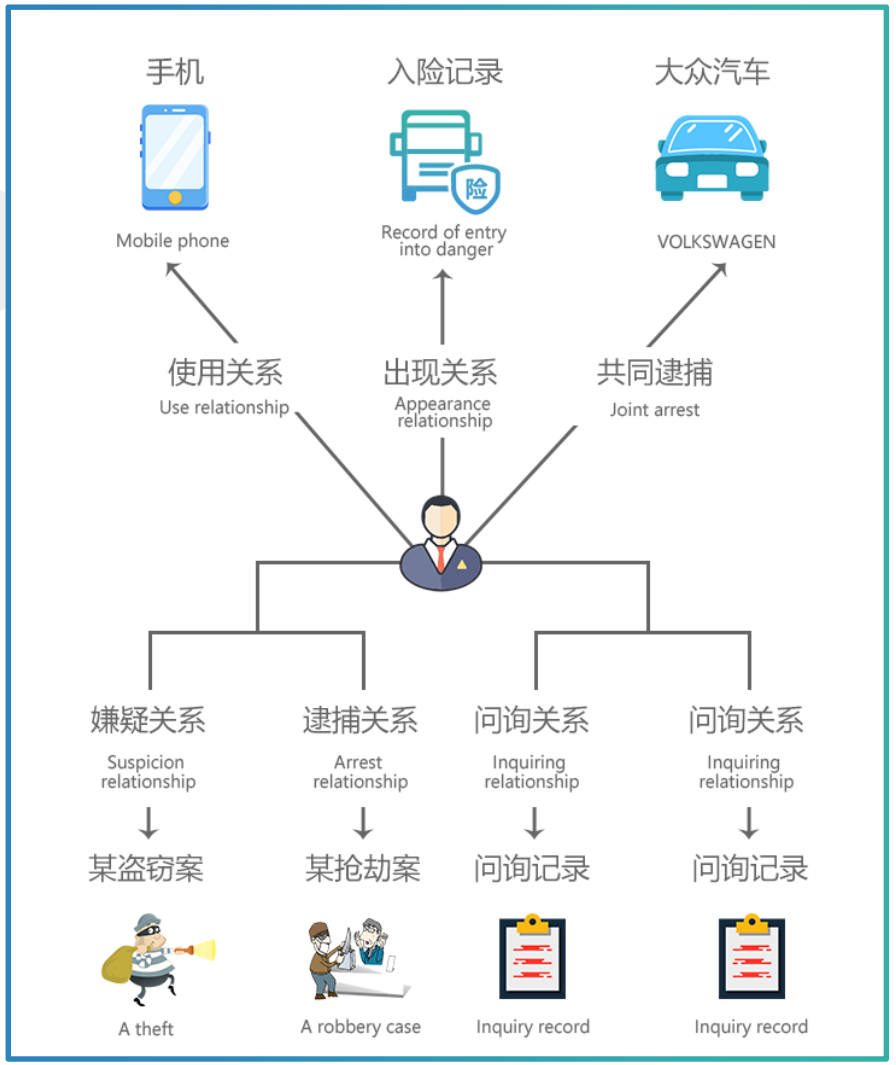
繁 拼

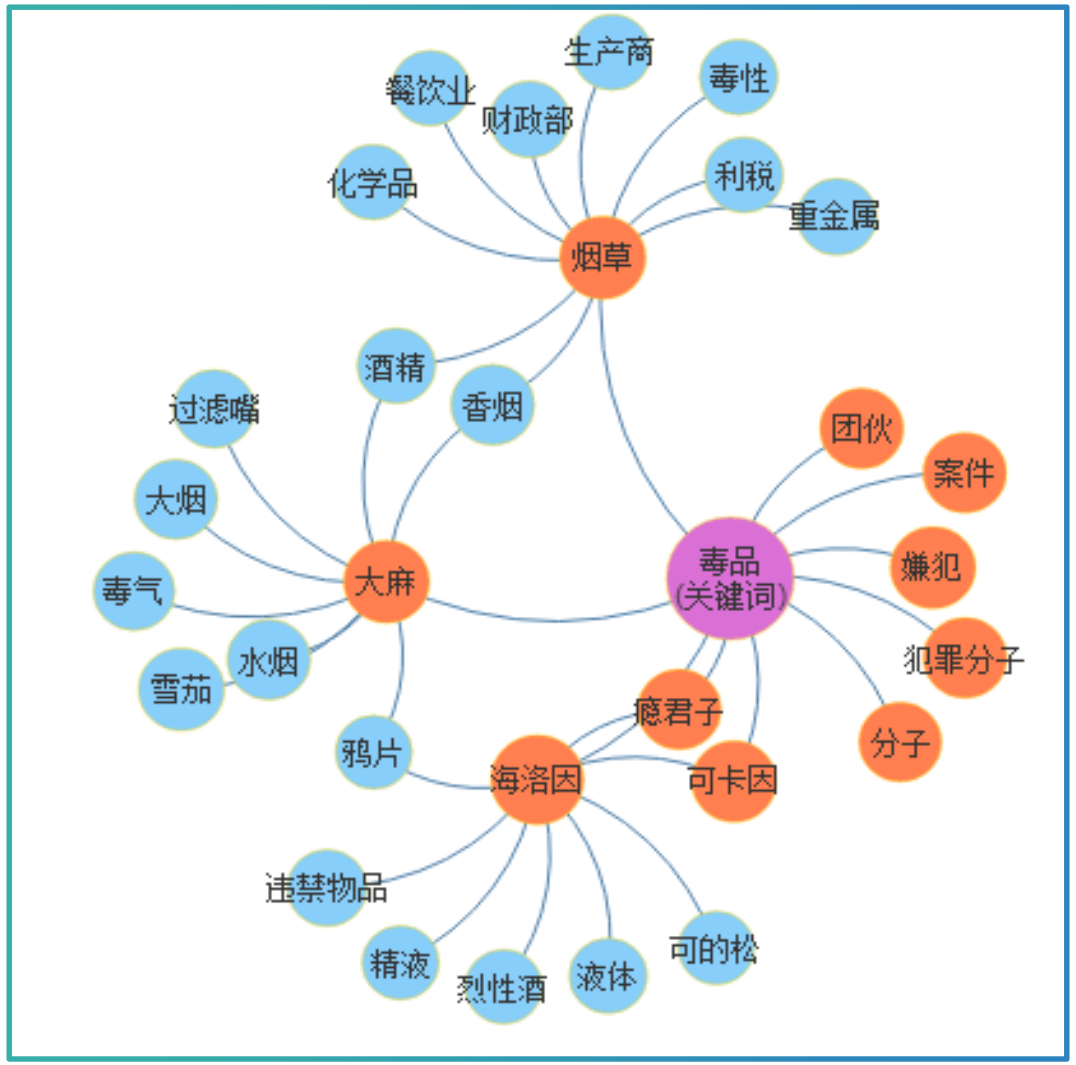
2014年，刘某到日本留学。
2015年底，刘某因与舍友相处不愉快，要求学校调整宿舍，被安排到姜某宿舍。
两人相识，并发现是老乡，两家相聚仅10公里。
5月起，刘某与结识不过一个月的陈某同居。
8月25日，双方发生矛盾，刘某被赶出家门，到打工店长家借住。
2016年11月2日下午，陈某找到江某住处，试图与刘某角色感情问题。
此时，刘某独自一人在房间里。应刘某的请求，善良勇敢的江某专程从外面赶回
