



数据结构与算法 (Python) -00/引子

刘云淮 Yunhuai.liu@pku.edu.cn

<http://www.yunhuai.net/DSA2025/CoursePage/DSA2025.html>

北京大学计算机学院

目录

- › 自我介绍
- › 这是一门什么课?
- › 课程内容与目标
- › 往年我们做什么?
- › 我们的教材
- › 联系方式
- › 有用的软件和网站



自我介绍

› 刘云淮

教授、博导、“博雅”特聘教授

- 清华大学计算机 学士
- 香港科技大学计算机 博士
- 中国科学院深圳先进技术研究院 副研究员
- 公安部第三研究所物联网中心 副主任、副研究员、二级警督
- 2016年加入北京大学
- 2019年获得国家杰出青年称号

› 研究兴趣

智慧城市

移动边缘计算

群智感知网络



这是一门什么课？

- › 首先，这不是一门关于编程的课
- › 本课为你展示：

如何把数据组织起来
进行有效的处理
以解决问题

每一天，有.....

2940亿封电子邮件发送

平均每个地球人每天发送42封

6288个新移动应用可被下载

日均下载量已达3500万

57万6000小时视频

上传到YouTube

3亿5000万张照片

上传到Facebook

如果把它们都印出来，叠起来能有
80个埃菲尔铁塔那么高

500TB数据上传到Facebook

如果用2TB的硬盘储存
每年Facebook要新购65吨硬盘

20亿小时电视与电影

在Netflix上观看

整个因特网的流量信息可以装满

24万亿张DVD光盘

需要6万艘10万吨的油轮运送

2亿3000万条tweets

在Twitter上发布

2020年

全年互联网总数据为1.93ZB

即1,930,000PB

每年数据量增长60%

其中非结构化数据增长80%

衡量数据量的方法

- 计算机基础存储单位：字节
 - 64位机器，一次最多处理8个字节
 - 一张高清照片，通常在2M-4M
 - 1 KB = 2^{10} = 1,024 字节
 - 1 MB = 2^{20} = 1,048,576 字节
 - 一部高清电影，大约在2G-10G
 - 1 GB = 2^{30} = 1,073,741,824 字节
 - 一台普通电脑的存储容量，通常可以容纳500-100部高清电影，或者10万张以上的照片
 - 1 TB = 2^{40} = 1,099,511,627,776 字节
 - BATD等大型互联网企业所拥有的数据量，通常在这个数量级，也是需要大型的数据中心才能处理的
 - 1 PB = 2^{50} = 1,125,899,906,842,624 字节
 - 1 EB = 2^{60} = 1,152,921,504,606,846,976 字节
 - 1 ZB = 2^{70} = 1,180,591,628,517,411,303,424 字节
 - 通常是在视频监控、流媒体等领域才能看到的数据量级，一般都需要分布式存储、分布式处理，很少有单一中心或者单一企业能处理
 - 640 KB 足够所有人用了！
比尔·盖茨，1981
- 通常是全球级的互联网流量

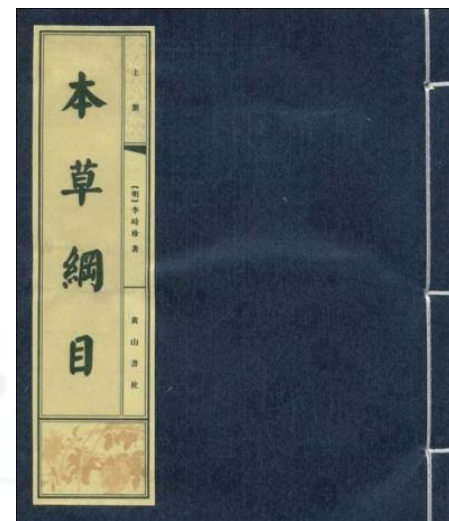
数据的前世今生

中国传统药学著作



神农本草经

上古，先秦，秦汉时期
多位医家集结整理
上中下三卷，载药 365 种



本草纲目

明朝李时珍
历时 27 年编纂，1590 年出版
共 52 卷，载药 1892 种
方剂 11096 个

美国海军上尉和他的大数据实践

- › 马修·方丹·莫里
- › Matthew Fontaine Maury
- › 1806 年出生于美国弗吉尼亚
- › 1824 年刚刚达到入伍年龄便进入了美国海军学校
- › 1839 年，已经晋升为海军上尉的莫里在一次事故中不幸腿部致残
- › 不适合于服役远航的莫里在 1842 年被任命为主管海图和仪器库的负责人



大航海时代的海图



六分仪



经典的书籍、教材



指南针



1459年毛罗地图



郑和航海图



哥伦布航海图

莫里的目标



大量快发霉的航海日志

前人视为垃圾

变废
为宝

莫里的目标



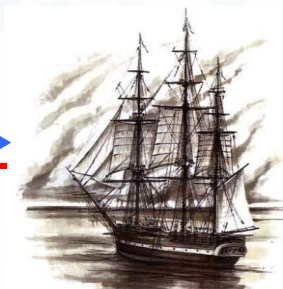
精确的图表

与商船交换信息，在自愿基础上以互利互惠的合作方式开创了国际气象界公开交换环境资料的传统



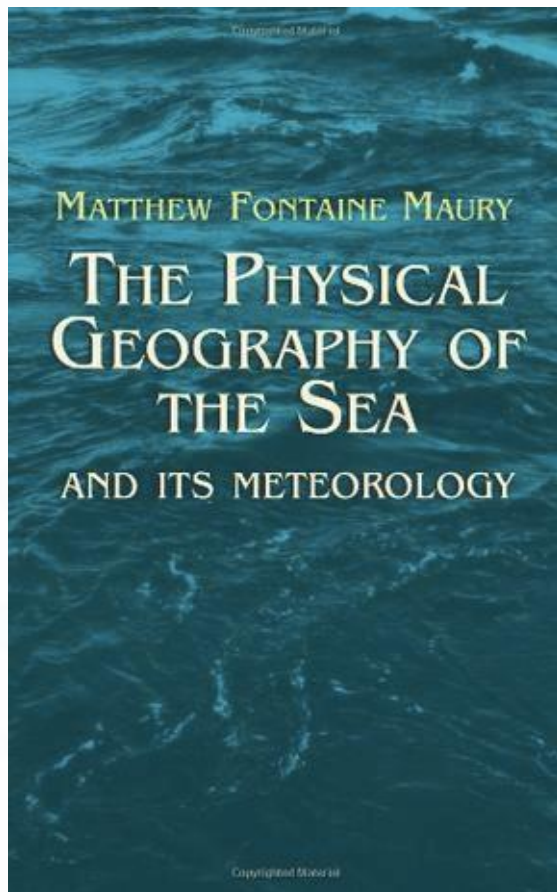
莫里

更新后的图表
航海日志

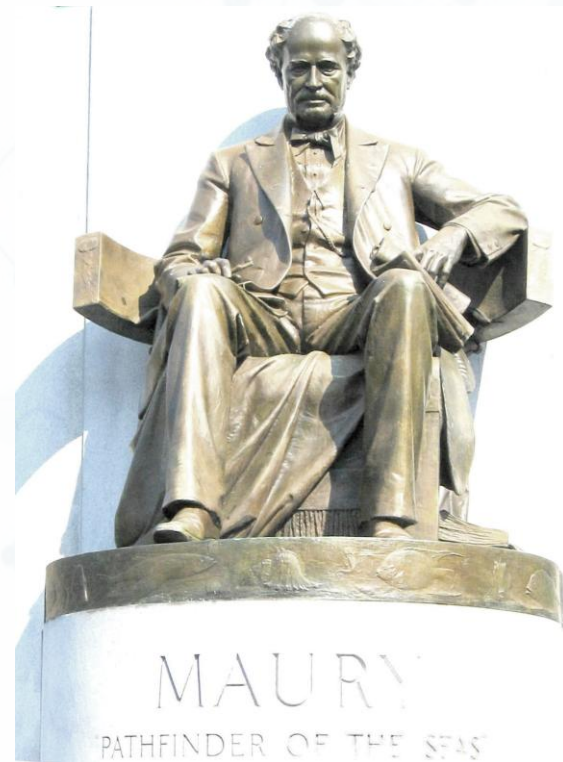


商船

名垂千古：海洋学的奠基人



- 1855 年,莫里出版权威著作《海洋物理地理学和气象学》，被誉为**海洋学的奠基人**
- 当时，他已经绘制了 120 万个数据点
- 四个国家授予了他爵士爵位，包括梵蒂冈在内的其他八个国家还颁给了他金牌奖章
- 即使到今天,美国海军颁布的导航图上仍然有他的名字



信息技术令人类获取数据的能力大幅提高

数据无处不在



交通数据



金融数据



物联网数据



零售数据



社交网络数据



科学数据

无处不在的数据

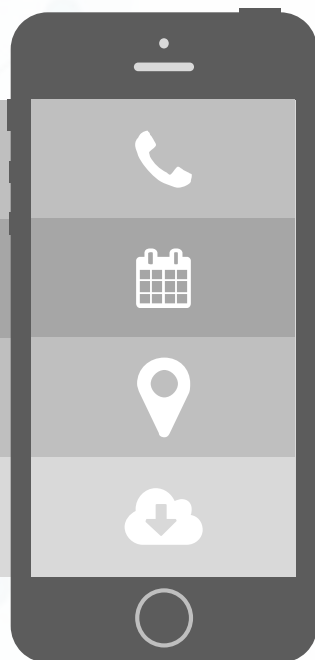
政务数据



商业数据



互联网数据



新闻信息

- ✓ 百度新闻
- 今日头条
- 搜狐新闻
- 网易新闻

...

社交媒体信息

- ✓ 新浪微博
- 网易云音乐
- QQ空间

...

专利信息

- ✓ 佰腾专利
- 中国专利局官网
- 中国专利信息网

...

房地产信息

- ✓ 链家
- 58同城网
- 贝壳二手房网
- 我爱我家租房网

...

公共资源招投标信息

- ✓ 各城市政府招投标监管网官网
- 各城市政府采购官网
- 各城市规划和国土资源局官网
- 各城市联合产业交易所官网

...

招聘信息

- ✓ 51job招聘
- 智联招聘
- 前程无忧

...

空气质量信息

- ✓ Aqistudy
- 各省环境保护厅官网

...

餐饮店铺点评信息

- ✓ 大众点评
- 美团网
- 猫途鹰网
- 百度糯米网

...

如果汽车工业发展也像信息技术一样快…

	容量（内存）	速度 (CPU)	价格
IBM7030 (Stretch)1961	128KB	1.2 MIPS	RMB: 1亿元
Core i9-900K 2000	16G (typical)	81666 MIPS	RMB: 5000块

	容量（乘客）	速度（公里/小时）	价格
一辆1960年的车	5	100	20万
一辆今天的车，应该能达到的性能 （如果按照信息技术发展）	64万人	80万公里/小时	1块

每一个智能手机...



典型和非典型物联网设备



智能设备：每时每刻收集数据



智能手机

地理位置数据
运动数据
环境亮度数据
图像数据
语音数据
...



空气质量数据
温湿度数据
气压数据
...



智能家居设备



身体状况数据
运动习惯数据
实时图像数据
...

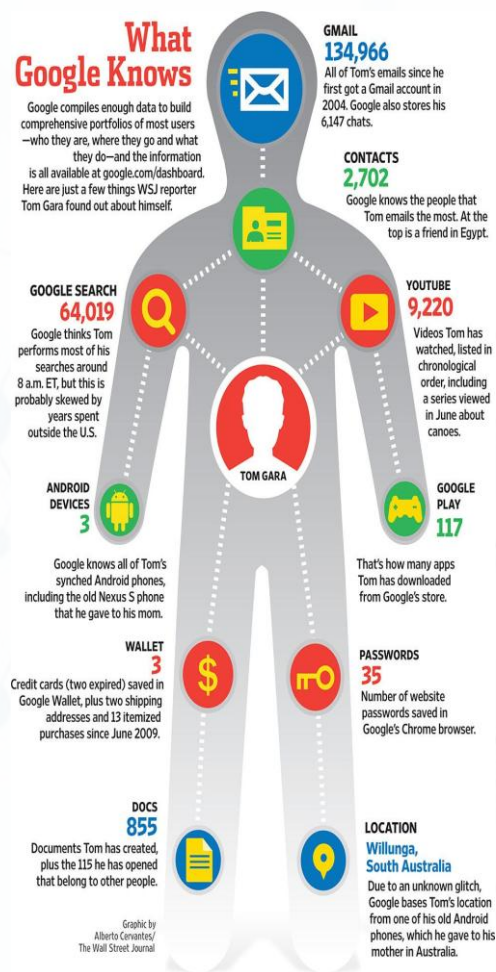


可穿戴设备

信息时代的数字脚印



互联网企业比你更了解你



GoogleTM
is watching



从你的数据中了解你！

从NBA总决赛说起



NBA总决赛，美国观众人数超过2700万

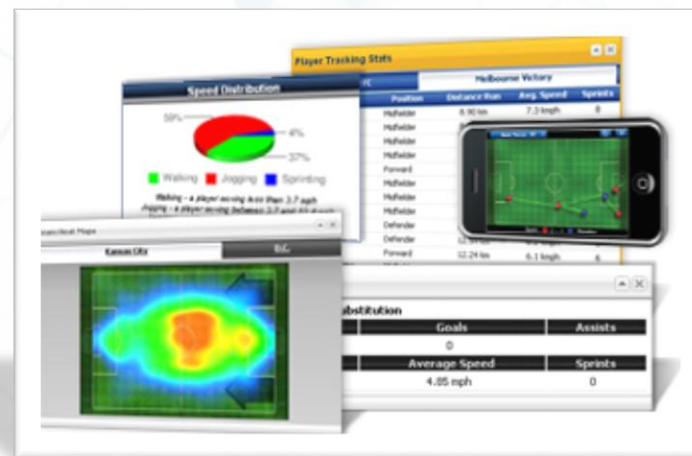


VS



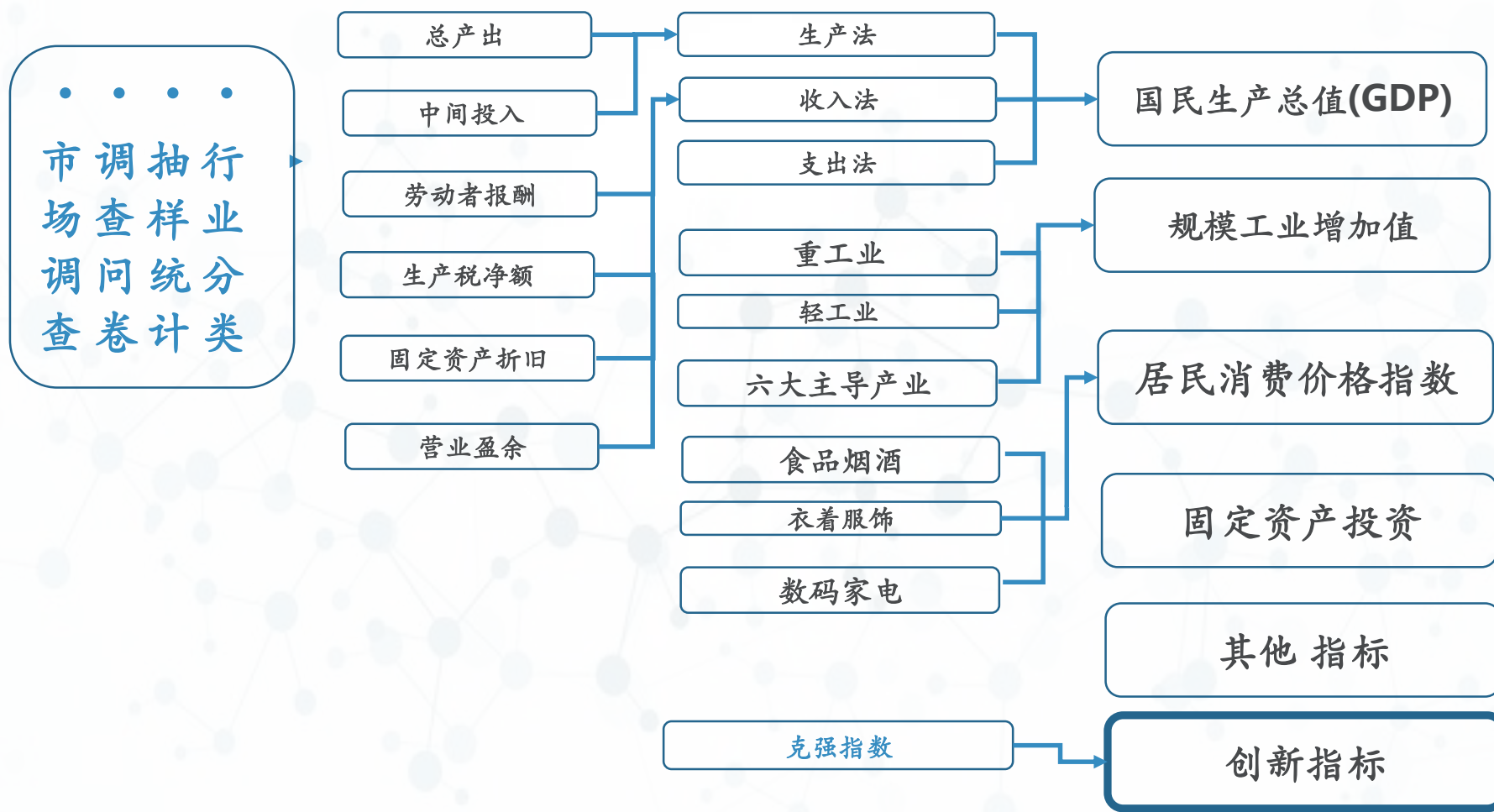


- 遍布传感器，获取海量数据，用于分析球员、比赛各类信息
- 金州勇士是最早应用该系统的NBA球队



宏观经济大数据典型案例

传统宏观经济统计监测



大数据监测 VS 传统统计监测

更快速

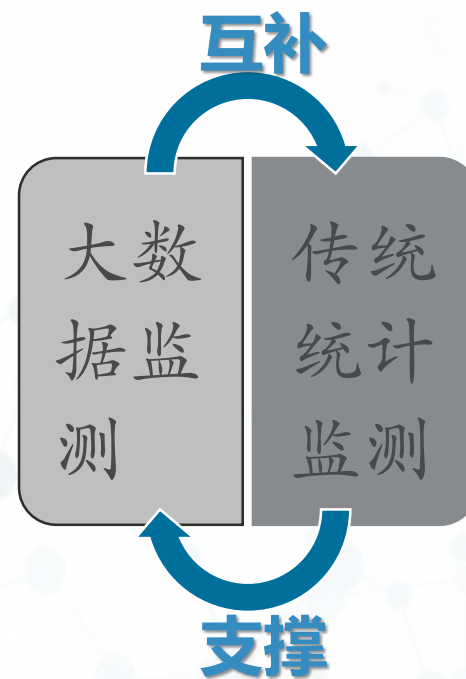
🕒 实时、日报、周报、月报

更精细

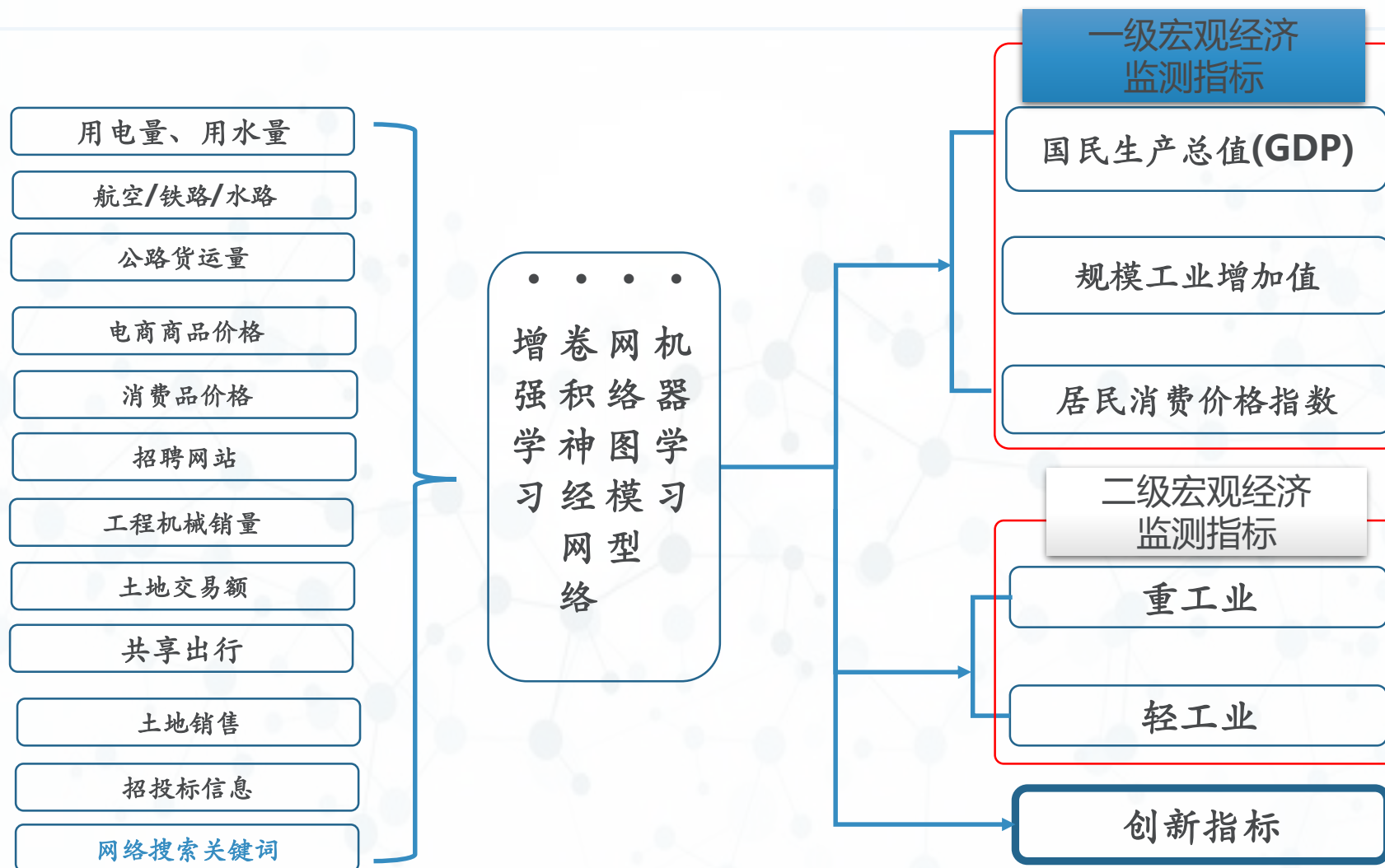
📊 细分行业、新兴产业

更灵敏

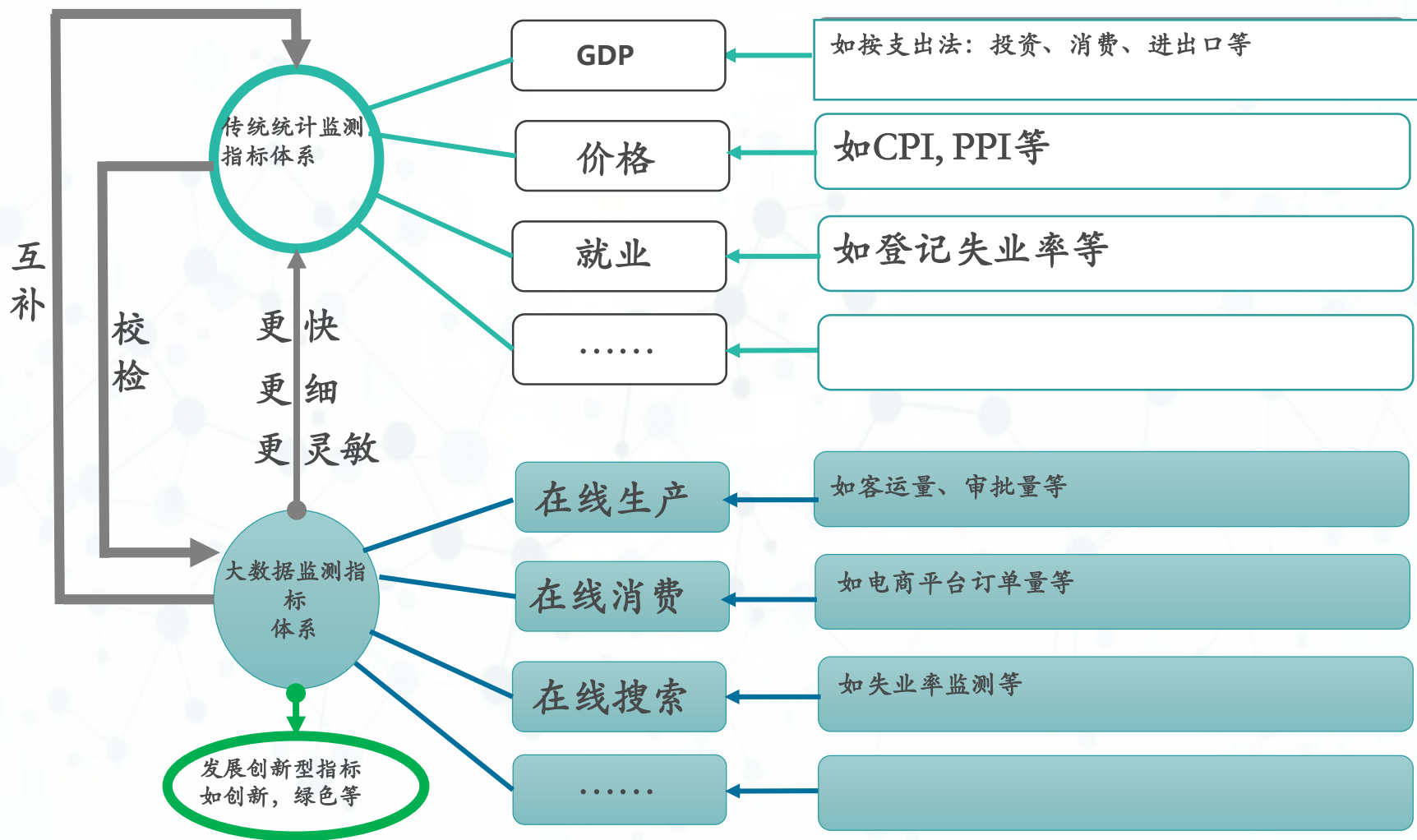
⚙️ 测量经济变动（流量）> 测量经济体量（存量）



大数据宏观经济监测预警的思路



新型宏观经济监测指标体系



多元化数据采集机制



爬取：海量互联网数据自动化采集



采购：与企业沟通，购买相关数据



合作：与政府合作，获得统计数据



众包：数据堂进行数据众包采集

居民失业率

数据科学入门案例 (Python)

登记失业率

城镇登记失业人员情况

表 号：A502001号
制表机关：人力资源和社会保障部
批准机关：国家统计局
批准文号：国统制字[2010]20号
有效期至：2010年10月
单位：人

地区	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
合计	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

注：1. 本表为城镇登记失业人员情况，不包括农村失业人员。
2. 本表为城镇登记失业人员情况，不包括农村失业人员。
3. 本表为城镇登记失业人员情况，不包括农村失业人员。
4. 本表为城镇登记失业人员情况，不包括农村失业人员。
5. 本表为城镇登记失业人员情况，不包括农村失业人员。
6. 本表为城镇登记失业人员情况，不包括农村失业人员。
7. 本表为城镇登记失业人员情况，不包括农村失业人员。
8. 本表为城镇登记失业人员情况，不包括农村失业人员。
9. 本表为城镇登记失业人员情况，不包括农村失业人员。
10. 本表为城镇登记失业人员情况，不包括农村失业人员。
11. 本表为城镇登记失业人员情况，不包括农村失业人员。
12. 本表为城镇登记失业人员情况，不包括农村失业人员。
13. 本表为城镇登记失业人员情况，不包括农村失业人员。
14. 本表为城镇登记失业人员情况，不包括农村失业人员。
15. 本表为城镇登记失业人员情况，不包括农村失业人员。
16. 本表为城镇登记失业人员情况，不包括农村失业人员。
17. 本表为城镇登记失业人员情况，不包括农村失业人员。
18. 本表为城镇登记失业人员情况，不包括农村失业人员。
19. 本表为城镇登记失业人员情况，不包括农村失业人员。
20. 本表为城镇登记失业人员情况，不包括农村失业人员。



手续繁琐，主动
登记的人较少

调查失业率



劳动力调查表

20 年 月

应在本户登记的人：

调查时正在本户的人：

本户人口中，外出半年以上的人：

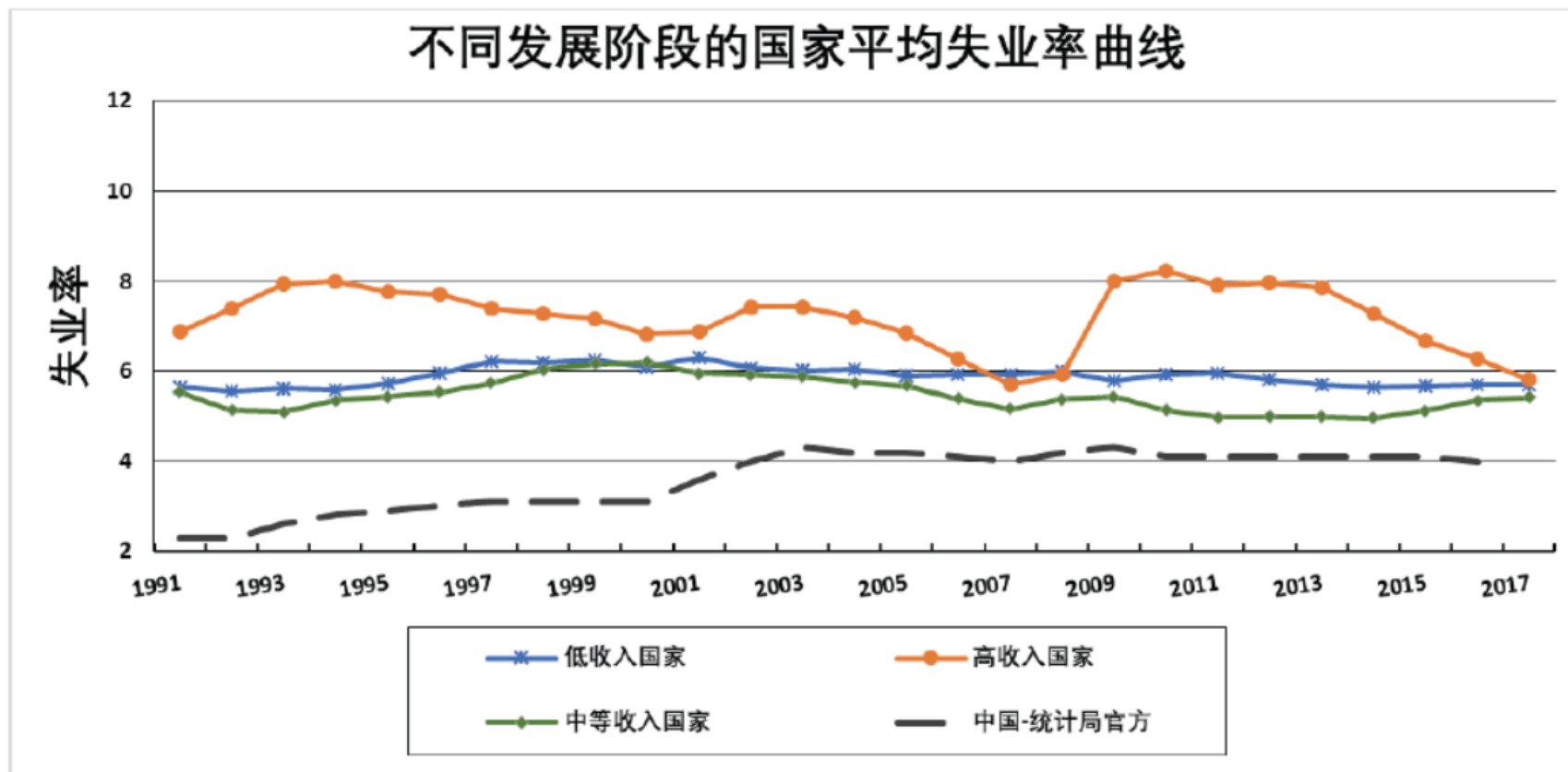
本户地址：_____ 户主姓名：_____ 户主性别：_____ 户主年龄：_____ 户主职业：_____

姓名	性别	年龄	职业	是否在本户居住
1. 户主	男/女	岁		是/否
2. 户主配偶	男/女	岁		是/否
3. 户主子女	男/女	岁		是/否
4. 户主父母	男/女	岁		是/否
5. 户主兄弟姐妹	男/女	岁		是/否
6. 户主其他亲属	男/女	岁		是/否
7. 户主其他人员	男/女	岁		是/否

一个拥有400万人的地
级市，每月调查样本仅
800户，共2100多人

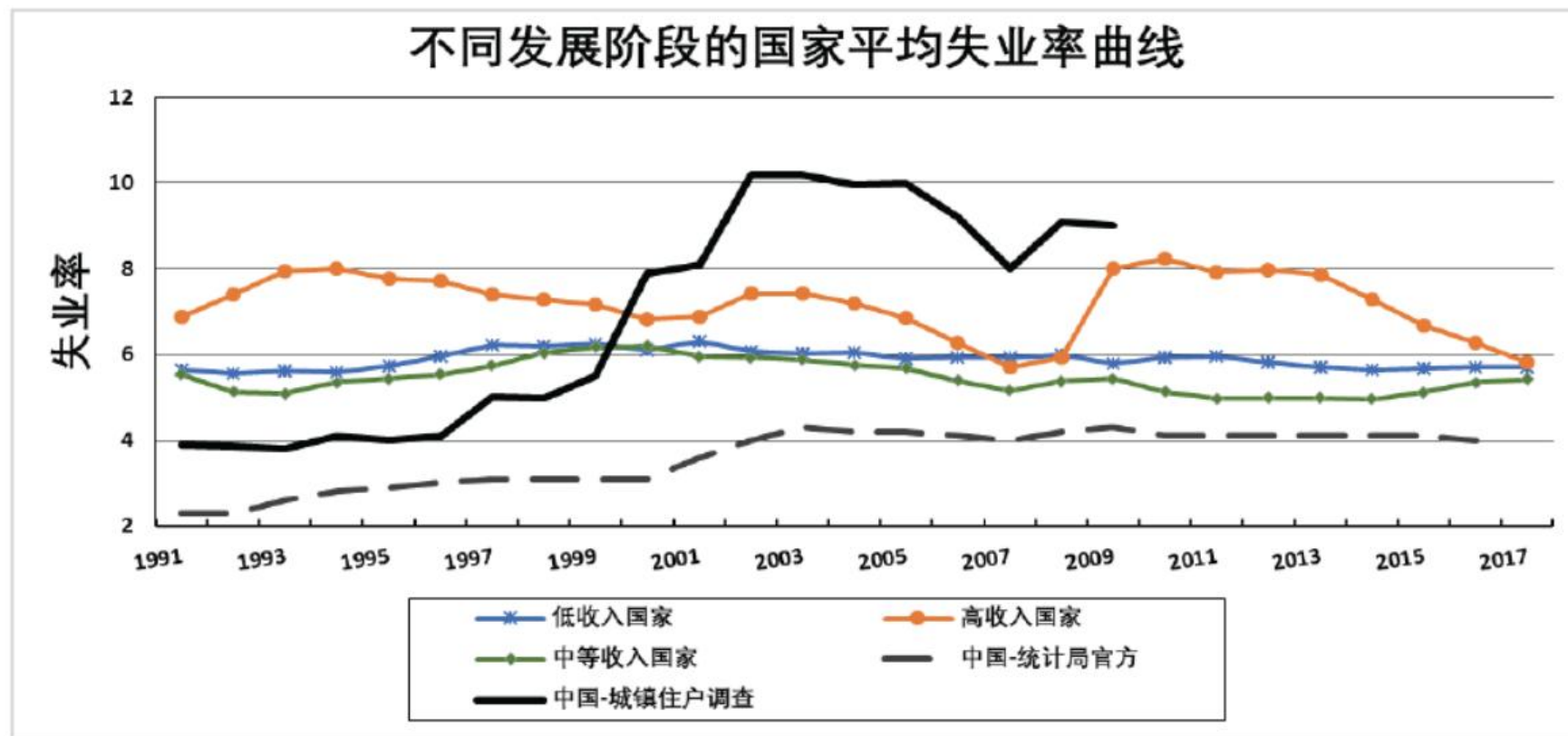
平均失业率（官方数据）

右图为不同发展阶段国家的平均失业率曲线
最下方虚线为中国官方数据



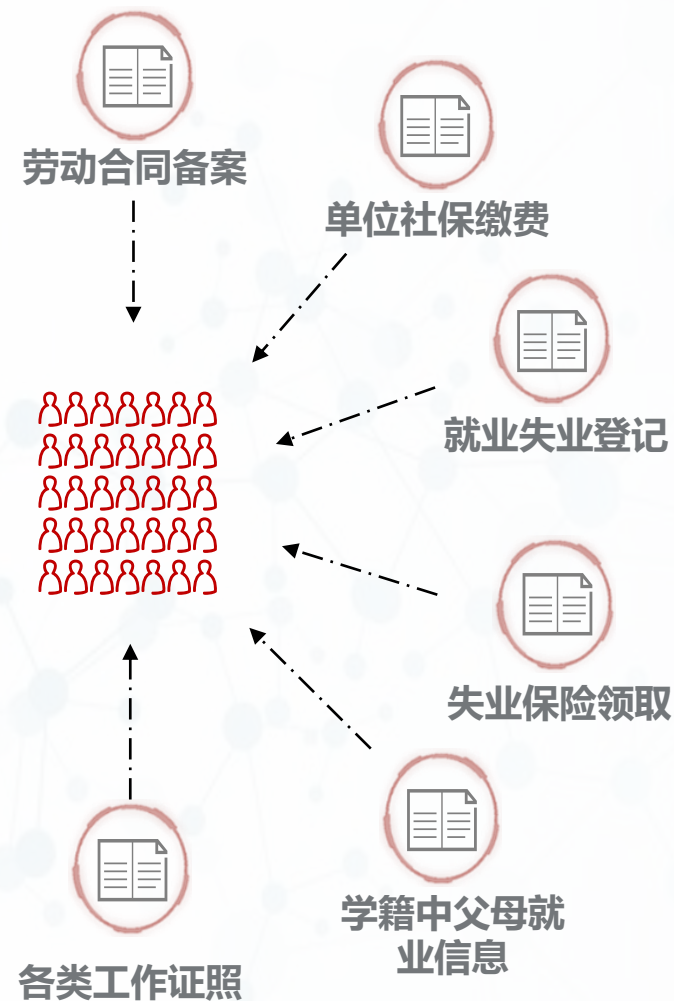
平均失业率（住户调查）

黑色实线为使用中国城镇住户调查数据计算的失业率



数据来源：世界银行、UHS调查、Feng et al. (2017) 与北京大数据研究院

个人层面：基于大数据的失业/就业分析



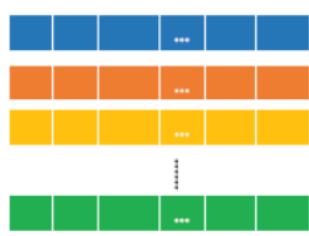
个人大数据模型构建

失业
人口

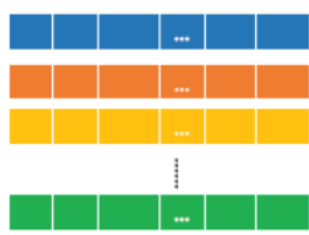
基本信息
婚姻父母
子女信息
五险一金
缴费
就医信息
违法信息

构造一百多个
跟失业 / 就业
相关的特征

就业
训练样本



失业
训练样本



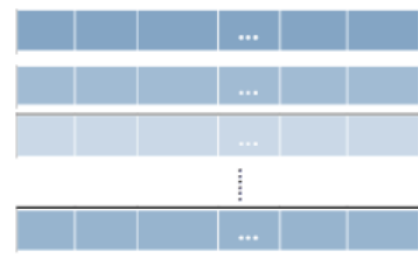
构造有标注的样本
(50万人)

机器学
习模型

用机器学习算
法训练模型

迁移学习

无标
注的
数据



(350万人)

训练模型在全样本上估计失业 / 就业情况

模型训练的结果

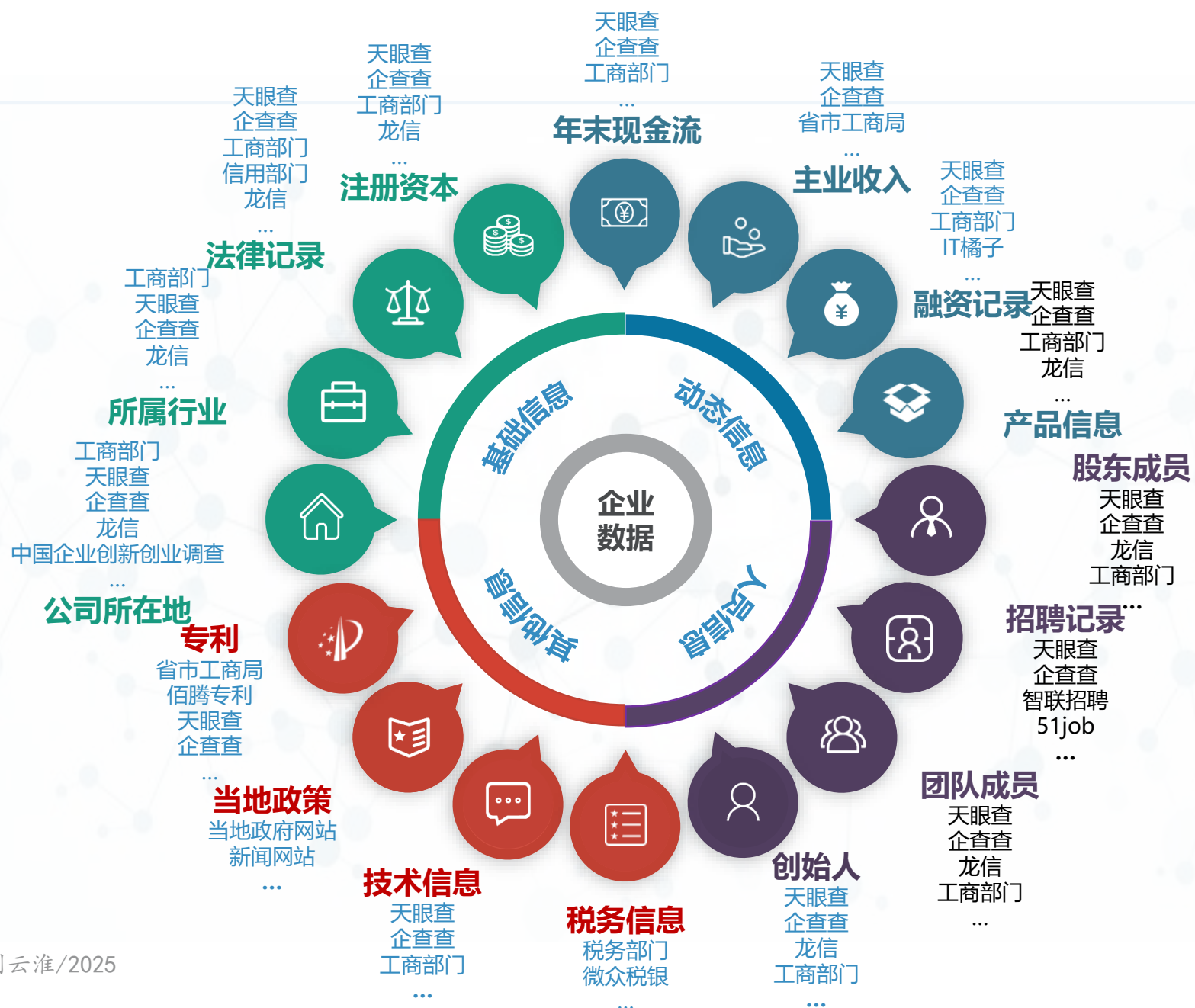


模型集成结果

交叉验证结果	预测为失业	预测为就业
真实失业样本	24624 😊	24 ☹️
真实就业样本	1164 ☹️	409925 😊

交叉验证准确率：99.72%，失业标签召回率：99.90%

企业



企业财务报表

数据科学计算 (Python)