2016年11月2日晚10时13分，江某到达东中野站。原本仅需10分钟即刻到家的江某
晚11时40分，陈某尾随刘某江某到达二人住所，约20分钟后开始敲门。刘某躲进
11月3日凌晨0时20分左右，在刘某两次报警间隙，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

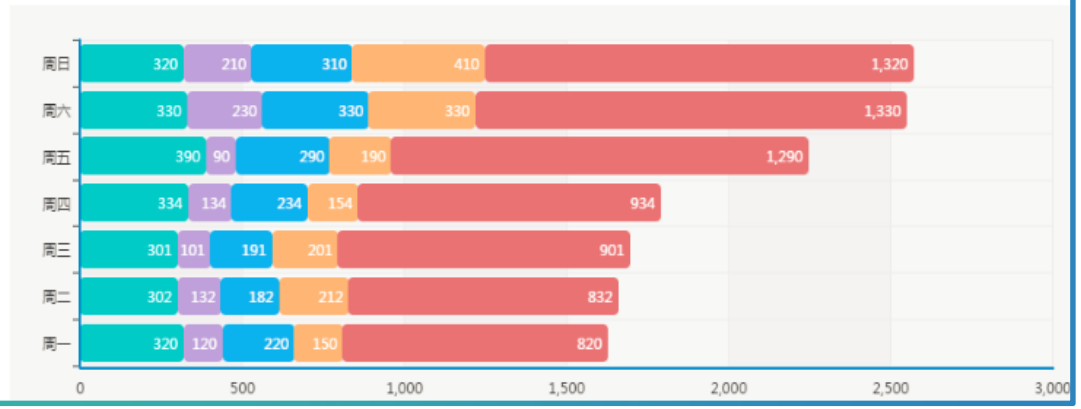

生产商
餐饮业
财政部
毒性
化学品
利税
重金属
烟草
酒精
香烟
过滤嘴
团伙
案件
大烟
毒品
(关键词)
嫌犯
毒气
大麻
犯罪分子
水烟
雪茄
瘾君子
分子
鸦片
海洛因
可卡因
违禁物品
精液
烈性酒
液体
可的松