营业收入

利润总额

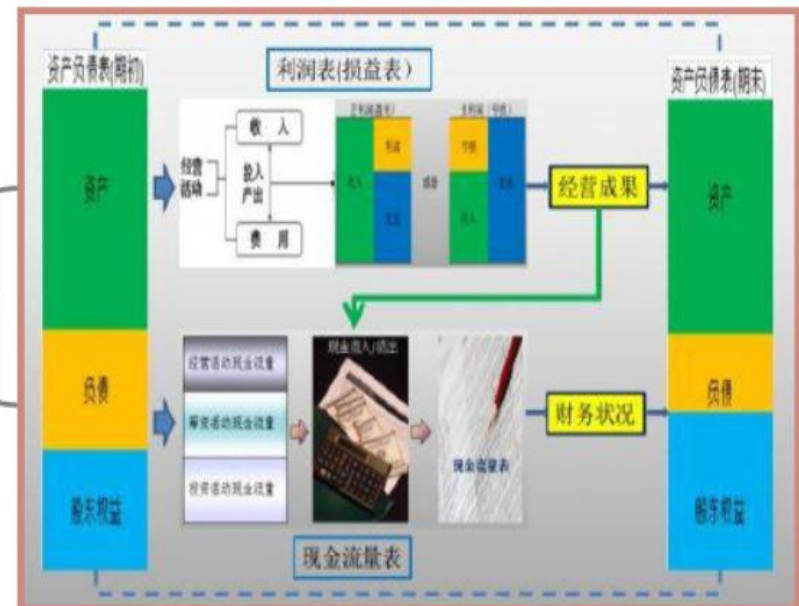
经营成本

从业人员数量

收集与计算

企业财报

利润表	2017-12-31	资产负债表	2017-12-31	现金流量表	2017-12-31
报告期	年报	报告类型	年报	报告类型	年报
期间跨度	12个月	报表范围	合并报表	报表范围	合并报表
数据来源	合并报表	负债资产		报表类型	合并报表
利润表摘要					
营业收入	23,801,400.00	现金及现金等价物	1,109,400	净利润	7,151,000
同比(%)	56.23	交易性金融资产	13,166,900	加：折旧与摊销	2,361,100
营业收入变动	17,153,600.00	应收款项合计	641,700	营运资本变动	2,366,200
营业利润	6,647,600.00	应收账款及票据	457,100	其他非现金调整	-1,094,400
同比(%)	32.13	其他应收款	184,600	经营活动产生的现金流量净额	10,614,000
税前利润	6,821,500.00	存货		投资活动：	
同比(%)	70.82	其他流动资产	189,900	出售固定资产收到的现金	2,800
净利润	7,151,000.00	递延资产合计	15,116,900	减：资本性支出	3,200,400
同比(%)	74.01	非流动资产：		投资减少	9,230,400
非经常性损益	1,666,900.00	固定资产净值	12,477,500	减：投资增加	16,223,400
扣非后归母净利润	5,484,100.00	权益性投资		其他投资活动产生的现金流量净额	551,400
同比(%)	59.80	持有至到期投资		投资活动产生的现金流量净额	-9,609,200
研发费用	1,745,600.00	可供出售投资		筹资活动：	
EBIT	6,647,600.00	其他长期投资	5,975,000	债务增加	5,019,300
EBITDA	9,008,700.00	商誉及无形资产	2,100,200	减：债务减少	2,118,100
利润表摘要(NON-GAAP)		土地使用费	27,100	股本增加	17,100
净利润(NON-GAAP)		其他非流动资产	706,100	减：股本减少	
稀释每股收益(NON-GAAP)		非流动资产合计	10,055,900	支付的股利合计	506,200
		总资产	25,172,800	其他筹资活动产生的现金流量净额	246,700
资产负债表摘要		总负债		筹资活动产生的现金流量净额	2,609,800
流动资产	17,844,600.00	流动负债：		现金净流入：	
固定资产		应付账款及票据		汇率变动影响	-256,100
权益性投资	14,458,100.00	应交税金	227,100	其他现金流量调整	
资产总计	55,467,200.00	交易性金融负债		现金及现金等价物净增加额	3,379,500
流动负债	15,174,000.00	短期借款及长期借款当期	775,400	现金及现金等价物期初余额	7,190,200
非流动负债	12,563,900.00	其他流动负债	7,203,200	现金及现金等价物期末余额	10,569,700
负债合计	27,737,900.00	流动负债合计	8,205,700	货币币种	CNY
股东权益	27,709,300.00	非流动负债：		原始币种	CNY
归属母公司股东权益	25,607,400.00	长期借款	3,581,200	记账汇率	



企业:基于大数据的企业活跃度分析

中国工商企业数据

- 注册数据: 注册时间、地点、行业、所有制类型、经营范围、股东总数、股东认缴资本
- 年报数据: 资产总额、营业总收入、主营业务收入、纳税总额、从业人数、利润总额、净收入

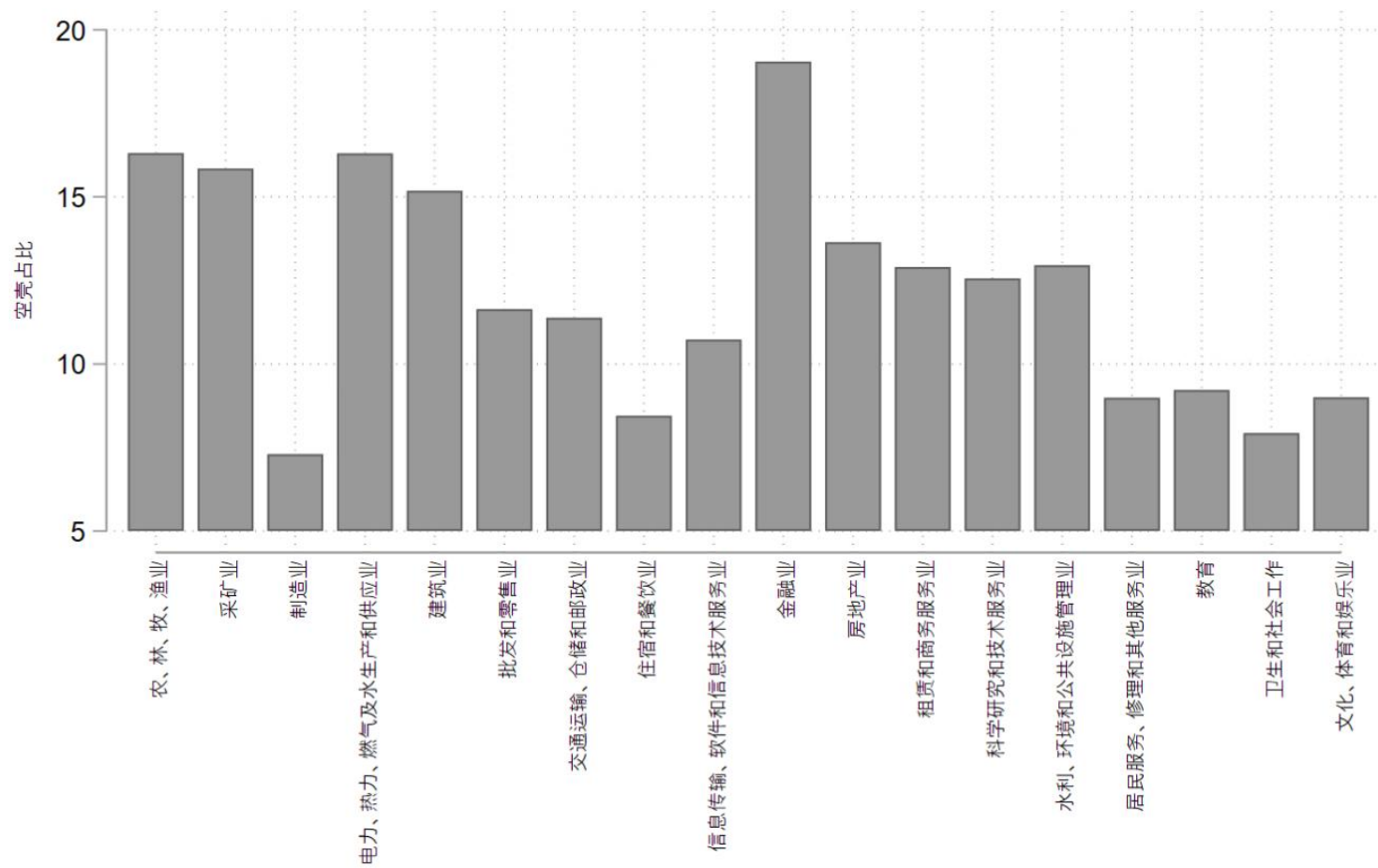
企业外部数据

- 税务数据: 税务信息
- 司法风险: 司法信息
- 经营风险: 司法执行信息
- 公司发展: 融资历史、核心团队、企业业务等
- 经营状况: 行政许可、资质证书、招投标信息等
- 知识产权: 商标、专利、软件著作权等

2018年中国企业创新创业调查 (ESIEC 2018)

- 辽宁、上海、浙江、河南、广东、甘肃6个省117个县
- 58500个样本, 11700家个体户, 46800家公司制企业
- 添加企业存续状态标签 (391个空壳)

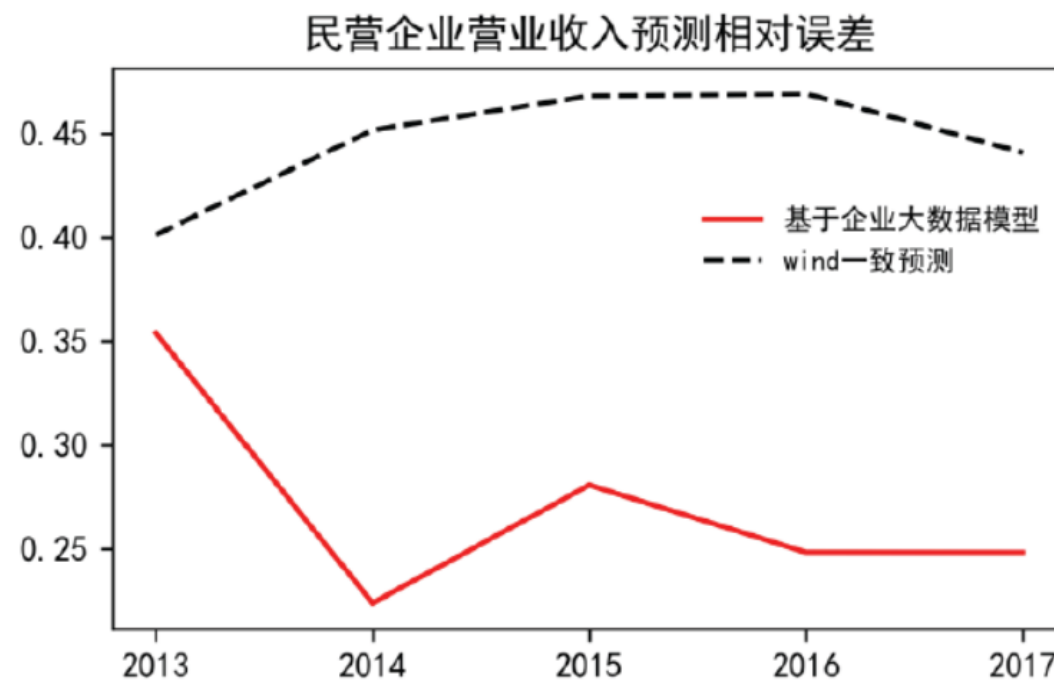
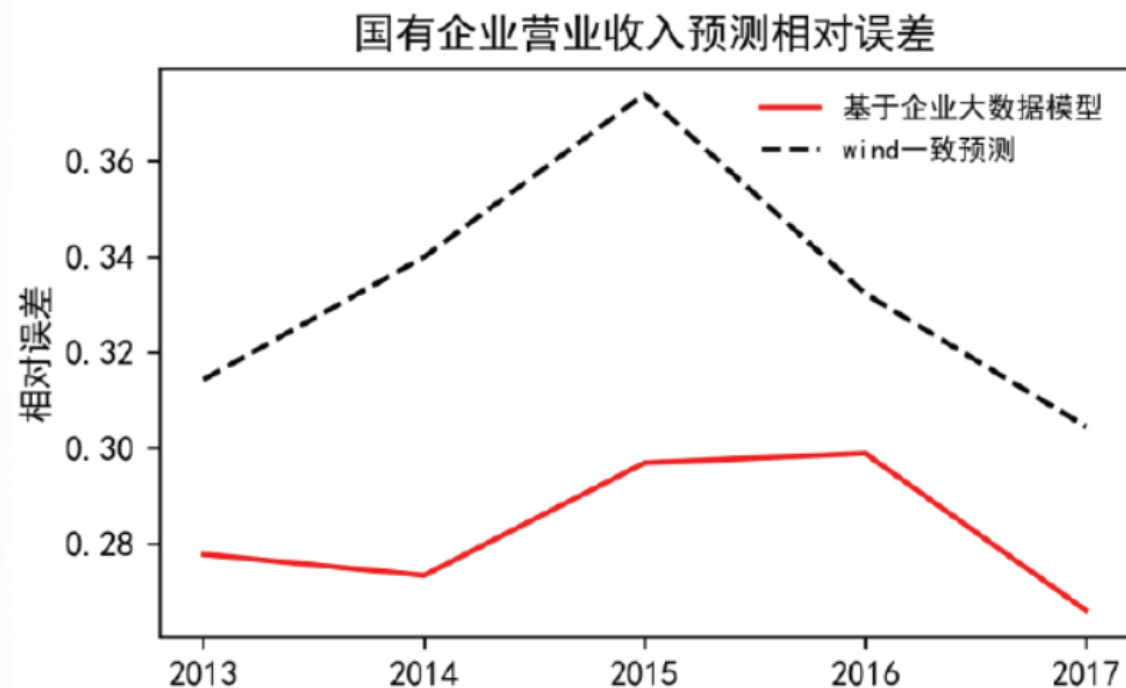
企业:基于大数据的企业活跃度分析



空壳企业概率行业分布

- 金融业
- 农、林、牧、渔业
- 采矿业
- 电力、热力等供应业
- 卫生和社会工作
- 制造业
-

企业大数据建模结果



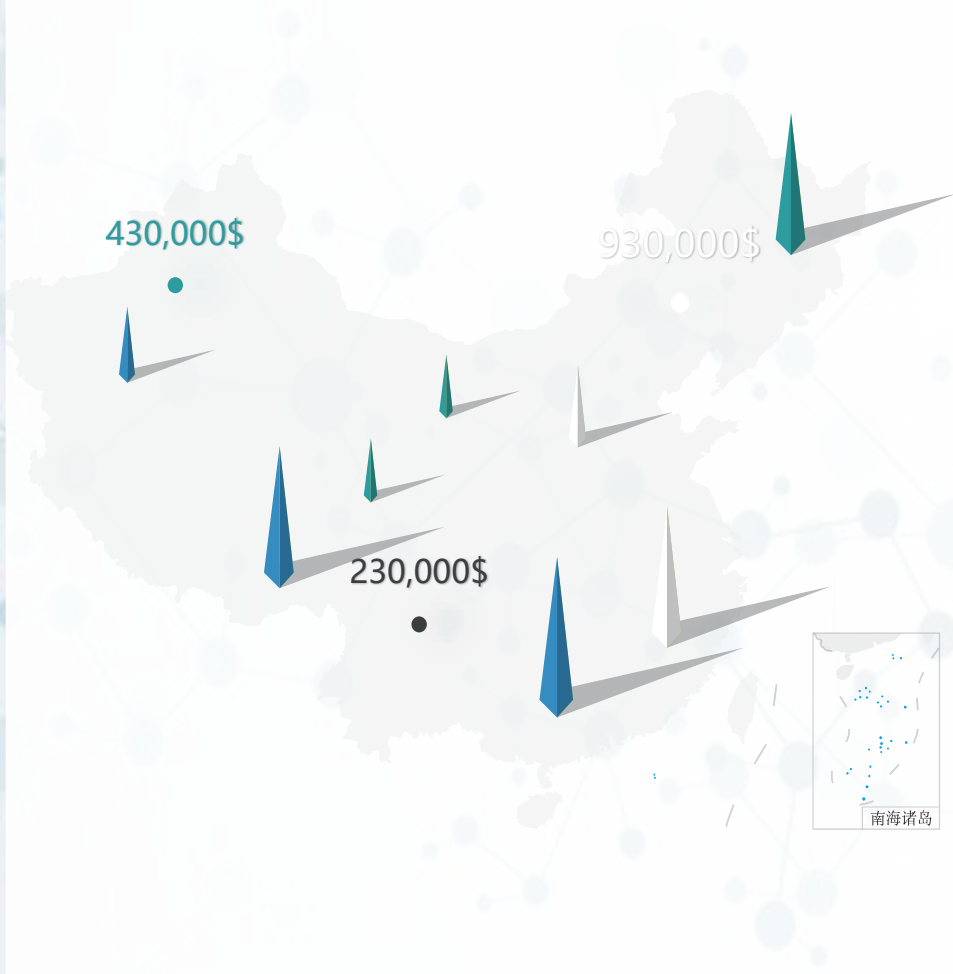
主要发现

- 基于大数据方法的营业收入预测误差明显低于专业分析师的一致预测
- 使用及时可得的企业大数据可以显著提高预测效果

地区



区域产业发展对标分析



重点地区定向对标

01

从国内、全省、重点对标城市三个维度进行呈现，国内、全省维度会针对各产业进行产业重点指标的横向对标，分析自身产业比较优势，便于在发展政策等方向进行重点部署

产业集群定型对标

02

针对产业集聚区，通过与全国领先的产业集聚区进行对标分析，以此借鉴同类型产业集群建设经验，保障区域产业优质高效发展。

重点企业定点对标

03

通过与国内外各产业链内重点企业进行对标分析，以此为依据对区域产业发展具有引领性的重点企业进行针对性扶持。

企业风险预警预测

对海量企业相关风险信息进行识别、分析、归纳处理，对企业风险进行评价和检测，对可能暴雷的企业进行前瞻性的评估与预警，做到早发现、早评估、早控制，最大程度降低企业风险。

01

僵尸企业

02

空壳企业

03

休眠企业

04

套利企业

05

虚开企业

06

域外经营企业



异常企业预警



企业黑名单



区域投资风险



中小企业信贷

交通大数据的探索

城市数据融合分析





500 K
10 TB

Transportation



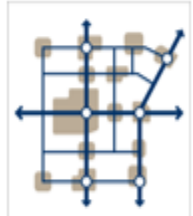
10 M
1 TB

Telecom



16 M
1 TB

Finance

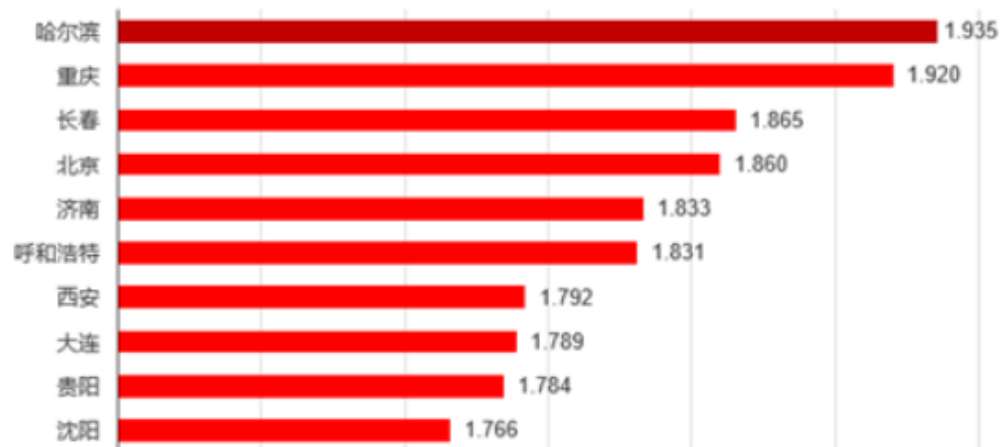


70 K
100 K

Geography

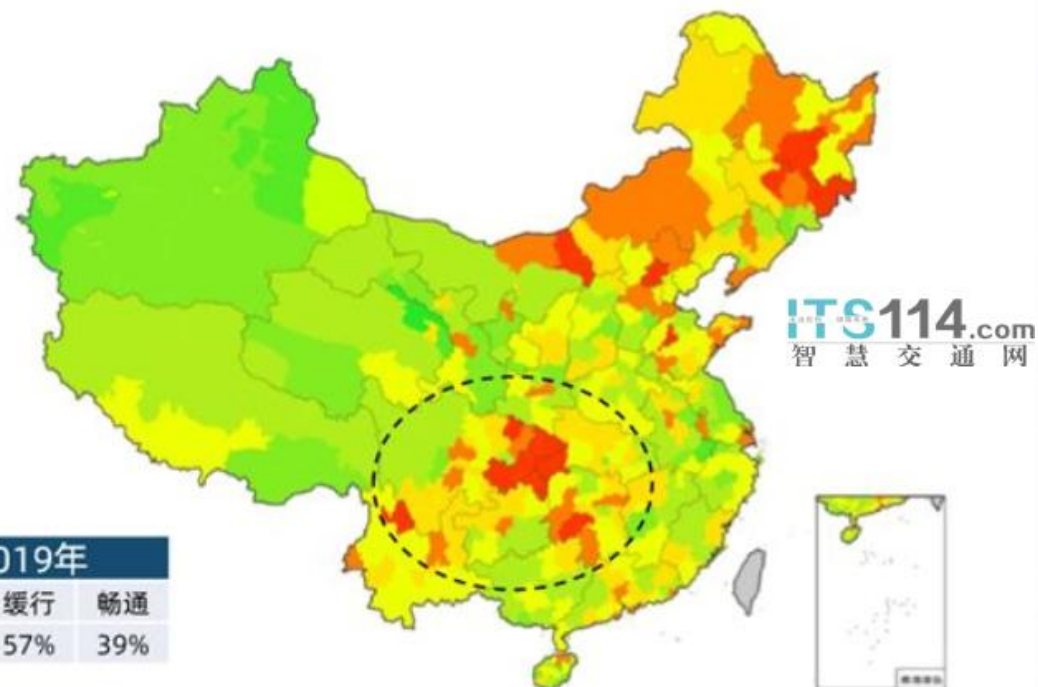
交通拥堵，人民心中的痛

年度高峰拥堵延时指数TOP10



- 1/3城市通勤拥堵
- 32城市拥堵指数 >1.8
- 线城市、省会城市拥堵严重
- 发达城市富贵病

2019年度全国拥堵地图

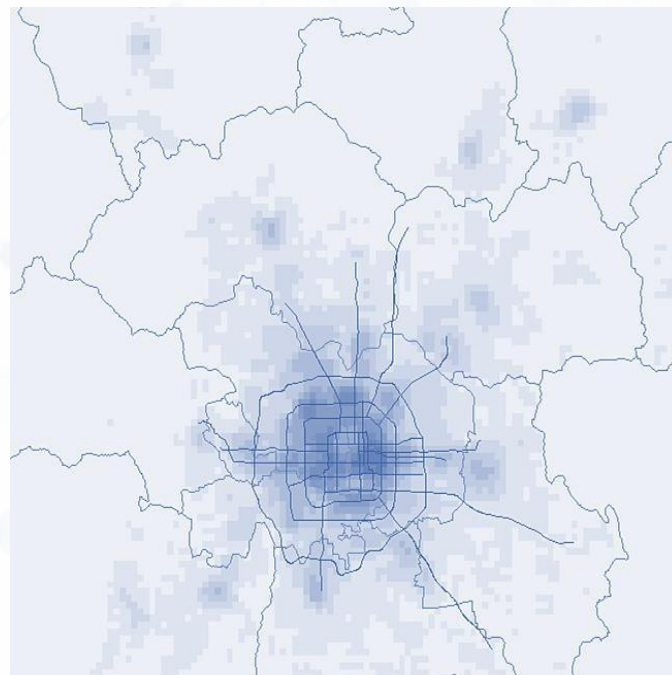
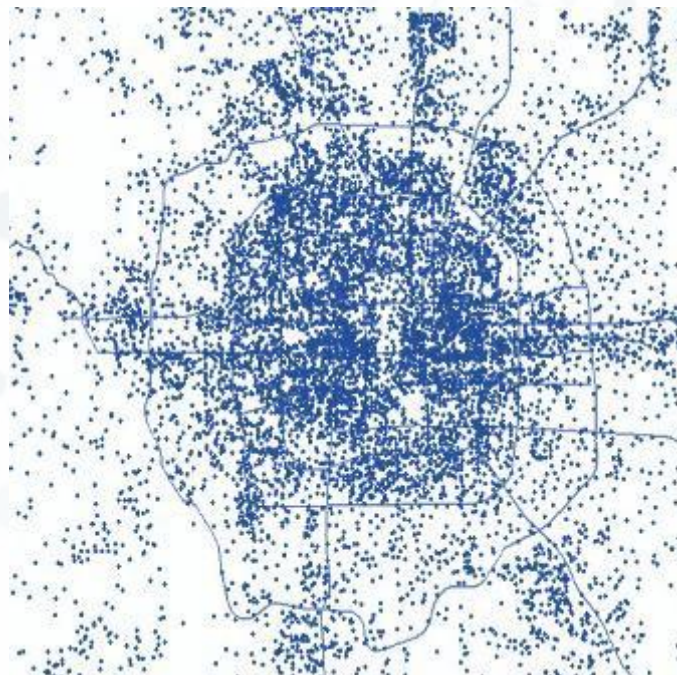
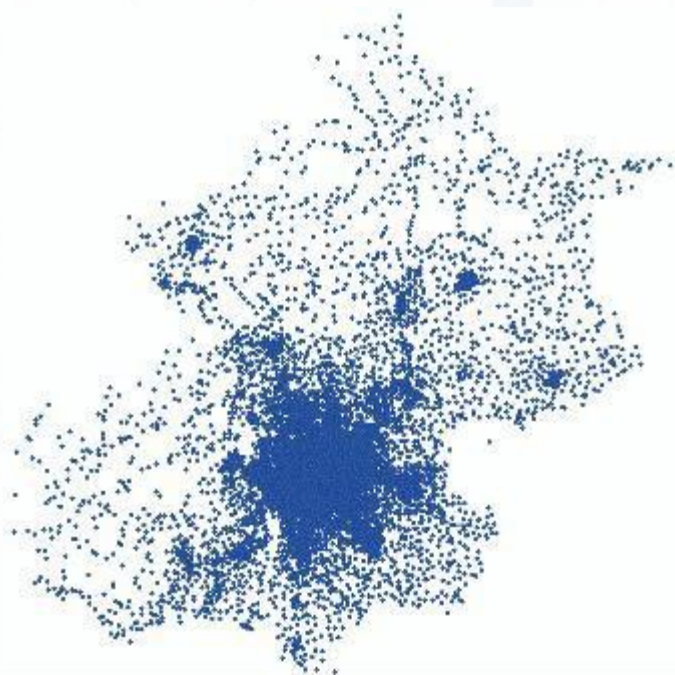


2019年

拥堵	缓行	畅通
4%	57%	39%

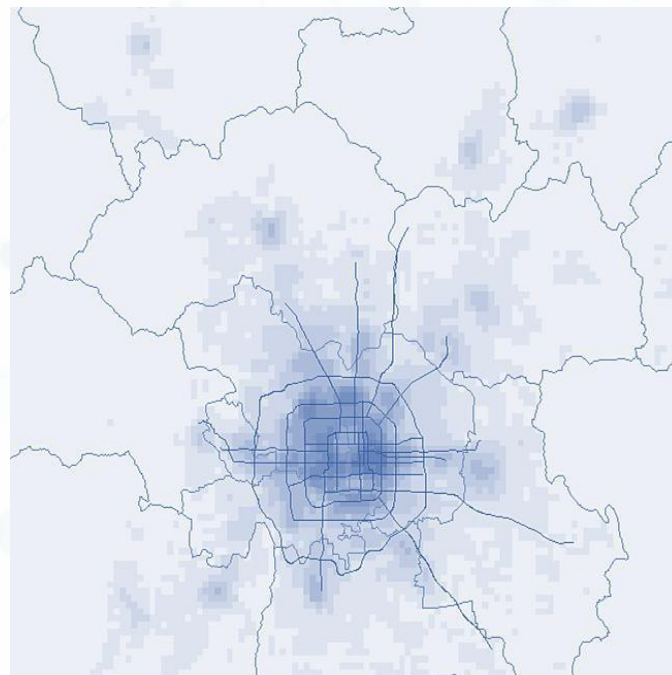
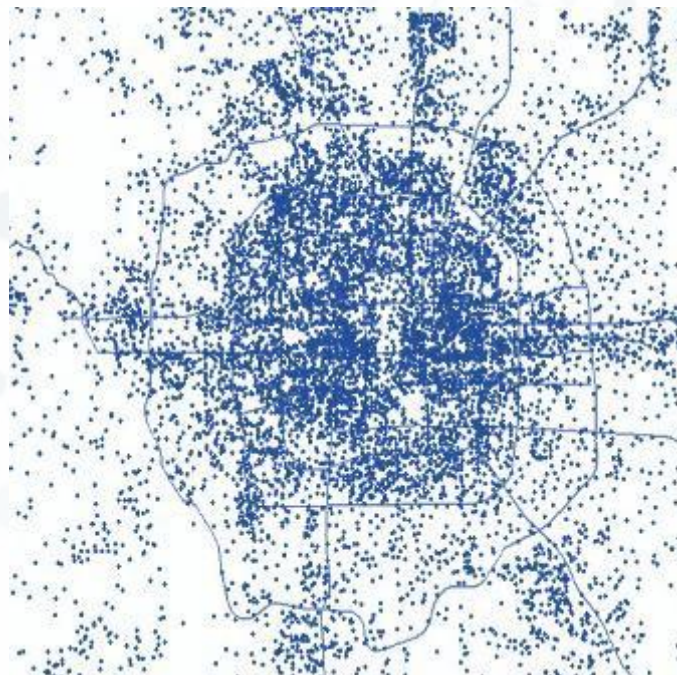
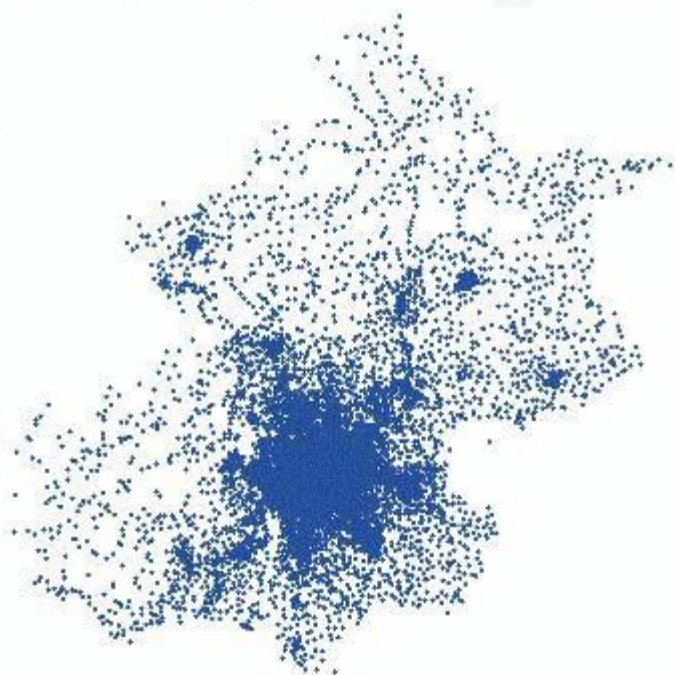
基于手机的交通出行感知

- 目前北京市中国移动手机用户数量高达**1700万**，约占总常住人口数的**75%**。
- 目前，北京市中国移动基站扇区总数超过**38,000个**，市区基站密度高达**44cells/km²**。
- 每日数据总量高达**12亿条**，占用空间约**70~110GB**。

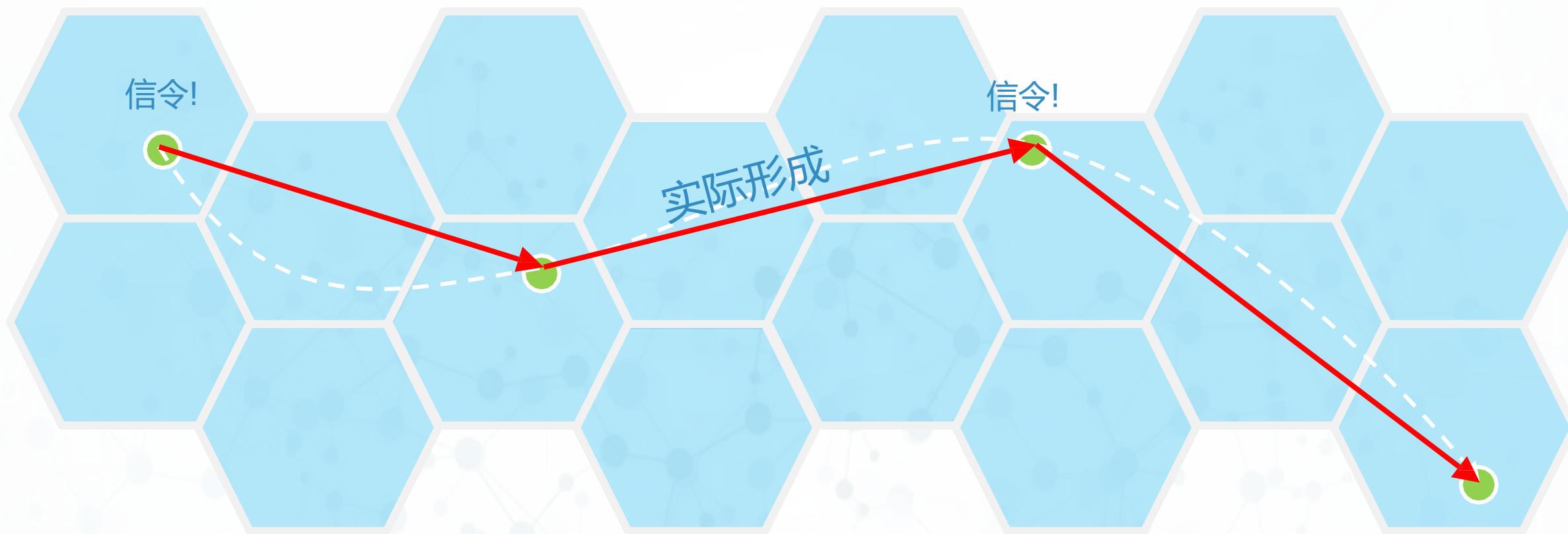


基于手机的交通出行感知

- 目前北京市中国移动手机用户数量高达**1700万**，约占总常住人口数的**75%**。
- 目前，北京市中国移动基站扇区总数超过**38,000个**，市区基站密度高达**44cells/km²**。
- 每日数据总量高达**12亿条**，占用空间约**70~110GB**。

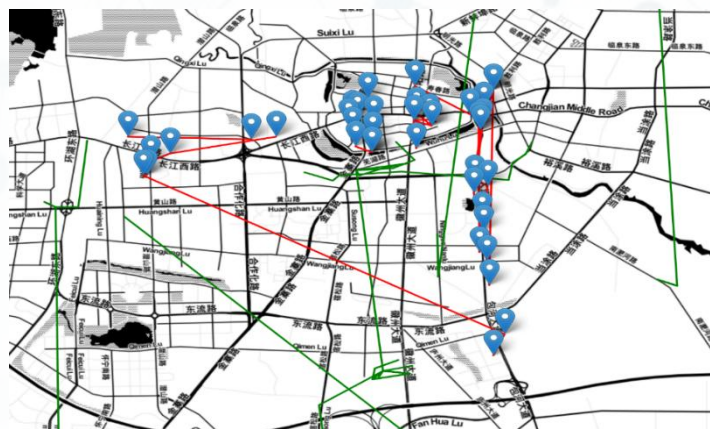
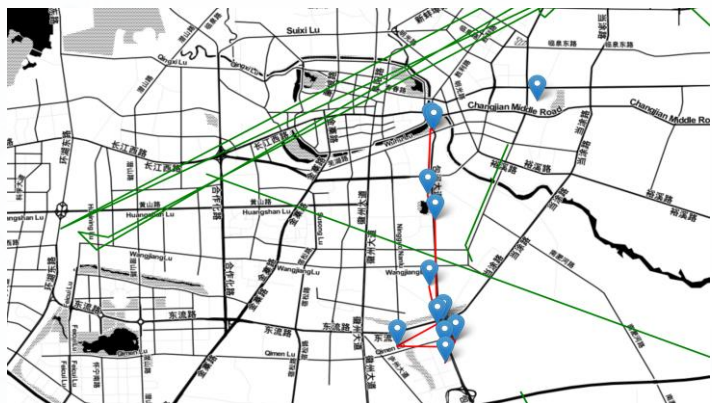


基于手机的技术原理



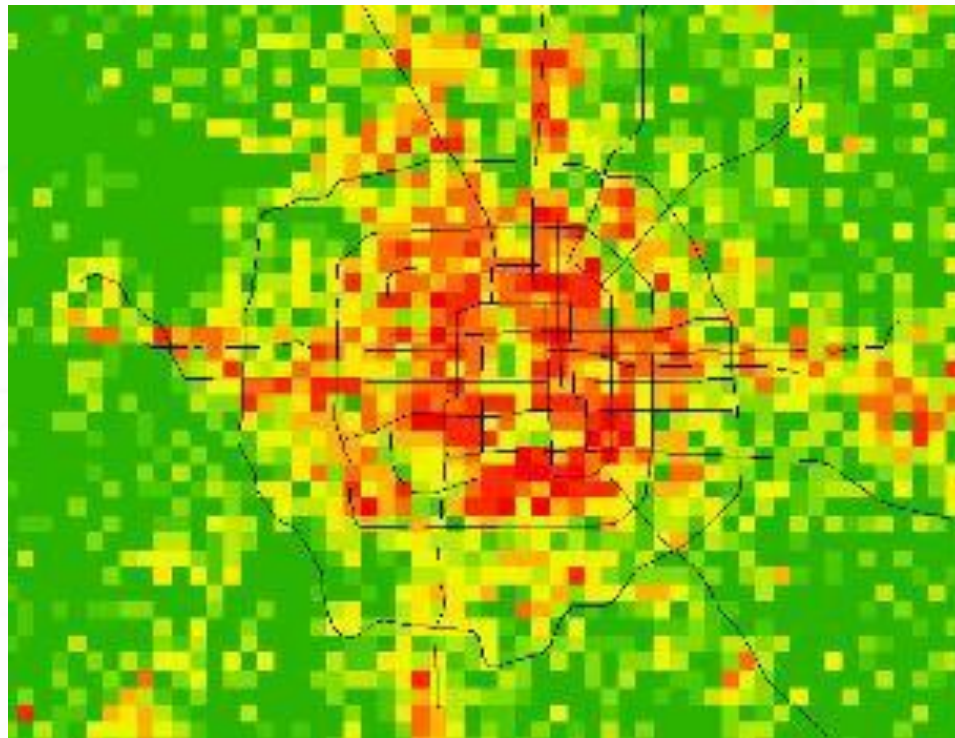
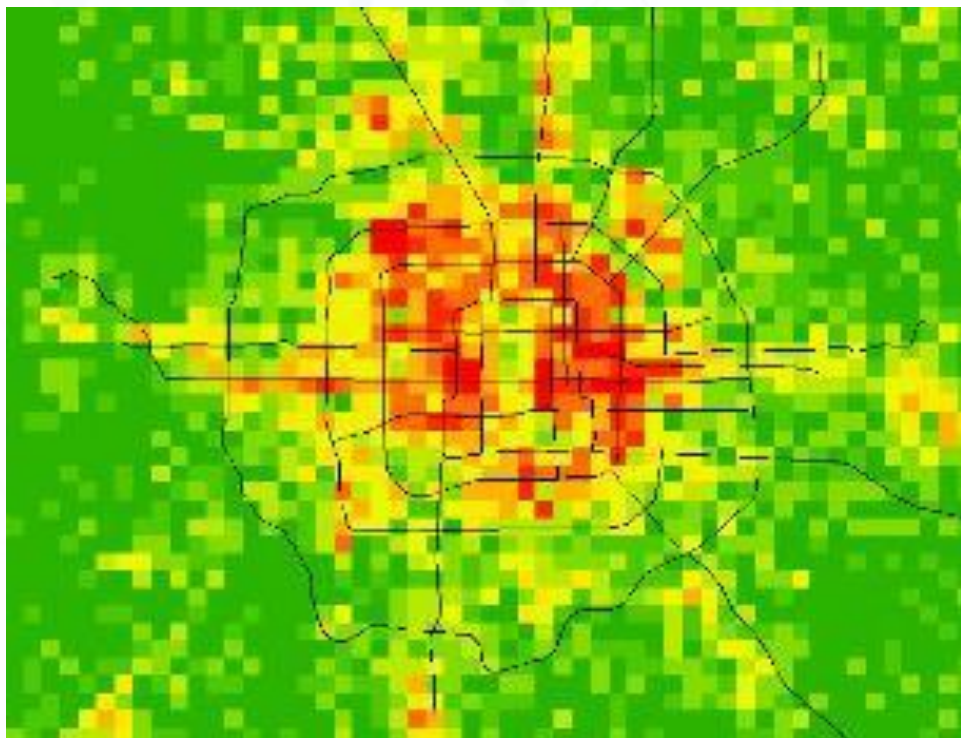
- 手机在日常使用中会与基站不断交互，即信令
- 主要的信令有7种，打电话、发短信、开机、关机、上网、2G/3G/4G切换、基站切换
- 即使没有任何行为，也会定期发送信令

出行类别分类（公交、地铁、私家车、出租车）

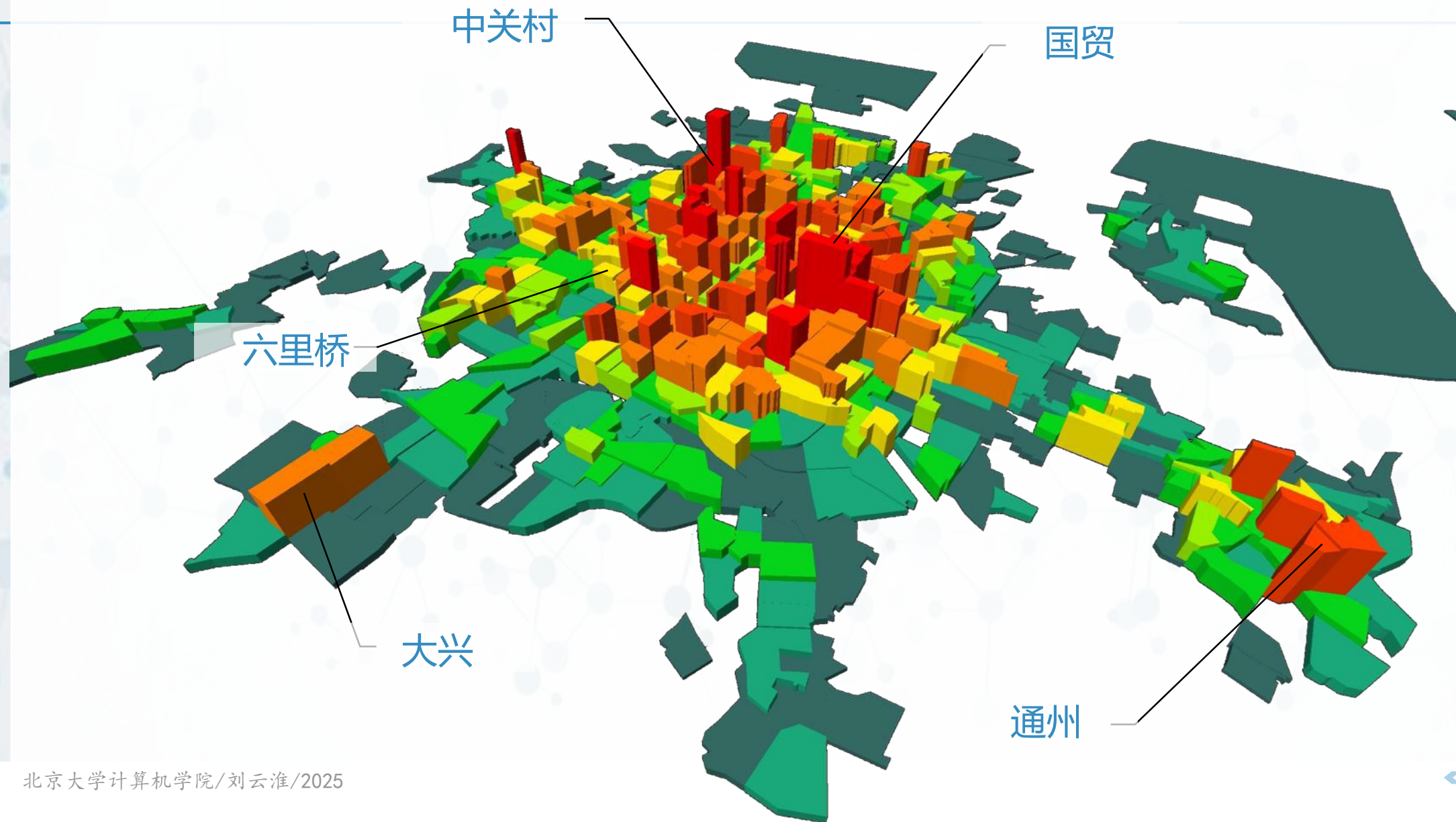


北京就业和居住分布

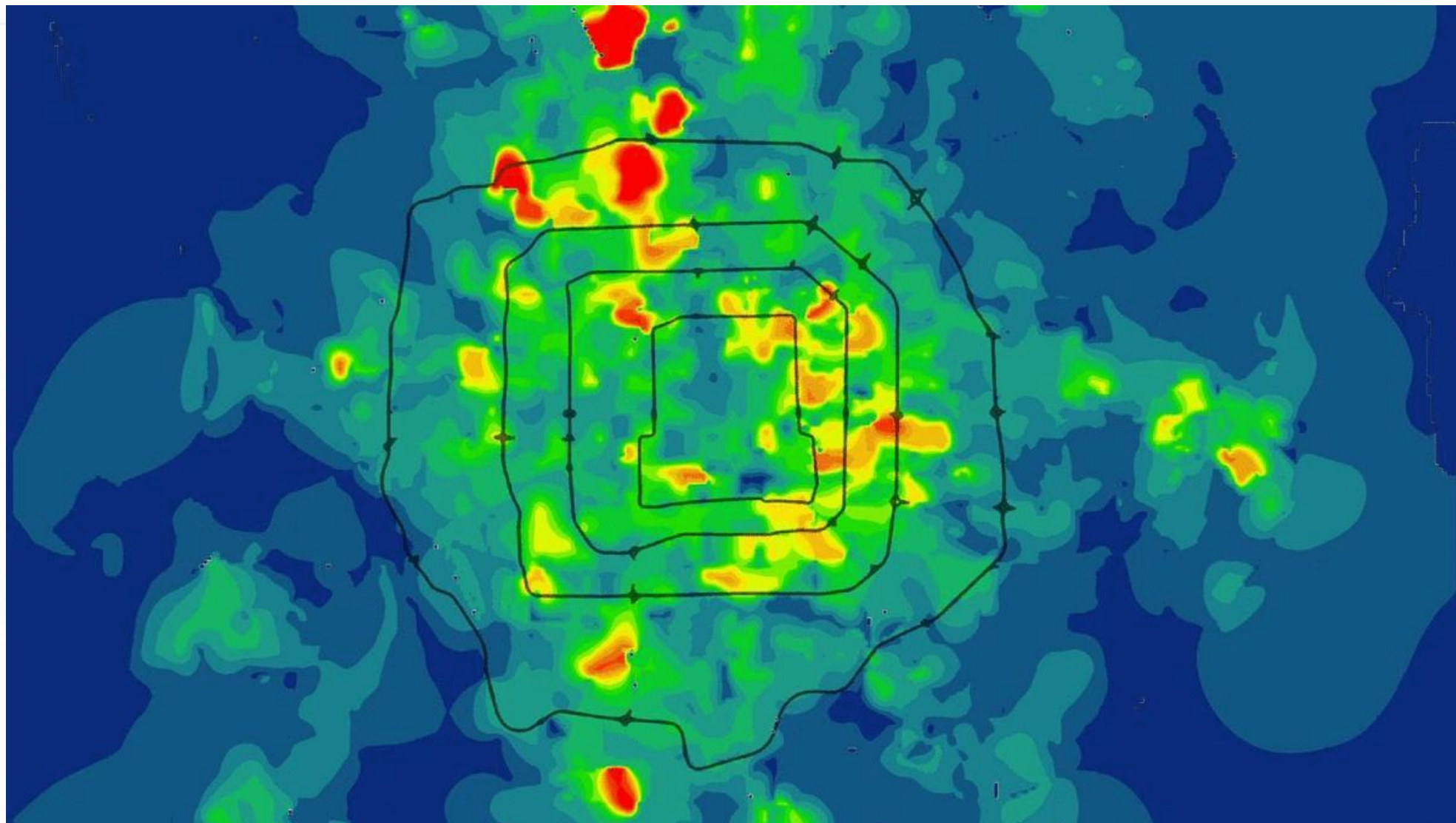
- › 根据信令行为，分析就业和居住的分布
- › 夜晚所在地为居住地，白天所在地为就业地