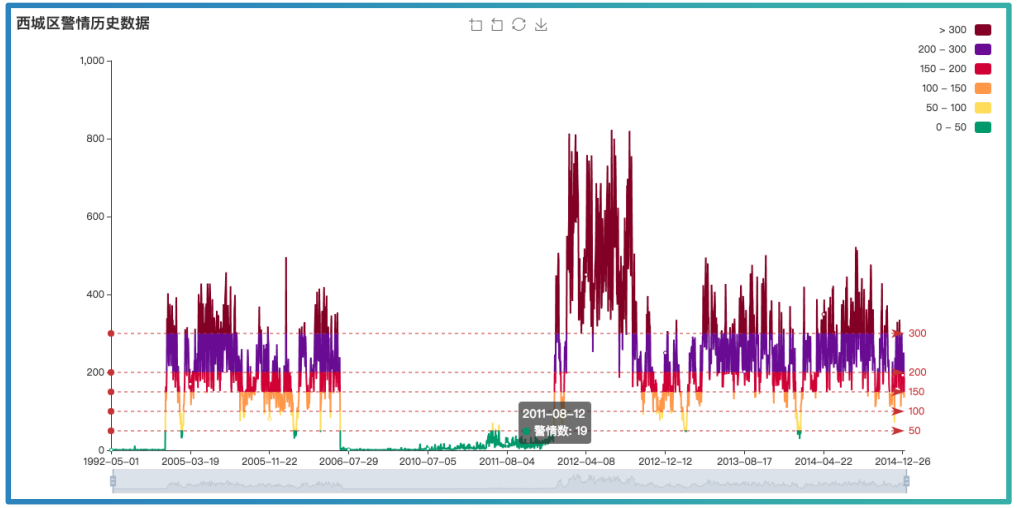
公安语义增强应用

聚类高发

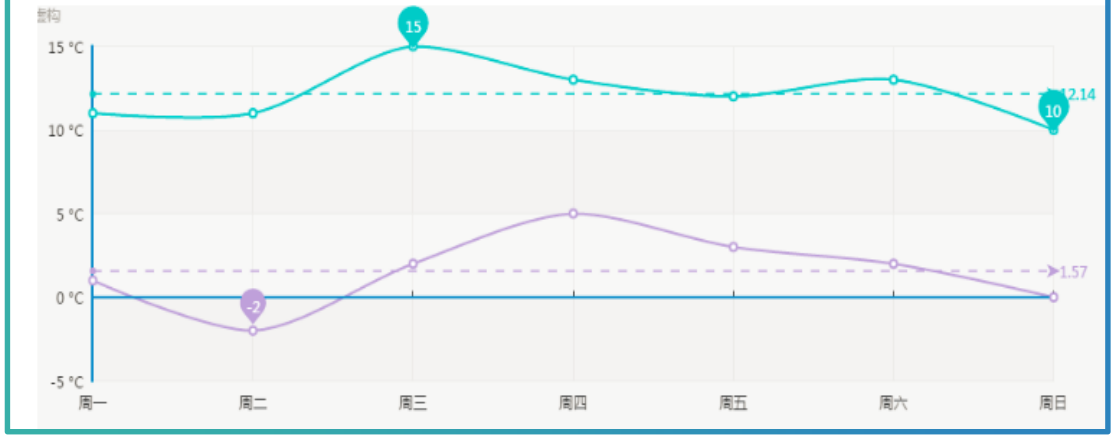
Cluster 0: 天作 使用费 非法占有
Cluster 1: 伪造 交通事故 赔偿金 保险公司
Cluster 2: 博亿 房屋买卖 虚构
Cluster 3: 鲁迅文学奖 作家协会 诗词
Cluster 4: 骗走 人民币 平房 公墓 灵泉 塔位 灵塔 地葬位
Cluster 5: 押金 天纺 柯东 冒用 丰体
Cluster 6: 为名 培训 假借 团伙 考证 帮助

现实高发





预警提示



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）

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。

ISIS: 记者隔离策略

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