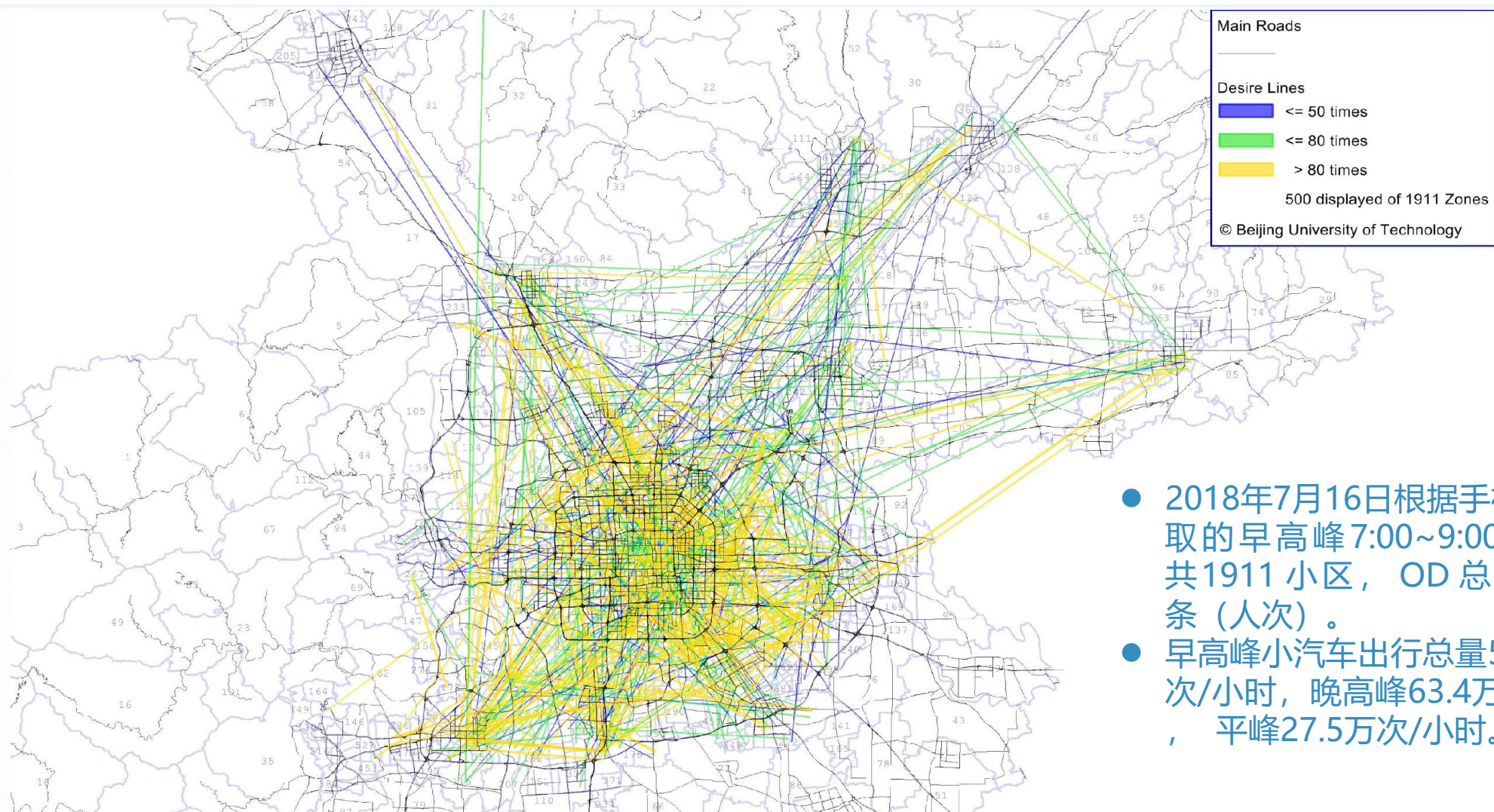
北京不同地区人口密度



北京地区不同时间人口分布

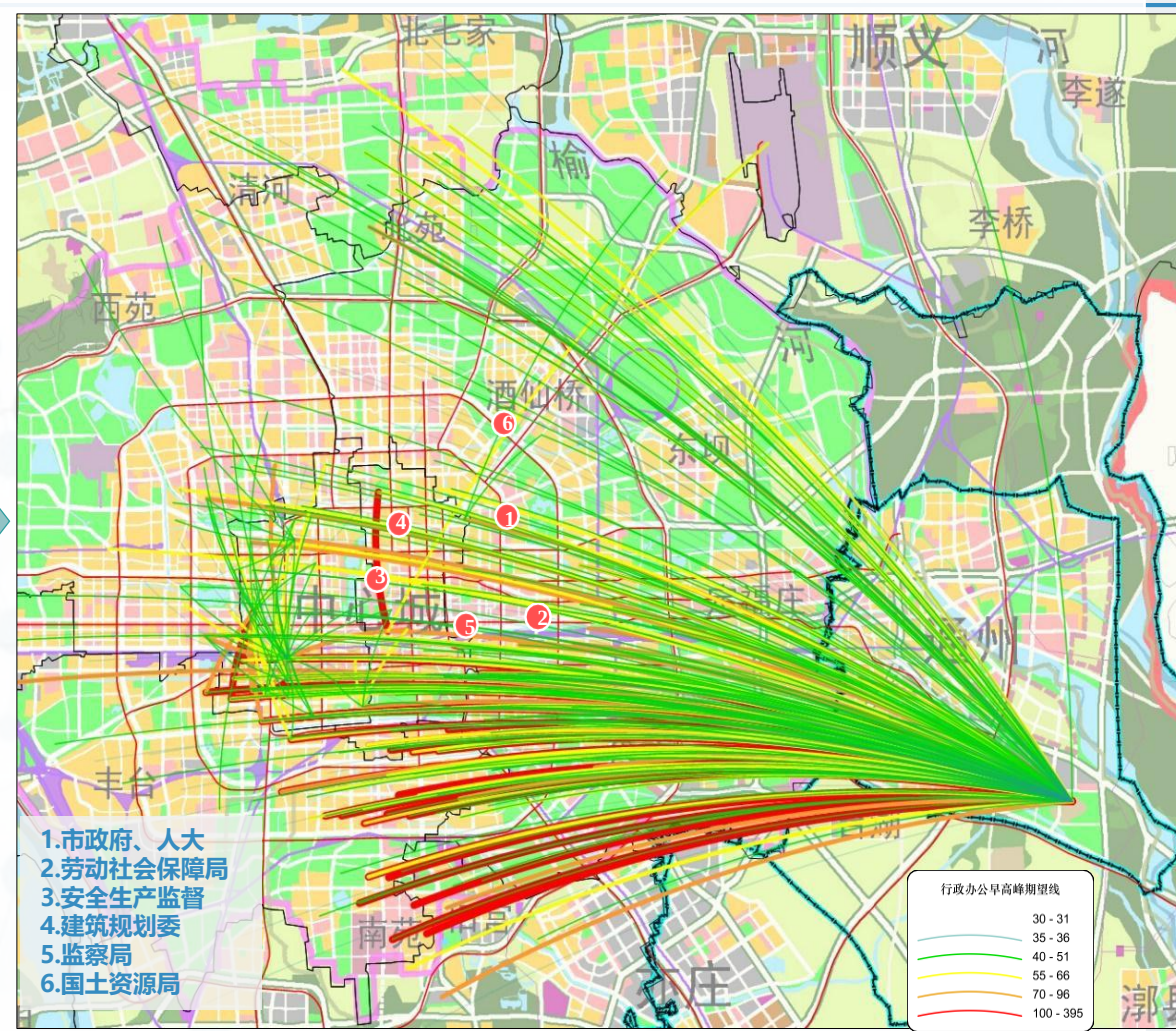
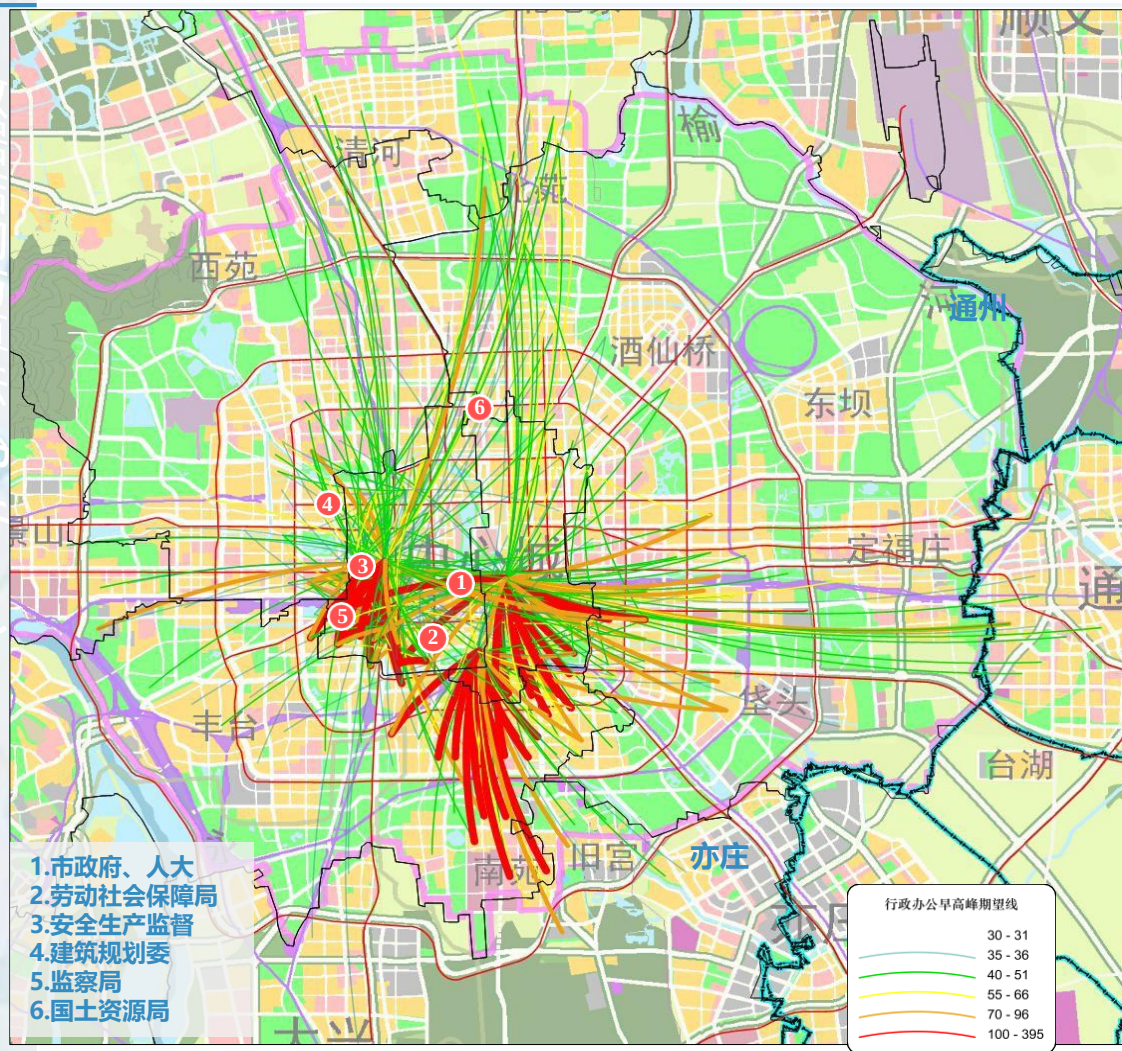


区域间出行

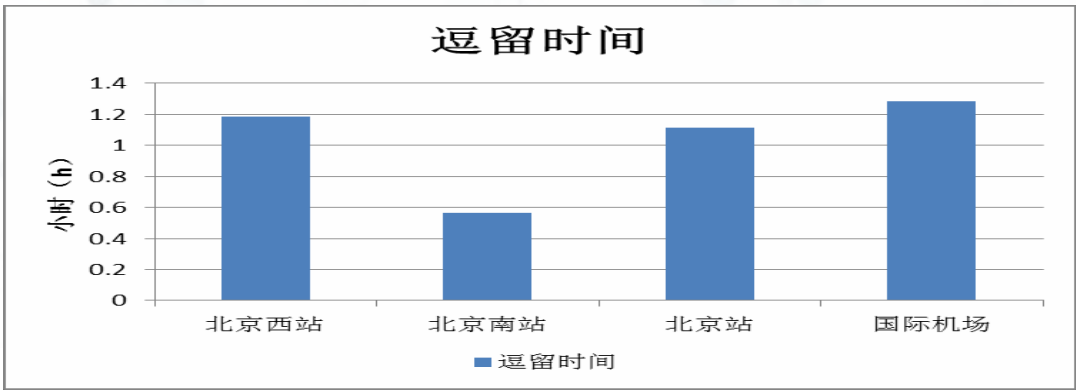
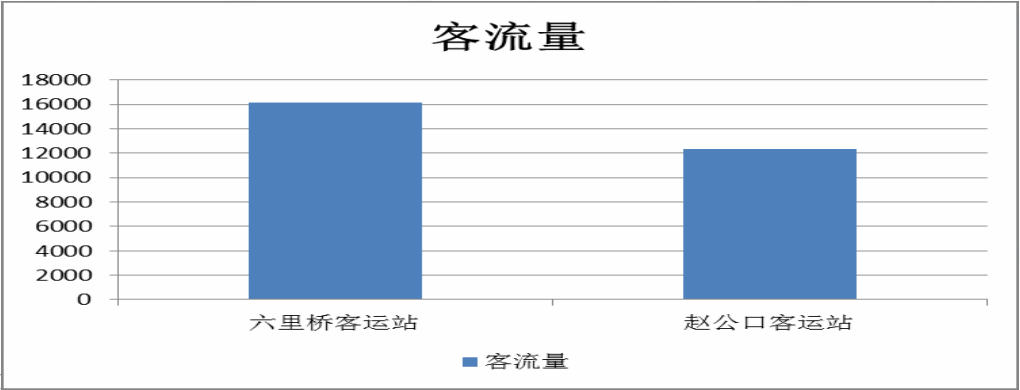
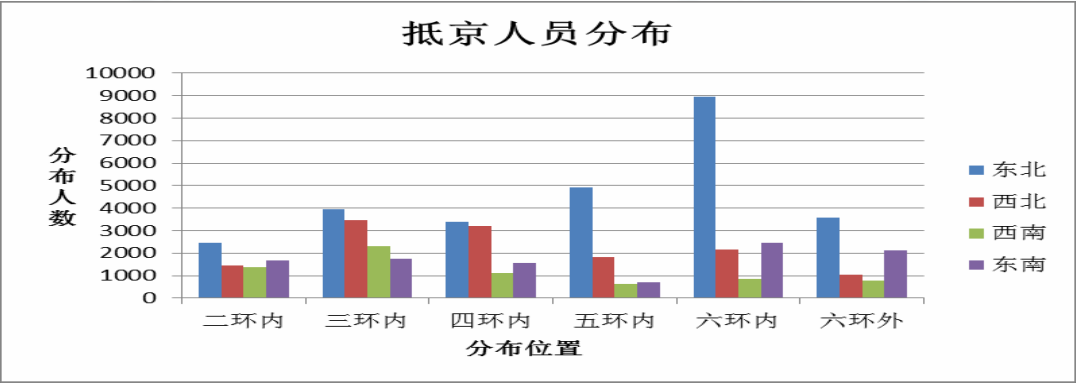
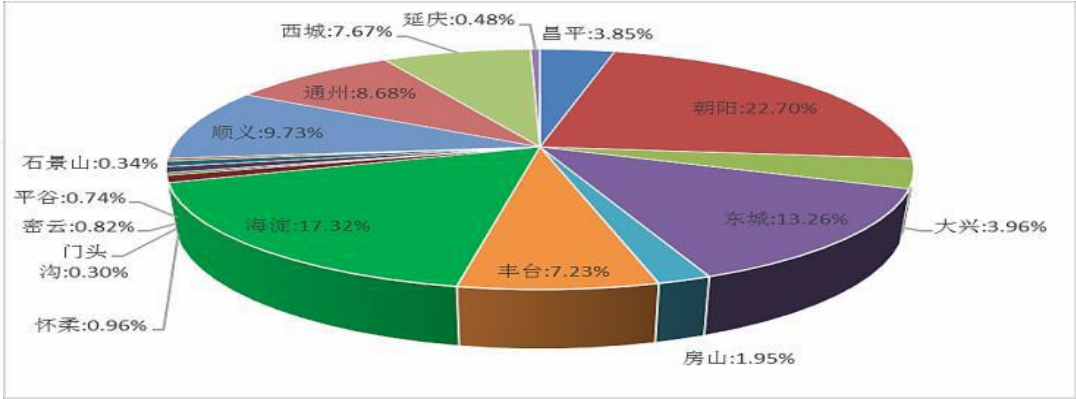
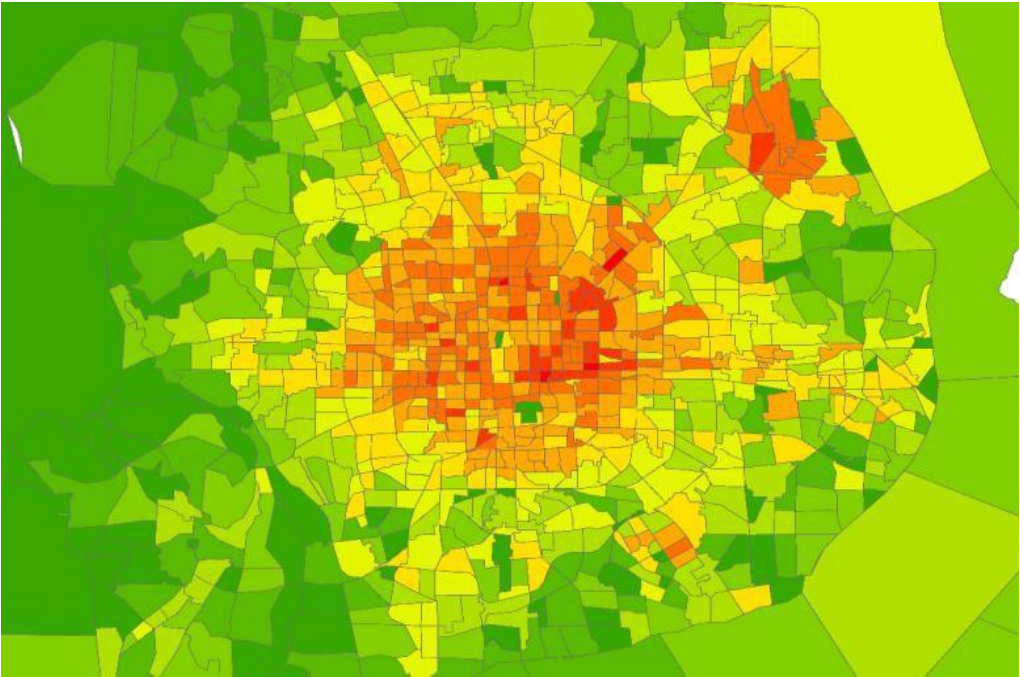


- 2018年7月16日根据手机数据获取的早高峰7:00~9:00数据。共1911小区，OD总量166万条（人次）。
- 早高峰小汽车出行总量51.8万辆次/小时，晚高峰63.4万次/小时，平峰27.5万次/小时。

应用案例：通州副中心对交通的影响



应用：第二机场选址论证





海关大数据典型案例

海关大数据的建设思路



一本底账

› 结构化数据记录1343.36亿条

来源于253个应用系统，其中：

- 内网应用系统162个
- 外网应用系统61个
- 交换应用系统29个
- 外购应用系统2个

› 非结构化数据

文件共计1.64亿个

大小为74.59TB

主要是报关单随附单证

统计司综合事务〔2018〕13号

海关总署公文处理专用纸

综合统计司关于发布《海关数据资产报告(2017年度)》的函

办公厅、国口办、改革办、政法司、研究室、关税司、监管司、加贸司、统计司、稽查司、缉私局、科技司、国际司、财务司、督审司、信息中心、数据中心：

根据《海关数据资源共享管理办法(试行)》(署科发〔2017〕184号)要求，为更好服务海关大数据应用，在科技司、信息中心和数据中心等相关司局的大力支持下，我司组织整理了《海关数据资产报告(2017年度)》(详见附件1、2)，现提供你司(局)，供工作参考。

附件：1. 海关数据资产报告(2017年度)

2. 数据资产报告附录

统计司

2018年2月28日

办理部门：统计司 承办人：刘寿吉 负责人：黄颂平 联系电话：4143

一本底账（元数据、数据资源目录）

序号	账号名称	使用部门	账号用途
1	crbrq_user	开发部	跨境限额限重
2	RSKMFT_R_GBASE	开发部	进出境货物运输工具风险作业子系统（陆路运输）
3	SJZL_META	治理处	元数据自动采集服务
4	cbec_read	开发部	跨境电子商务进口统一版数据上报子系统
5	cbec_user	开发部	跨境电子商务进口统一版数据上报子系统
6	emsmoniter_read	开发部	金关工程二期加工和保税管理系统
7	etl_loader	开发部	ETL数据加载
8	etl_schedule	开发部	税管中心项目中etl周期调度任务
9	irp_admin	开发部	信息资源共享服务平台
10	kjdstj	开发部	跨境电子商务进口统一版数据统计子系统
11	qlik_reader	开发部	qlikview
12	sjzl_loader	治理处	大数据平台数据加载
13	sjzl_tb	治理处	大数据平台数据同步
14	tmc_read	开发部	税管中心在线分析系统
15	tmc_rir	开发部	税管中心信息系统
16	tmc_write	开发部	税管中心信息系统
17	xyjx_user	开发部	金关工程二期企业信用管理绩效评估子系统
18	yth_read	开发部	税管中心全国通关一体化项目
19	irp_reader	开发部	信息资源共享服务平台
20	gbasebi_limit	开发部	稽查司业务BI

序号	用户名	显示名	序号	用户名	显示名
1	a0302	a0302	31	lishuo01	lishuo01
2	a0302_user	a0302_user	32	liuhaocheng	刘浩成
3	anquan_ali	安全日志测试阿里	33	liutao	刘韬
4	anquan_test	安全数据测试	34	menghao	孟昊
5	anquandata	安全数据	35	qmx	qmx
6	appana	海关大数据应用分析	36	qware_admin	qware_admin
7	bdcq_qyxyypg0702	企业信用评估模型完善升级(生产)	37	qyxyxxdc	企业信用信息调查子系统
8	bdcq_qyxyypg0702_dev	企业信用评估模型完善升级(开发)	38	qyxyxxdc_admin	企业信用信息调查子系统
9	bdcq_shk_jgl	上海跨境贸易管理(生产和测试)	39	rc01	初始化账号
10	cangdan	舱单风险测试	40	rptuser1	rptuser1
11	caojin	曹锦	41	rptuser2	rptuser2
12	ccic_yiruiting1	开发部伊瑞婷	42	s_jdp	数据底座建设
13	chencheng	陈晨	43	songke	宋克
14	chenzechao	陈泽钊	44	suijittest	suijittest
15	eciq	eciq	45	sunsenlin	孙森林
16	fengxin	冯鑫	46	tgyth	通关一体化
17	fkzxshfx	fkzxshfx	47	wangjian	汪健
18	haisou_admin	haisou_admin	48	wlfx_kf	wlfx_kf
19	hangzhouguan	hangzhouguan	49	wlfx_sc	物流风险生产
20	hanwei	韩伟	50	wudi	吴迪
21	hl2008_admin	HL2008_admin	51	wuliudz	wuliudz
22	hlwqb	互联网情报	52	xushaonan	徐少楠
23	houfangqing	侯芳清	53	yaoyinghong	姚颖虹
24	hulianwangqingbao	互联网情报系统	54	ydp	演示大屏
25	hzkj_admin	杭州跨境电子商务分析系统资源	55	zhanglei	张磊
26	jimin	金敏	56	zhangyun	zhangyun
27	jixiaofei01	季晓飞	57	zhaoxiao	zhaoxiao
28	kuajing_admin	kuajing_admin	58	zhengqingsyuan	厦门风险布控
29	kuajing_kaifa	kuajing_kaifa	59	zhouminjie	周敏捷
30	ligang	李刚			

编码: 名称: 是否落地: 是否有效:

	编码	名称	所属资源	所属用户
1	rmft_arrival_road_conta	公路运抵报告集装箱表	GBASE_gods	gods
2	rmft_arrival_road_list	公路运抵报告提运单表	GBASE_gods	gods
3	rmft_binding_road_conta	公路进出境承运确报集装箱表	GBASE_gods	gods
4	rmft_binding_road_seal	公路进出境承运确报封志表	GBASE_gods	gods
5	rmft_binding_road_transport	公路进出境承运确报运输工具表	GBASE_gods	gods
6	rmft_binding_transport_tray	公路进出境承运确报运输工具...	GBASE_gods	gods
7	rmft_change_off_conta_seal	公路落装改配集装箱封志表	GBASE_gods	gods
8	rmft_change_off_road_conta	公路落装改配集装箱表	GBASE_gods	gods

详细信息

基本信息		结构		物化信息		显示数据	
序号...	字段	名称	类型	数据物理类型	长度	所属维度	
1	manifest_id	货物运输批次号	字符串	varchar(128)	128		
1	manifest_id	货物运输批次号	字符串	varchar(128)	128		
2	i_e_flag	进出境标志	字符串	varchar(1)	1		
2	i_e_flag	进出境标志	字符串	varchar(1)	1		
3	decl_traf_mode	运输方式代码	字符串	varchar(4)	4		
3	decl_traf_mode	运输方式代码	字符串	varchar(4)	4		
4	customs_code	进出境海关代码	字符串	varchar(4)	4		
4	customs_code	进出境海关代码	字符串	varchar(4)	4		

关闭

一份数据（内部）

- 内部业务系统共完成采集**113个系统**
- 共计**4900多张表**

	业务系统中文名	前缀	抽取表数
1	H2010通关系统（58个子系统）	无前缀	857
59	e-CIQ主干系统	ECIQ	1067
60	邮递物品管理系统	POST	107
61	CIQ2000数据	CIQ	491
62	财务系统	CW	152
63	ATA	ATA_或 LPS	18
64	新舱单	MFT_	85
65	公路舱单	RMFT_	38
66	行邮物品管理系统	NCAD_	37
67	行邮旅客管理系统	PGS_	84
68	行邮舱单管理系统	PMS_	35
69	运输工具	YSGJ_	127
70	金关工程二期多式联运管理系统	MTM_	96
71	金关工程二期智能卡口管理系统1.0版	RTR_	14
72	免税店免税商品管理系统	DUTY_	95
73	外交外商常驻机构公自用物品进出境监管系统	LPS_	18
74	新快件	EXP_	48
75	金关工程二期监管场所管理系统	CSA_	64
76	海关通关作业辅助系统	HP_	22
77	企业信用调查子系统	JC_E_	64
78	企业信用调查子系统	JC_M_	279
79	企业信用调查子系统	JC_W_	1
80	缉私	JS_	184

	业务系统中文名	前缀	抽取表数
81	跨境进口全国统一版系统	CBEC_	2
82	跨境进口全国统一版系统	CBEC_	21
83	跨境进口全国统一版系统	CBEC_	11
84	估价系统	PR或PRL	141
85	涉案财物	FM_	65
86	税管中心作业平台	TMC_	11
87	选查离线三期	CCTS3_	62
88	中国海关公式定价进口货物备案管理系统	FMLA_	67
89	风险分析作业舱单风险分析子系统（在线）	MRA_	1
90	通关系统-船舶吨税电子文件试点项目	CUST_	30
91	新综平(交互协调平台)	HG_	19
92	小船系统	XC_	2
93	查验管理二期	RSK2_	4
94	跨境出口地方版	OCBEC_	77
95	跨境出口全国统一版	CBEE_	65
96	税管2.0模型库	TMC_MA_	4
97	税管2.0风险研判库	RIR_	4
98	税管2.0条码库	BC_	3
99	税管中心风控管理平台	RM_	2
100	商品库	GOODS_	23
101	进出口商品归类系统	CLASSIFY_	10
102	H2000 APP	H2K_	33
103	知识产权	ZSCQ_	7
104	海关票据管理系统	BM_	16
105	海关行政复议管理系统V2.0	XZFY_	20
106	海关行政诉讼案件系统	XZSS_	40
107	原产地管理系统	OMIS_	13
108	中国海关实验室信息管理系统（LIMS）全国联网项目	LIMS_	17
109	风险分析作业舱单风险分析子系统（离线）	CCTSM_	35
110	进出口商品规范申报管理系统	GFSB_	7
111	金关工程二期报关单批量复审系统	PLFS_	39
112	海关综合业务管理平台	无前缀	16
113	关务保障系统二期	GB2	596

一份数据（外部）

› 商务部、外汇管理局、工商总局、税务总局、国家信息中心**5类外部数据**

› 共计19张表6.16亿条数据

英文表名	中文表名	记录数
市场监管总局企业登记注册数据		
GSZJ_NEWENTINFO	企业注册信息表	95415502
GSZJ_REVENTINFO	吊销企业表	624167
GSZJ_CALENTINFO	注销企业表	1100474
外汇管理局数据		
SAFE_QRY_LOG	外汇局查询日志	7146915
SAFE_DPSIT_BSC	外汇局存款表	82218
SAFE_DRAW_BSC	外汇局取款表	182041
SAFE_TRADE_PFCURR_BSC	外汇局贸易付汇表	11833326
SAFE_TRADE_RCVECH_BSC	外汇局贸易收汇表	40097757
SAFE_TRADE_FEC_BRKRUL_BSC	外汇局贸易外汇违规表	665
SAFE_TRADE_FEC_EXPDT_DT	外汇局贸易外汇支出明细	2218750
税务总局数据		
RTX_REFUND_LIST	启运港已退税报关单	318025336
国家信息中心数据		
PUB_PENALTY	行政处罚双公示信息表	1100474
PUB_PERMISSION	行政许可双公示信息表	95415502
TB_SXBZXRMD	失信被执行人名单	624167
TB_YCJYML	异常经营企业名录	16639954

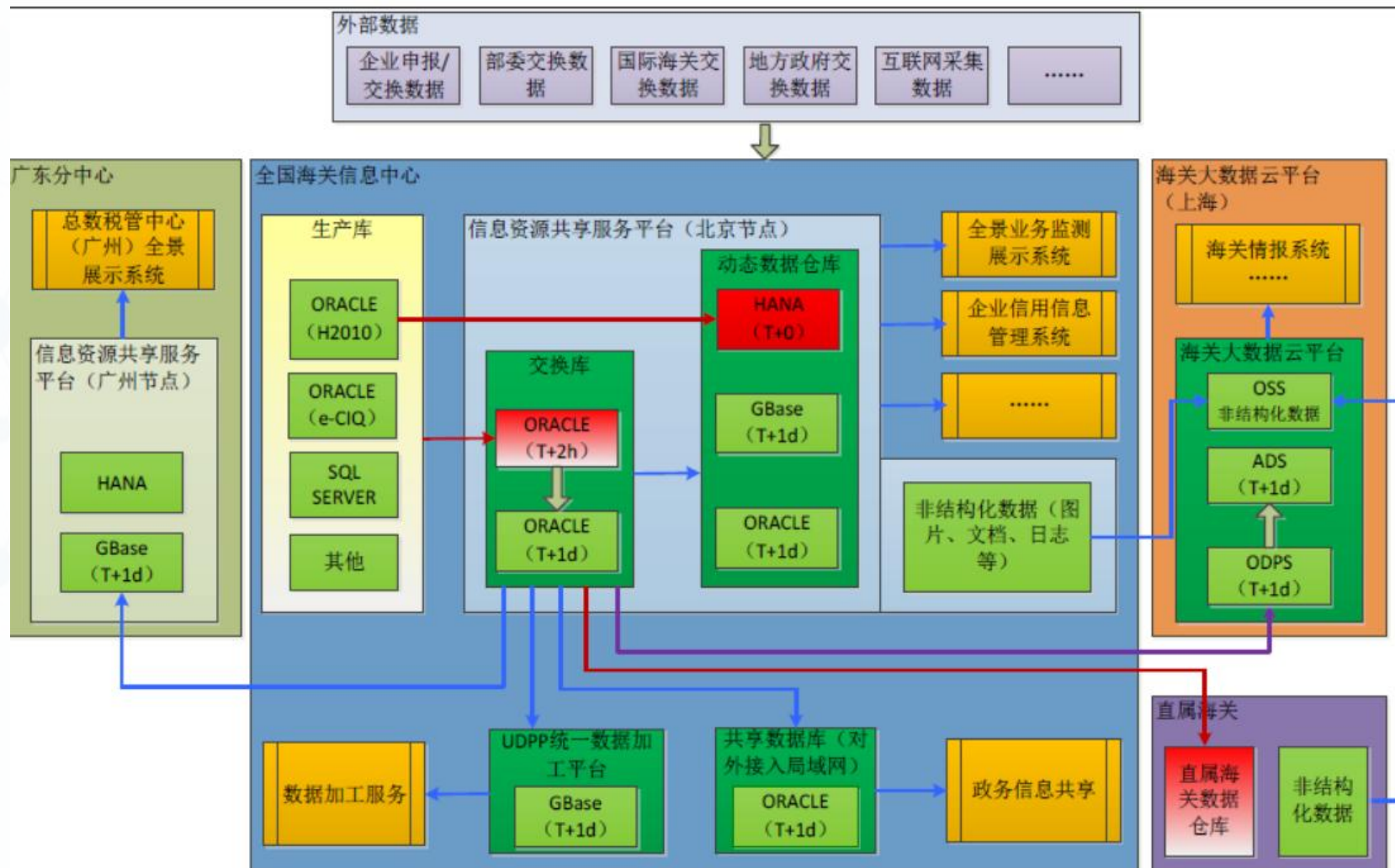
一份数据（商业）

- › 完成BVD、亿海蓝等2类商业采购数据的采集工作
- › 包括41张表共77.06亿余条数据记录，使用空间约为669.50TB

序号	数据表物理名称	数据表中文名称	数据记录条数
1	yhl_dynamic	船舶轨迹静态信息	5,772,289,222
2	yhl_static_ship	船舶基本信息表	327,857,869
3	yhl_reachport	船舶挂靠港口数据表	8,721,045
4	yhl_ihsports	港口数据表	6,252
5	bvd_financial_strength_more	活动	74,226,857
6	bvd_stock_exchanges_and_indexes	所有地址	81,135
7	bvd_dmc-current_only	所有当前股东的第一层级	65,710,083
8	bvd_dmc_previous	所有子公司的第一层级	42,142,172
9	bvd_all_current_shareholders_first_level	当前审计	43,538,538
10	bvd_all_subsidiaries_first_level	之前银行经理	4,045,606
11	bvd_basic_shareholder_info	当前银行经理	15,666,018
12	bvd_beneficial_owners_10_10	基本股东信息	40,066,956
13	bvd_branches	最终受益人10-10	5,783,369
14	bvd_controlling_shareholders	黑名单实体	15,525,504
15	bvd_additional_company_info	实体地址	18,446,544
16	bvd_all_addresses	实体进入类型（个人/组织）	102,439,892
17	bvd_bvd9	为了将黑名单数据与企业信息数据中的企业对应起来	92,266,277
18	bvd_contact_info	实体子分类	92,266,287
19	bvd_identifiers	国际/国家/省（州）	160,988,738
20	bvd_legal_info	黑名单中实体间的关系	92,266,287
21	bvd_industry_classifications	实体关系定义	248,694,695

一个平台

大数据平台数据架构

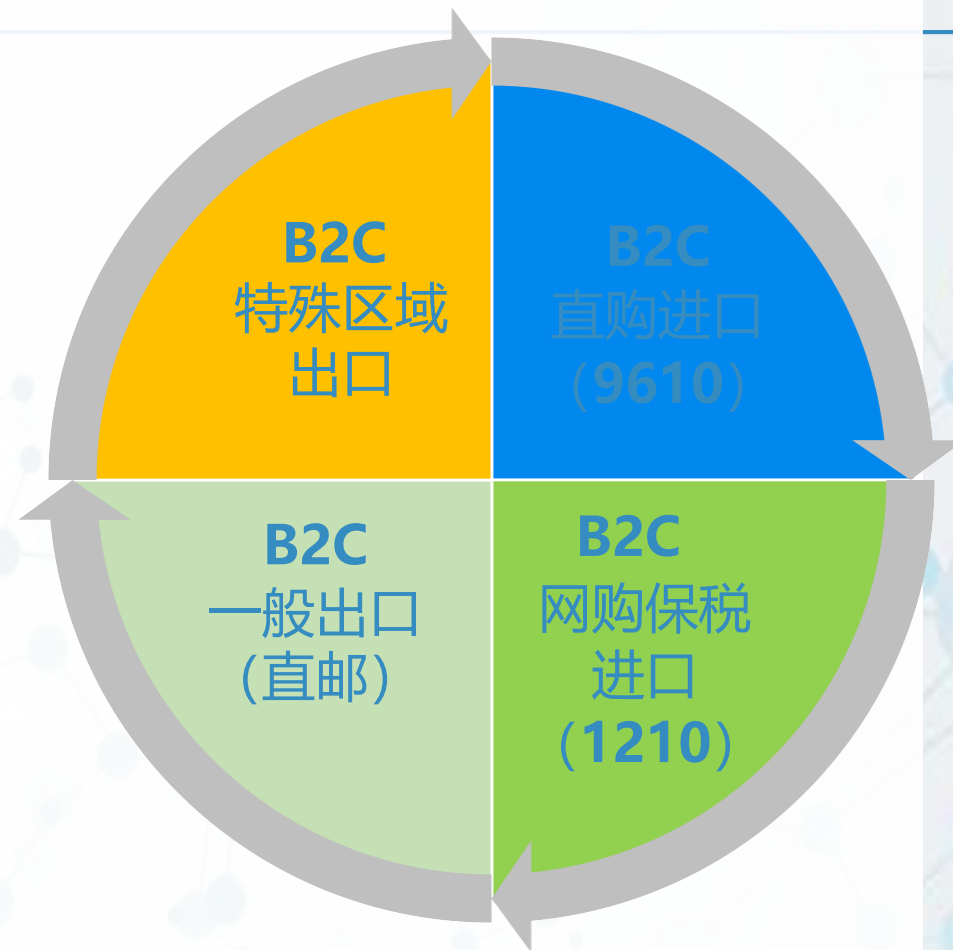


一套模型



跨境电商

- › 中国境内消费者通过跨境电商第三方平台 **自境外购买**
“网购保税进口”、“直购进口”运递进境
- › 境外消费者通过跨境电商第三方平台 **自境内购买**
“特殊区域出口”“一般出口”运递出境
- › 试点区域
上海、杭州、宁波、郑州、重庆、广州、深圳、天津、福州、平潭、合肥、成都、大连、青岛、苏州等**15个试点城市**
在杭州、天津、上海、重庆、合肥、郑州、广州、成都、大连、宁波、青岛、深圳、苏州、北京、呼和浩特、沈阳、长春、哈尔滨、南京、南昌、武汉、长沙、南宁、海口、贵阳、昆明、西安、兰州、厦门、唐山、无锡、威海、珠海、东莞、义乌等**35个城市综合试验区**



跨境电商重要意义

我国跨境电子商务呈现井喷式增长势头。据统计，2018年通过海关跨境电商平台零售进出口**商品总额1347亿元，增长50%**，其中**进口785.8亿元，增长39.8%**。

年份	2017	2018	2019（上半年）
规模			
进口跨境电商交易额（亿元）	565.9	785.8	456.5
进口网购保税交易清单数（亿）	2.2	3.59	2.03
进口直购进口交易清单数（亿）	0.39	0.54	0.23
进口跨境电商交易清单数（合计，亿）	2.59	4.13	2.26

跨境电商模型概述

安全准入 风险防控子模型

针对**涉爆、涉毒、涉濒危**等安全准入风险，从企业、人员、商品、交易行为等4个维度构建风险特征，利用机器学习技术，实现跨境电商渠道安全准入高风险人员、企业、商品的识别，并对接风险作业系统实现自动识别下达**布控规则**。

虚假交易 风险防控子模型

对交易信息、交易特征、购买偏好进行挖掘，通过申报清单、订单、支付单、物流运单等“四单”中姓名、身份证件、地址、电话等、价格相关性数据，甄别**异常交易行为**，甄别存在风险的人员身份证号、地址及联系电话、以及跨境电商企业，可以对实时申报清单进行风险甄别，也可以推送风险信息实施人工**布控**。

模型目标

› 现有无干预随机布控

布控率 $<1\%$ （每100单核查一单）

查获率 $<1\%$ （每100个核查中，有一个查获）

综合查获率 $<0.01\%$ （每10000单一个查获）

› 模型建设目标

布控率 $<1\%$ （每100单核查一单）

查获率 $>20\%$ （即每5单核查，有一个查获）

综合查获率 $>0.2\%$ （每500单查获一单）

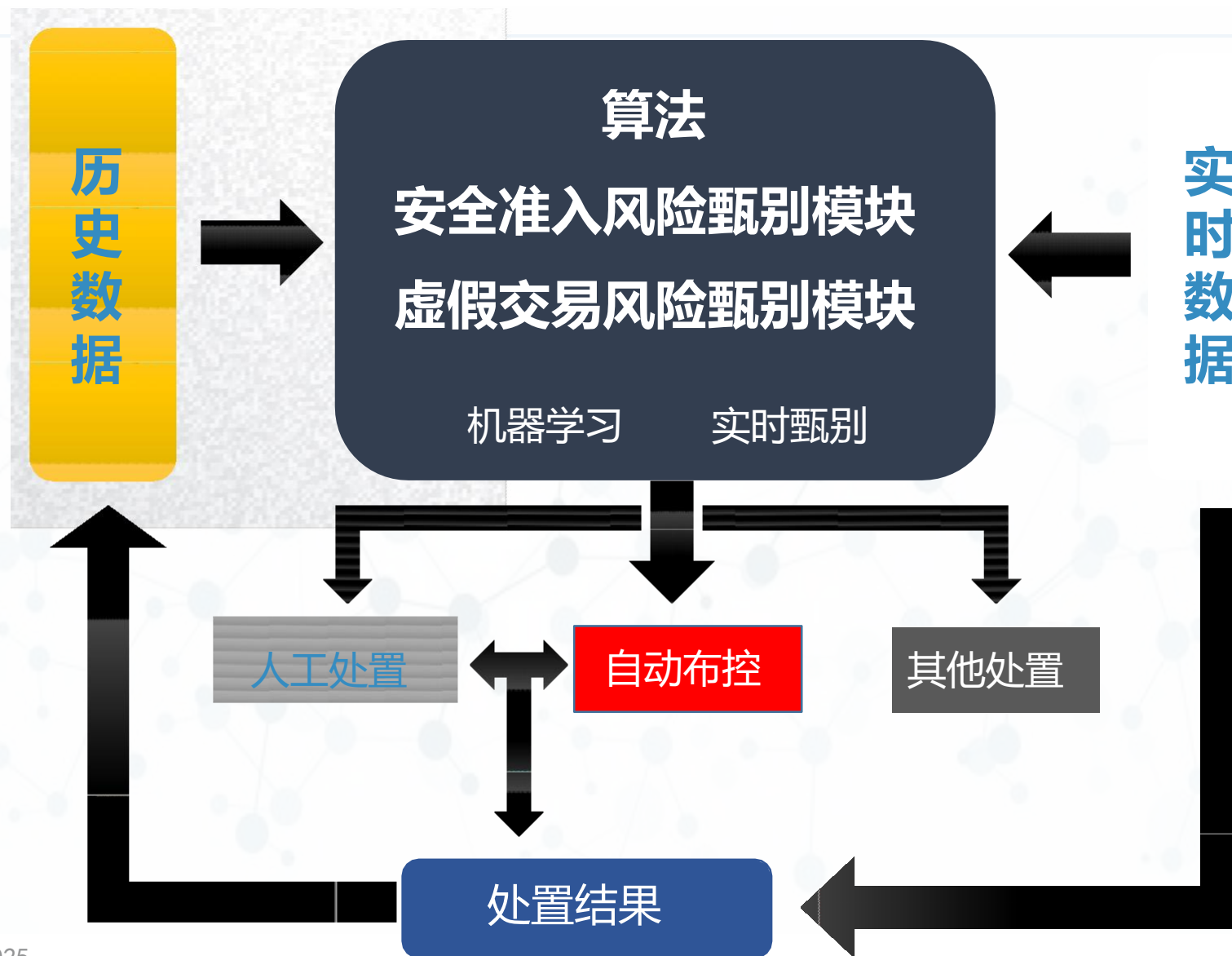
工作效率提升20倍以上

数据资源

› 清单、订单、运单、支付单数据6表126个字段

使用数据库	数据来源	子表数
跨境电商进口统一版系统（核心数据）	云平台	8
工商总局企业登记注册数据	云平台	4
海关缉私信息综合应用系统（行政案件数据）	云平台	4
海关缉私信息综合应用系统（刑事案件数据）	云平台	4
失信企业名单	云平台	3
企业画像评分表	云平台	1
国家信息中心企业信用信息共享服务平台	云平台	4

工作模式



建模思路



- **特征抽取**
 - 基础特征、经验特征、统计特征
- **特征处理**
 - 异常值处理、缺失值处理、数值型特征除偏、分类型特征编码、重采样、欠采样
- **特征选择**
 - 卡方检验、方差分析、相关性分析、方差阈值筛选法、嵌入式选择
- **模型设计**
 - 虚假交易模型：预训练 + 微调
 - 安全准入模型：无监督模型 + 有监督模型
- **效果评估**

特征提取

- **基础特征（79个）**
 - 清单、订单、运单、支付单等包含的基础信息
- **经验特征（42个）**
 - 借鉴海关业务经验，对特征进行抽取
- **统计特征（27个）**

设计思路：捕捉样本中的异常信息，含有更多异常的样本或更具风险性

基础特征（原始数据表项）

特征	计算逻辑	数据来源
企业性质	企业类型	可国有控股、混合所有制等、外资、民营、其他
注册资本	注册资本	十万级及以下、百万级、千万级、亿级
注册日期	注册日期	1-2年、3-5年、6-10年、10年以上
注销日期	注销时间是否在业务发生时间之前	注销日期
吊销日期	吊销时间是否在业务发生时间之前	吊销日期
经营场所	是否为一线城市	经营场所
注册地	是否为一线城市	注册地
海关信誉评级	级别	高级认证企业、一般认证企业、一般信用企业、失信企业
失信企业名单	是否在失信企业名单内	处罚对象名称

经验特征（根据历史经验总结的特征）

特征	计算逻辑	数据来源
周期内同一ID使用次数	计算同个订购人ID周期内订单数量	订购人证件号码、清单编号、电商企业代码
周期内同一ID对应订单金额分布	计算同个订购人ID周期内订单总金额及各订单金额分布	订购人证件号码、清单编号、订单金额
周期内同一ID使用的联系号码数量	同个订购人ID对应收件人电话计数	收件人电话号码、订购人身份证号码、电商企业代码
周期内同一ID对应收件人姓名数量	同个订购人ID对应收件人姓名计数	收件人姓名计数、订购人身份证号码、电商企业代码、
周期内同一ID对应收件地址的数量	同个订购人ID对应收件地址计数(到省市区)	收件地址、订购人身份证号码、电商企业代码
周期内同一ID对应收件地址与ID所属地的差异	订购人ID前6位，与收货地址（到省市区）比对	订购人证件号码、收件地址、电商企业代码
周期内同一ID对应电商（非大型电商）的数量	同个订购人ID对应电商企业计数	订购人证件号码、电商企业代码
周期内同一ID对应物流的数量	同个订购人ID对应物流企业数计数	物流企业代码、订购人证件号码、电商企业代码
周期内同一ID对应支付的占比	同个订购人ID对应支付企业数计数，计算每个支付企业 支付金额除以支付总金额	支付企业代码、订购人证件号码、电商企业代码、支付 金额

经验特征（续）

特征	计算逻辑	数据来源
进口收件人是否有发生过安全准入情事	收件人身份证号是否在安全准入身份证列表内、是否在 缉私案件信息库内	收件人身份证号码
进口联系电话是否有发生过安全准入情事	进口收件人电话是否在安全准入电话列表内、是否在缉 私案件信息库内	收件人电话
进口商品是否安全准入枪爆影子商品	优先级：起运国（直购）或原产国（保税）+税号+品 名、起运国（直购）或原产国（保税）+品名、品名	起运国（直购）或原产国（保税）、税号、品名
进口商品是否安全准入涉毒影子商品	优先级：起运国（直购）或原产国（保税）+税号+品 名、起运国（直购）或原产国（保税）+品名、品名	起运国（直购）或原产国（保税）、税号、品名
进口商品是否安全准入涉濒危影子商品	优先级：起运国（直购）或原产国（保税）+税号+品 名、起运国（直购）或原产国（保税）+品名、品名	起运国（直购）或原产国（保税）、税号、品名
跨境直购进口商品是否安全准入属日本核辐射区食品	优先级：起运国+税号+品名、起运国+品名、品名	起运国、税号、品名
进口商品是否属于禁限宣传品影子商品名录	优先级：起运国（直购）或原产国（保税）+税号+品 名、起运国（直购）或原产国（保税）+品名、品名	起运国（直购）或原产国（保税）、税号、品名
商品品名、规格型号出现麻醉品精神药品特征词	商品品名、规格型号与名录名词比对	品名、型号
商品品名、规格型号出现两用物项特征词	商品品名、规格型号与名录名词比对	品名、型号

统计特征 (27个)

- 样本异常特征数目
 - 基于Z-score, 对特征的异常性进行判断; 针对样本所含的所有特征, 统计异常特征个数
- 样本清单物品相对于同时期同类物品的差异信息
 - 样本清单物品价格 (或质量等) 相对于过去15天同类物品的平均价格(或质量等)差异
- 字段历史发生问题占比
 - 某字段历史发生问题的比例

特征处理

› 异常值处理

对于同一清单对应正负两个标签的情况，对该样本进行舍弃
考虑到离群值或许包含有效信息，故未对离群值进行处理

› 缺失值处理

根据特征含义与分布情况，对于缺失字段填充0值、极大值、极小值或均值
针对不同特征，具体的填充方式经实验确定

› 分类型特征编码

编码原因：部分分类型特征类型过多，且每类样本数较少，模型对类别难以进行有效学习

编码方法：将出现次数较少的类别合并为“其它”大类

Catboost不需要对分类型特征进行编码

特征选择

› 卡方检验

卡方检验：统计样本的实际观测值与理论推断值之间的偏离程度，以判断某分类型变量对另一分类样本造成影响
针对分类型特征两两进行卡方检验，明确分类型特征之间的相关性

› 方差分析

方差分析：对多个样本总体的均值进行检验，以研究某分类变量的不同水平是否对观测数值变量产生显著影响
针对分类型特征和数值型特征进行方差分析，分析特征之间的相互影响关系

› 相关性分析

相关性分析：利用皮尔逊相关性系数，分析特征的相关性
针对数值型特征两两进行相关性分析，找与标签强相关的特征，用于扩充样本

特征选择

› 方差阈值筛选法

针对数值型特征计算其方差，剔除方差为0的特征

› 嵌入式选择

将特征选择和学习器结合，在学习器训练过程中自动进行特征选择

利用特征重要性与shape-value对特征进行分析理解

综合以上特征选择方法

- 虚假交易特征维度由**148**维降低为**42**维
- 安全准入特征维度由**148**维降低为**49**维

特征选择

- 虚假交易模型保留的特征为：

特征	含义
支付时间倒挂	直购进口模式下，计算支付单支付时间与进境时间的差值
商品编码评分	该商品编码历史黑样本中出现次数/该商品编码历史出现总次数
电商企业代码	电商企业的编码
毛重（kg）	货物连同包装的重量
净重（kg）	货物本身的重量
物品数量	该清单包含物品的数目
涉毒风险	根据涉毒影子商品库，判断该清单内的商品是否有涉毒风险
商品总价	该清单全部商品的价格

虚假交易模型部分特征表

特征选择

- 安全准入模型保留的特征为：

特征	含义
异常特征数量	清单中异常特征的数量
商品编码评分	该商品编码历史黑样本中出现次数/该商品编码历史出现总次数
包装种类	包装的种类，包含木箱、纸箱、桶装等
税款总额	清单商品收的税款总额
支付时间倒挂	直购进口模式下，计算支付单支付时间与进境时间的差值
实际支付金额	清单商品的货款金额+税款总额-折扣
毛重（kg）	货物连同它的包装的重量
货款金额	货款的总金额

安全准入模型部分特征表

模型设计

针对虚假交易和安全准入问题的特点，在基本模型的基础上，我们分别选择不同的训练学习策略，以实现精准预测。

› 基本模型

树形模型

- 优点：树形模型可以较好地刻画特征之间的非线性关系，模型的可解释性较好
- 尝试模型：CatBoost, LightGBM, XGBoost

线性模型

- 优点：线性模型简单、参数量较少，对于线性可分数据有较好地表现效果
- 尝试模型：Lasso, SVM

深度学习模型

- 优点：具有针对大样本集的强拟合能力
- 尝试模型：样本数量少的情况不适用

效果评估

评价指标

准确度 Precision: $TP / (TP + FP)$: 认为正常的中, 有多少是真正正常的

召回率 Recall: $TP / (TP + FN)$: 正常的中, 有多少被正确识别了

F1: $2 * Precision * Recall / (Precision + Recall)$

AUC: ROC面积, 在0.5-1.0之间

举例: 100单交易, 50单正常, 50单虚假的

某虚假识别模型	识别出60单正常的, 其中40个真的是正常, 20个其实是虚假的 识别出40单虚假的, 其中30个真的是虚假的, 10个其实是正常的
	准确度: $40/60=0.66$ 召回率: $40/50=0.80$ F1=0.72
另一模型	识别出80单正常的, 其中50个真的是正常, 30个其实是虚假的 识别出20单虚假的, 全部都真的是虚假的
	准确度: $50/80=0.625$ 召回率: $50/50=1$ F1=0.769

效果评估

虚假交易模型

Precision	Recall	F1	AUC
0.776	0.923	0.843	0.989

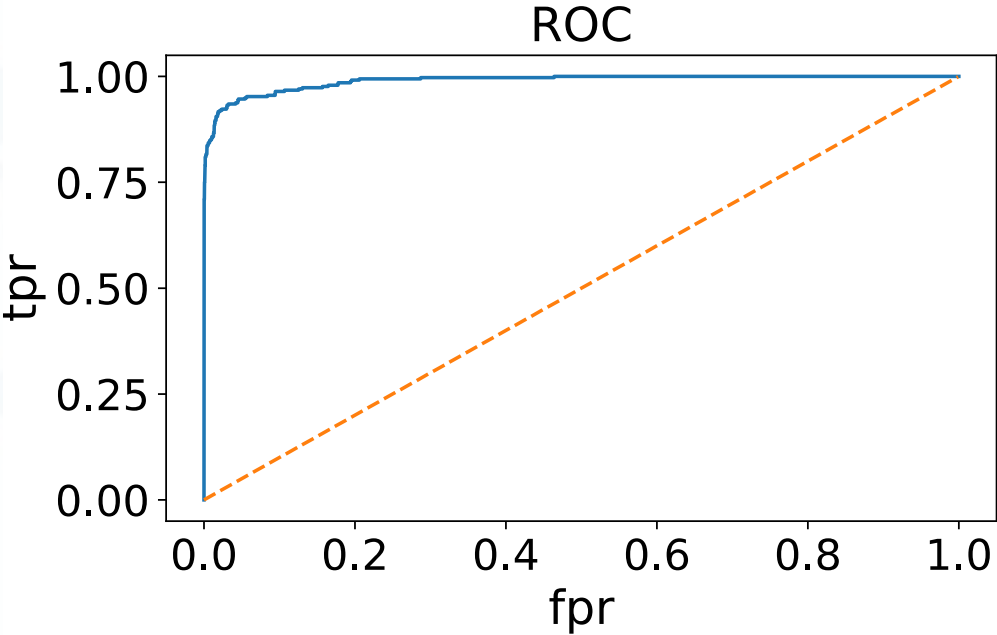


图1 虚假交易模型ROC图

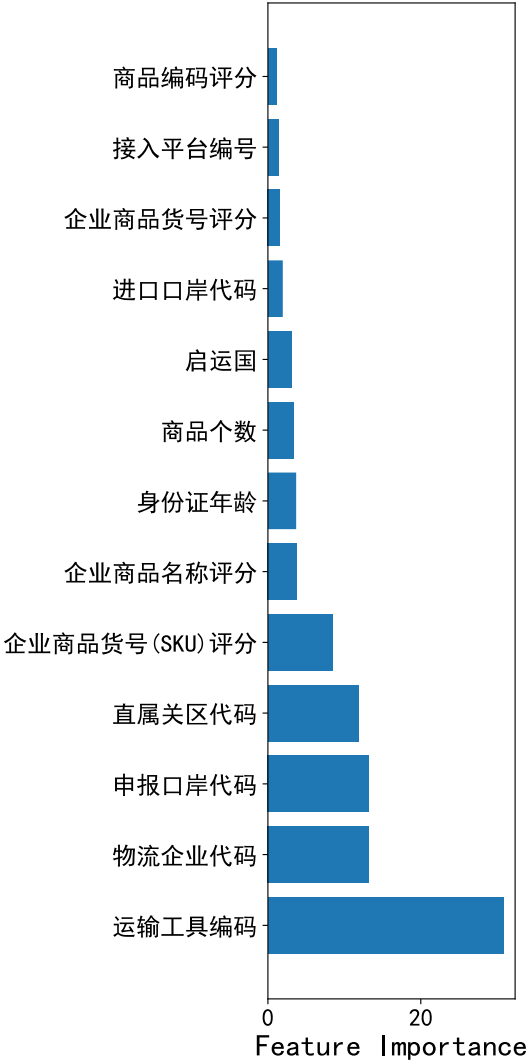


图2 虚假交易模型特征重要性图

效果评估

安全准入

Precision	Recall	F1	AUC
0.482	0.397	0.435	0.989

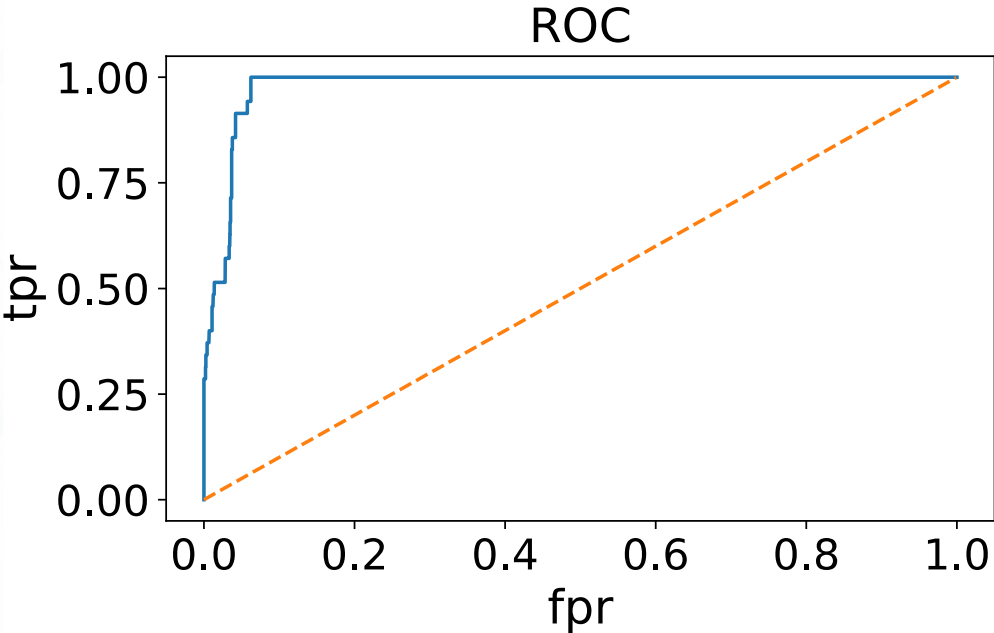


图3 安全准入模型ROC图

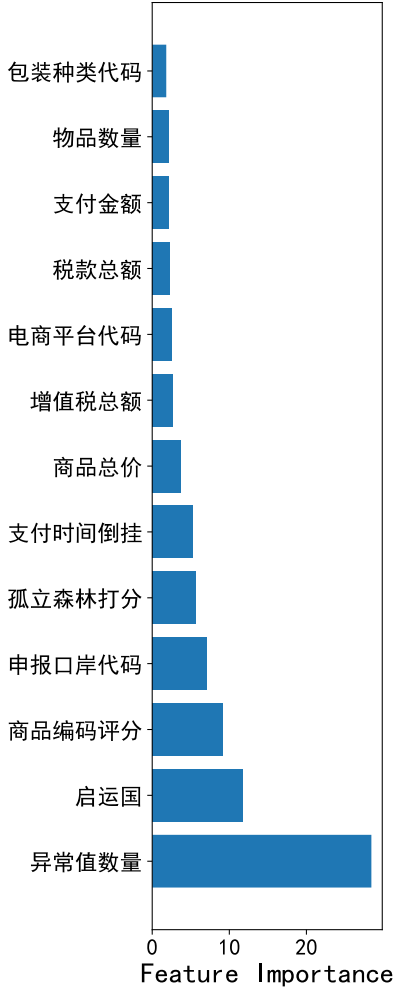
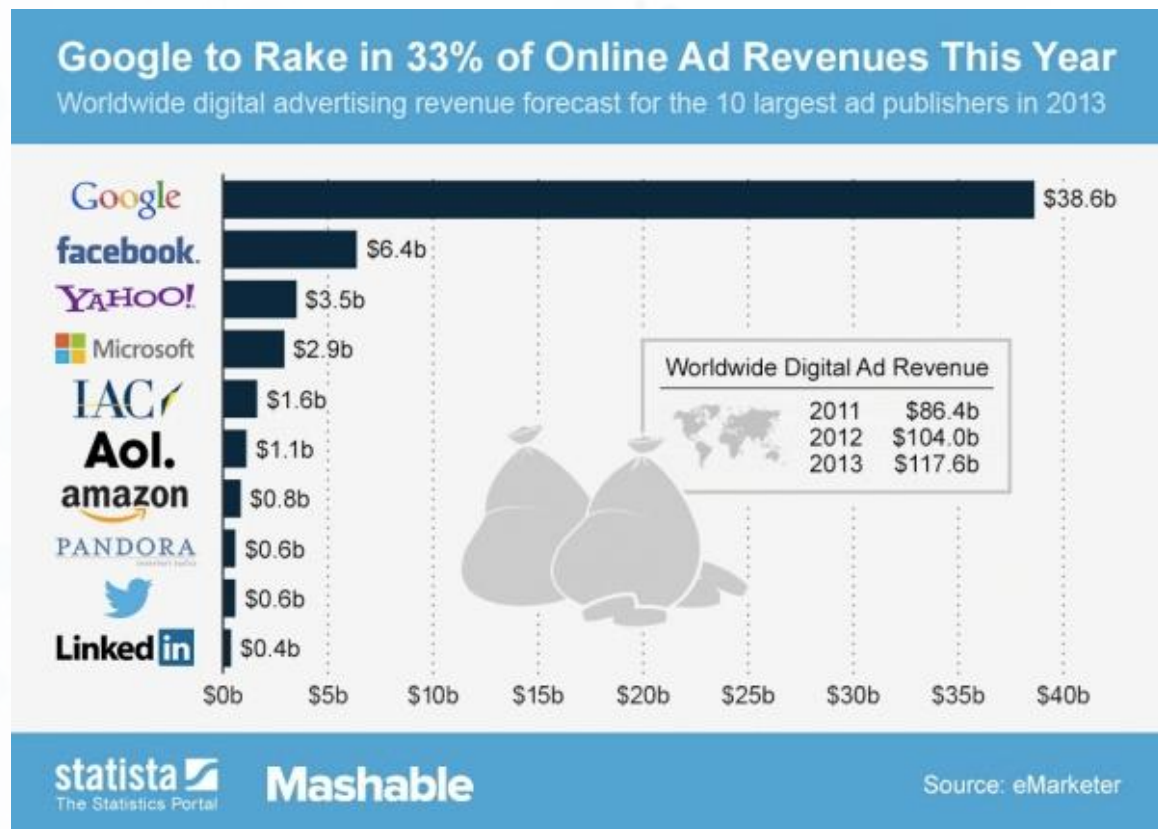


图4 安全准入模型特征重要性图

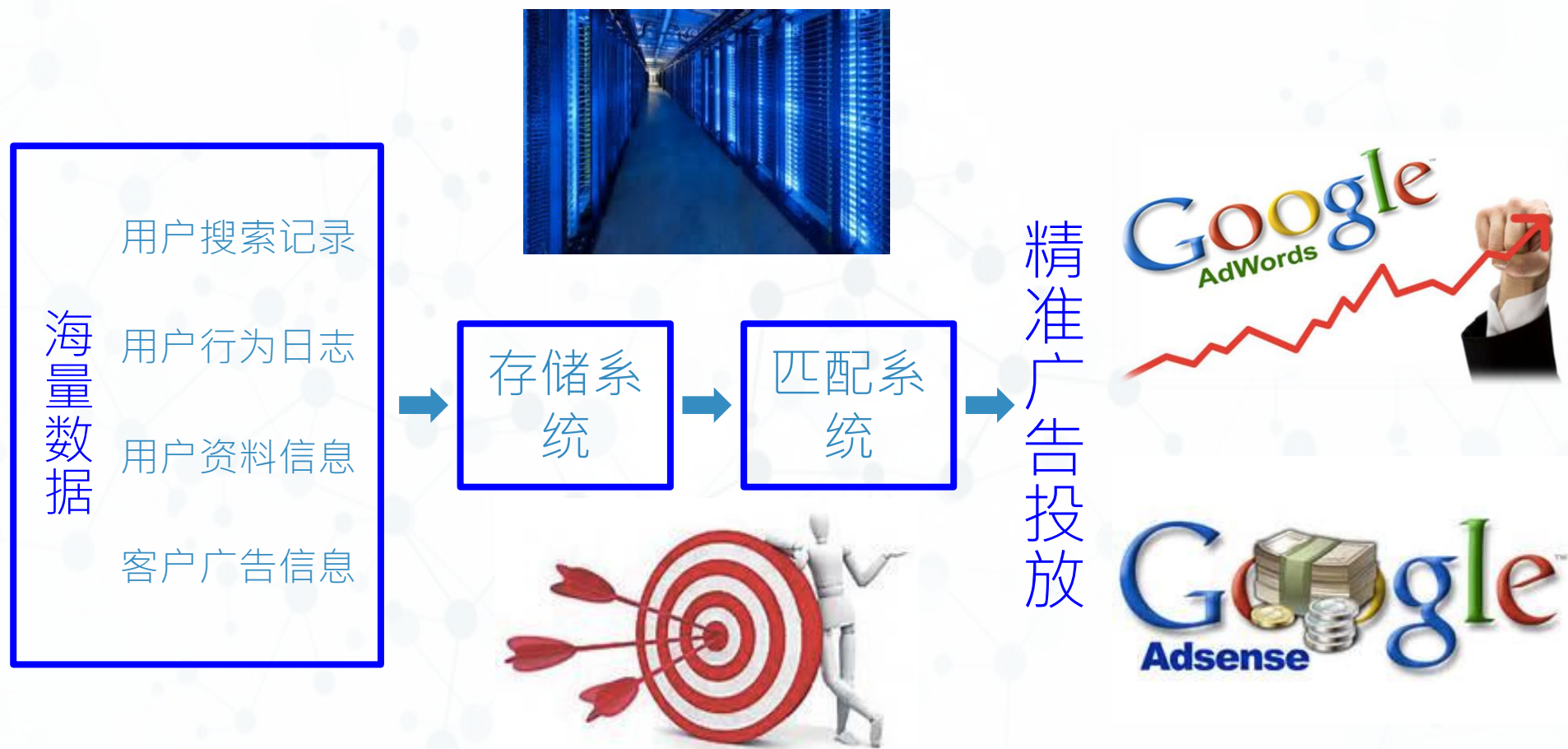
大数据的商业价值

谷歌的金库：精准广告投放



2022年全年，谷歌服务贡献营收**2535亿美元**，同比增长7%，**收入占比约为90%**，其中谷歌广告业务共实现营收2245亿美元，收入占比约为80%；谷歌云贡献营收263亿美元，同比增长37%，收入占比约为9%；其他业务实现营收10.7亿美元，对冲收益贡献营收约20亿美元，二者合计营收占比约为1%

谷歌的金库：精准广告投放



大数据下淘宝“千人千面”

2018年11月12日凌晨，阿里公布了淘宝天猫“双十一”购物狂欢节全天的销售额：

支付宝全天成交金额为：

2135亿

比2015年的 1682 亿增长：

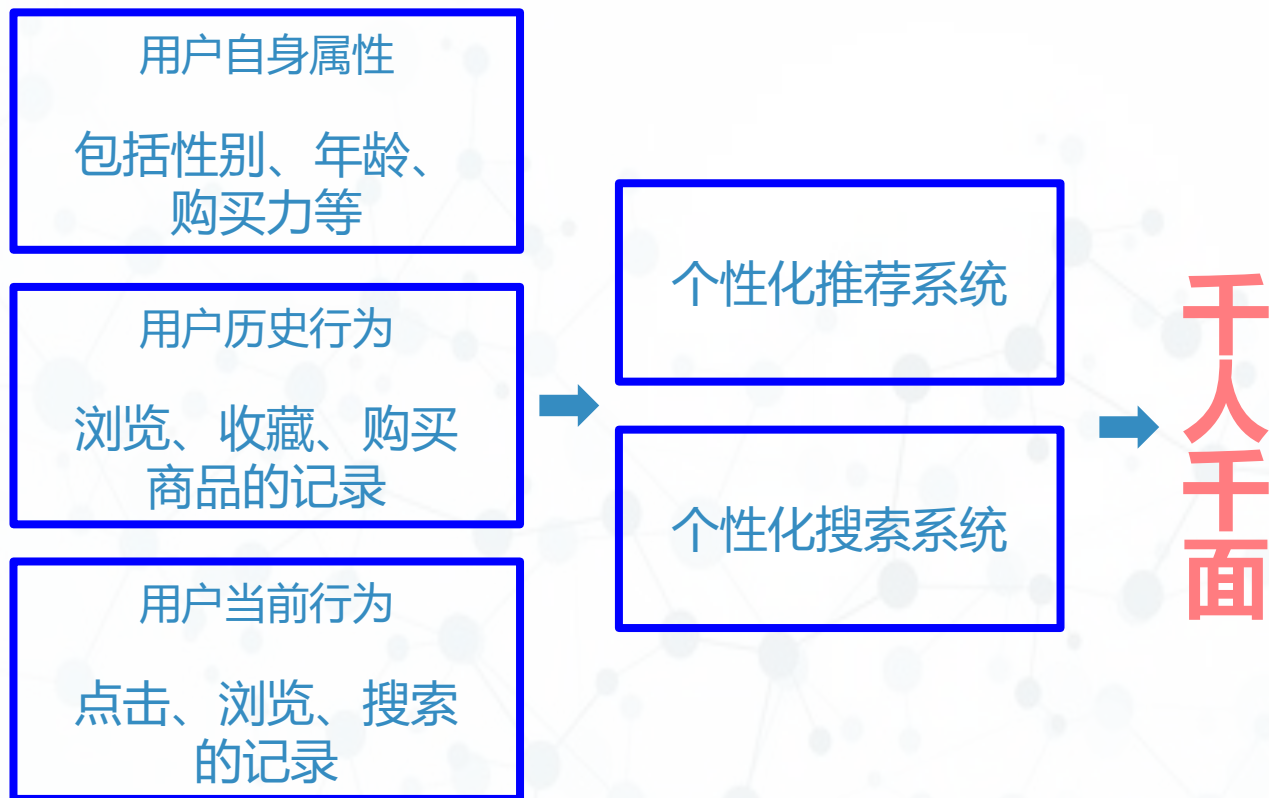
27%

订单：

12.35亿



大数据下淘宝“千人千面”



腾讯视频的大数据-用户规模

- › 超过**10亿**的客户端下载量，**3.4亿**全平台月度活跃用户
- › 每个活跃用户平均每天使用**2小时45分钟**
- › 所有活跃用户一天使用总时长约为**10980年**
- › 重大直播事件**1000万**用户同时在线



腾讯视频的大数据



《大奉打更人》 上线第一天，就带来了100万新增注册会员，超过3000万的收入，远超视频版权支出

服务器日志 (10+T/日)
Web server, LB, CDN server, 广告投放, 搜索, P2P server, application trace...

客户端日志 (10+T/日)
PC客户端, Flash/网页客户端, 移动App, P2P引擎...

大数据的价值 - 商业运营

- › **古装穿越剧**在上个月产生了多少次播放？上周呢？昨天呢？前一个小时呢？
- › 不同渠道带来的独立用户有多少？他们的停留时间和留存率如何？
- › 每一部视频的投入（版权、带宽）和产出（广告收入，付费点播）比如何？分地区分析？分终端分析？
- › 过去6个月**徐七安、临安公主**在视频观众里受关注程度变化趋势如何？



一个馒头引发的…

- › 淘宝：我吃完一个馒头，问我，你要不要来一个馒头？
- › 豆瓣：我买了个馒头，他问我，你要不要来碗米饭？
- › 百度：“老板，给我俩馒头。湖南株洲馒头机制造厂供应优质馒头机”
- › 谷歌：我想买个馒头-Z馒头无敌好吃！(Ad) 为你找到X家馒头店，(正文)。提示：是否使用Google shopping？
- › 腾讯：正当我要买馒头时，在后面拍了拍我，“同学，来我这买，一模样，还有豆沙馅，充值还有馒头皮换装！”
- › 360：让我摸一下，免费送馒头。
- › 小米：馒头便宜啦，就是卖完了，预订吧。

蚂蚁金服 “让天下没有难借的钱”

- › 传统银行借贷，即使不计其实体营业点运营成本和各种人力成本，借贷运营成本很高
- › 倘若借贷一元，其涉及借贷数据的硬件及软件花费将远超其借贷利息所得收益
- › 阿里小微信贷，利用大数据技术，使得信贷门槛降低，真正一元起贷



蚂蚁金服（网商贷）

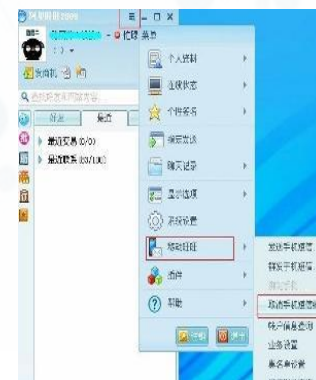
- › 基于大数据的放贷
- › 2013年阿里小贷全年新增投放贷款1000亿元
- › 截至2014年2月中旬，阿里小贷累计投放贷款超过1700亿元
- › 服务小微企业逾70万家，户均贷款余额不超过4万元，不良率小于1%
- › 远低于传统银行的10%-15%的不良率
- › 秒级放贷

阿里小贷发展路径图



蚂蚁金服的水文模型

- 某河道水位达到某个值，但人们无法依据这个数值采取应对，是准备防汛还是不做任何动作？下月河道水位走高还是走低，会否影响防汛等河道管理的措施？
- 但如果将这个值放到历史数据及周边河道数据中，就可以做出一定判断：如比过往同期，这个数据是否变高了，高了多少；以往这个时期后，河道水位又是怎么变化的。
- 通过该店铺自身数据的变化，以及同类目类似店铺数据的变化，判断客户未来店铺的变化。
- 如过往时点，该店铺销售会进入旺季，销售额就会增长，对外投放的额度就会上升，结合这些水文数据，系统可以判断出该店铺的融资需求；
- 结合该店铺以往资金支用数据及同类店铺资金支用数据，可以判断出该店铺的资金需求额度。
- 判断的数据甚至包括购物评价，阿里旺旺聊天内容



大数据的思维变革

大数据的核心要素

- 不同应用都有自己的模型
- 模型经常是数据引导的
- 难以发现从未出现
- 需要数据科学家
- 同时需要专业知识



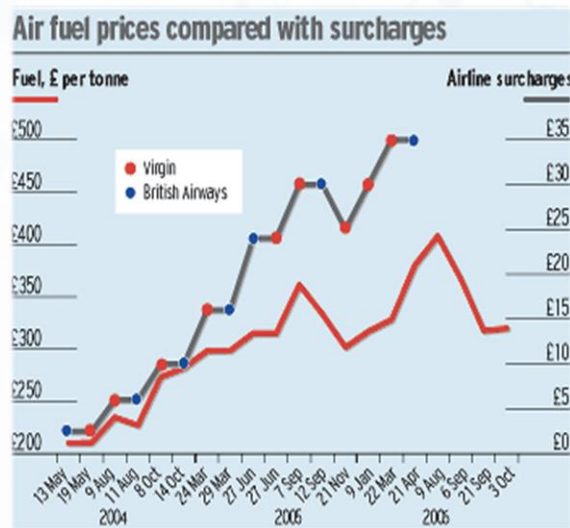
- 尽可能多的数据
- 不同来源的数据
- 不同系统的数据
- 不同种类的数据
- 数据价值随时间递减

- 监测、预测、预警、仿真、决策支持
- 不同行业有不同的应用
- 更多的关注要解决的问题
- 需要行业专家的支持
- 有时需要专业的解读

大数据之思维的变革

一旦完成了数据本身的目的之后，数据就没有用处了吗？

- › 在飞机降落之后，票价数据就没用了吗？
- › 一个检索命令完成之后，检索数据就没用了吗？
- › 用户完成购物后，订单数据就没用了吗？



不会做饭的裁缝不是个好司机

发现·好货

定制我的偏好



NBA球衣詹姆斯韦

根据您"浏览"过的"篮球服"推荐

588 收藏

收藏

不喜欢

我要自定义



根据您"浏览"过的"拖鞋"推荐

964 收藏



NBA 可调节球队

根据您"浏览"过的"篮球球迷用品"推荐

80 收藏

数据资源与传统资源（石油）的区别

› 越用越**多**
你有、我有、大家有

› 价值的增加通过**横向扩张**
单一数据资源用处有限
不同来源、不同类型的数据增加价值

› 随着时间流逝，价值**降低**

› 应用**松耦合**

› 越用越**少**

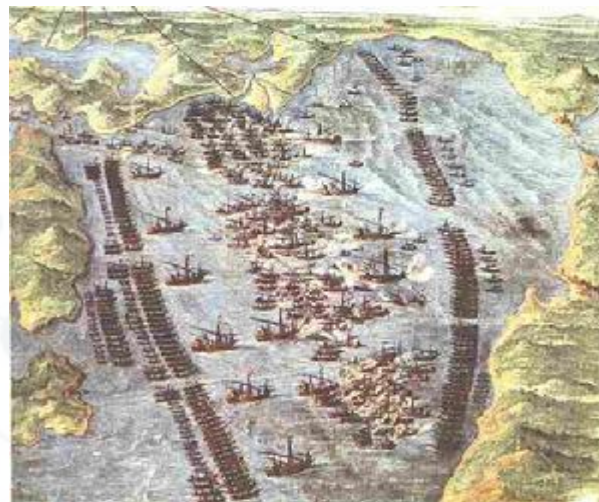
› 价值的增加通过纵向**拉长**产业链
石油烧掉最浪费
做成塑料、纤维、合成革、纺织服装

› 随着时间流逝，价值**越来越高**

› 应用**紧耦合**

勒班陀海战

- › 发生在1571年10月7日
- › 欧洲联军：13000水手，28000步兵，127条战船
- › 奥斯曼军队：13000水手，34000步兵，278条战船
- › 战争结果：基督教联合舰队大胜，共击毁土耳其舰队舰船113艘，俘获117艘，缴获火炮274门以及无数金银财富和细软，击毙土军将士3万人，俘虏8000人，使土耳其舰队几乎全军覆没。而联合舰队只损失舰船12艘，被俘1艘，死伤1.5万人。
- › 作战部署比较严密；规模较大，场面惊人；火力战、冷兵器战、接舷战并用，战斗异常激烈、残酷，



勒班陀海战结果的不同解读

西班牙人的解读

- 要加强接舷战的能力
- 增加水手数，增加接舷训练
- 强化装甲
- 提高冲撞能力



英国人的解读

- 接舷很重要，但火炮是未来
- 提高射程，
- 提高射击精度
- 提高战舰速度



英西班牙大海战

- 1588年7月-10月
- 西班牙130艘各类战舰，30000水手和士兵，1124门火炮
- 英国197艘战舰，10000人，200余门火炮
- 战争持续了3个多月，集中战斗1周多
- 西班牙惨败，沉40多艘战舰，阵亡5000余人
- 英国无一沉船
- 英国从此走上了“日不落帝国”道路
- 西班牙从此开始衰落



三块钢板的故事

- 故事来自一位数学家。二战后期，美军对德国和日本法西斯展开了大规模战略轰炸，每天都有成千架轰炸机呼啸而去，返回时往往损失惨重。美国空军对此十分头疼：如果要降低损失，就要往飞机上焊防弹钢板；但如果整个飞机都焊上钢板，速度航程载弹量什么都要受影响。
- 怎么办？空军请来数学家亚伯拉罕·沃尔德。沃尔德的方法十分简单。他把统计表发给地勤技师，让他们把飞机上弹洞的位置报上来，然后自己铺开一张大白纸，画出飞机的轮廓，再把那些小窟窿一个个添上去。画完之后大家一看，飞机浑身上下都是窟窿，特别是机翼，发动机。只有飞行员座舱、尾翼、机箱几个地方几乎是空白。
- 于是解决方法看起来很简单了



什么是大数据（技术的观点）

- › 当数据的**规模和性能要求**成为数据管理分析系统的**重要设计和决定因素**时，这样的数据就被称为大数据
不是简单地以数据规模来界定大数据
要考虑数据查询与分析的复杂程度
- › 以目前计算机硬件的发展水平看
针对**简单查询**（如关键字搜索），数据量为**TB至PB级**时可称为大数据，10亿-1万亿
针对**复杂查询**（如数据挖掘、音视频非结构化数据），数据量为**GB至TB级**时即可称为大数据

神农尝百草，中医在这里

$$F_1 = F_2 = G \frac{m_1 \times m_2}{r^2}$$

A photograph of the atomic bombing of Nagasaki on August 9, 1945. It shows a massive, dark, and billowing mushroom cloud rising from the city, with a bright fireball at its base. The surrounding landscape is visible in the foreground, showing the sea and some structures.

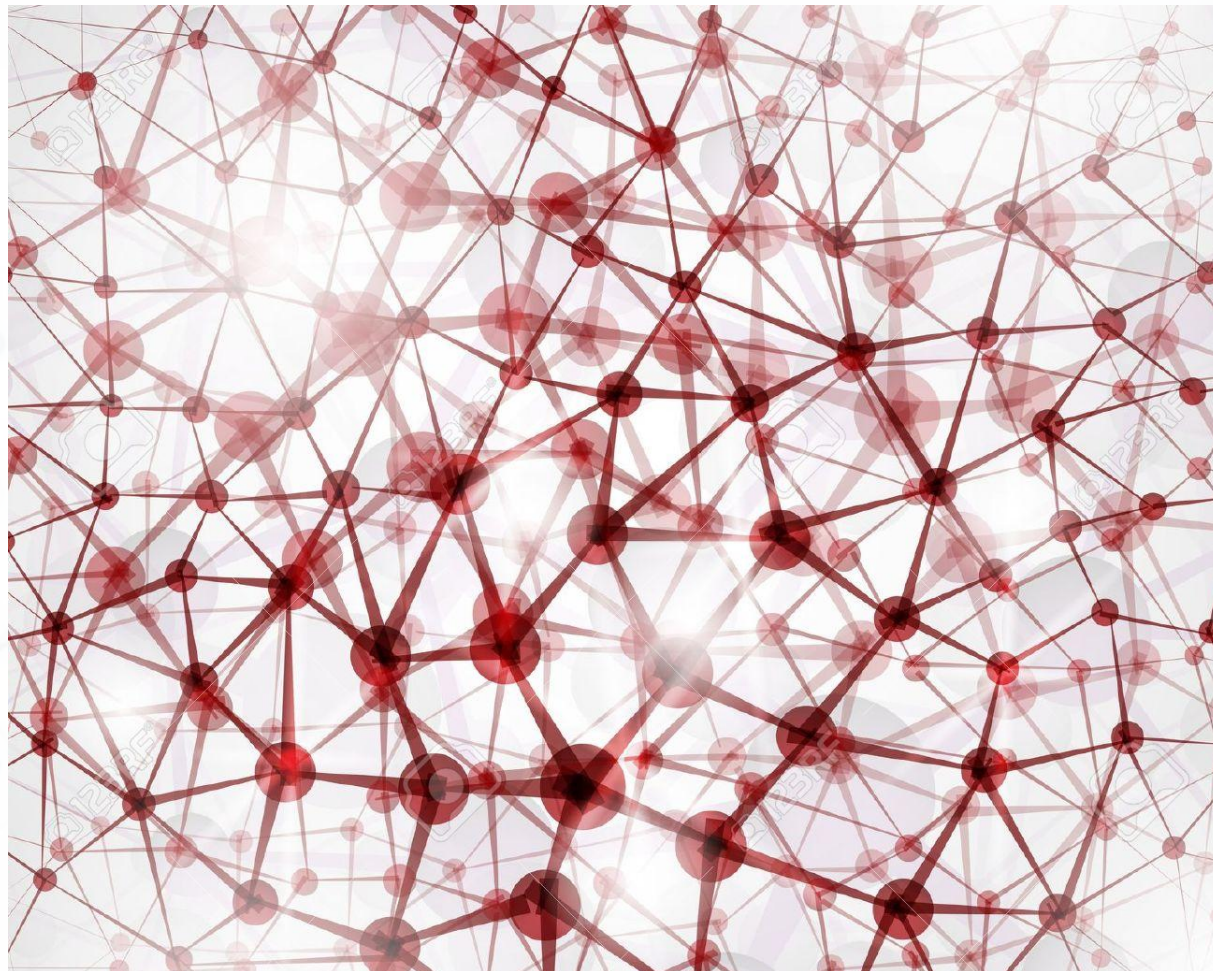
氢弹爆炸仿真计算

• 数据直接说话



课程主要内容

- › Python入门
- › 算法分析初步
- › 基本数据结构
(线性表、链表、栈、队列)
- › 递归
- › 排序与查找
- › 树和树相关算法
- › 图和图相关算法



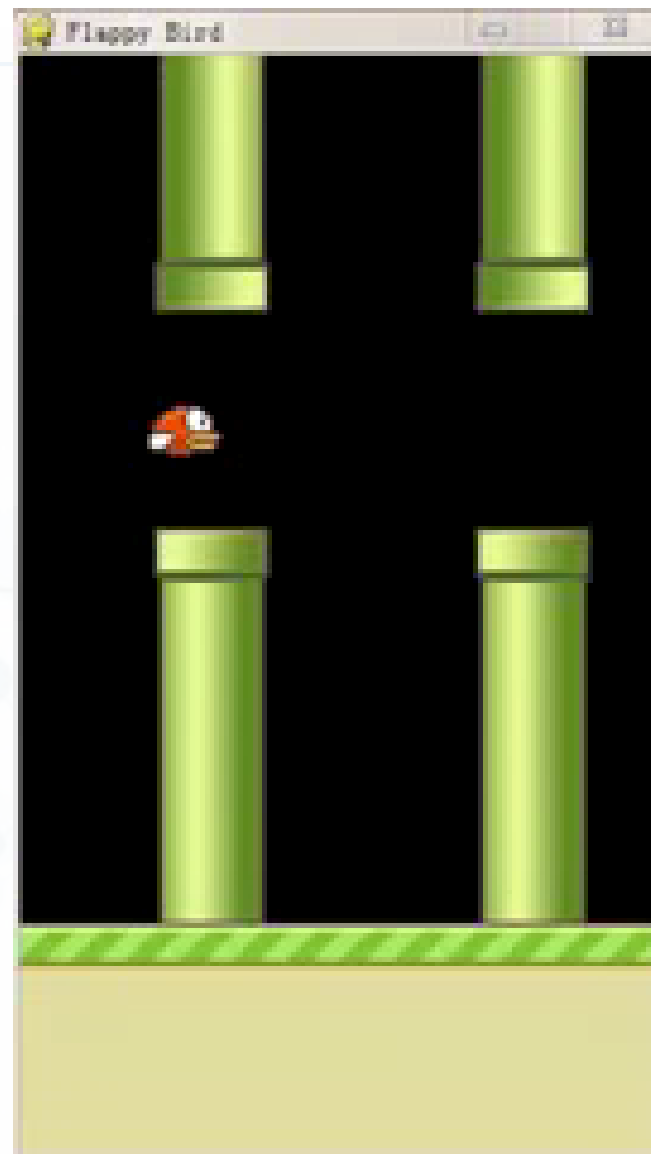
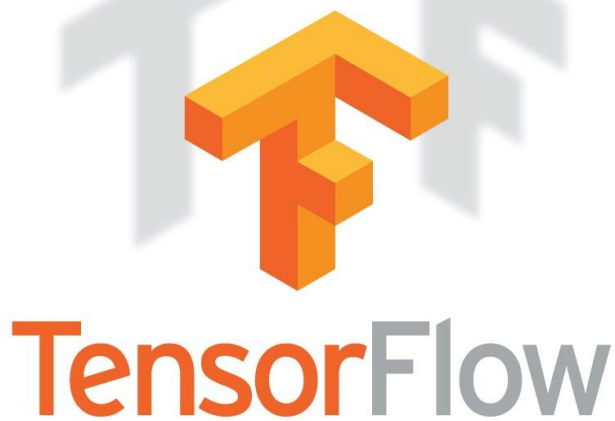
为什么选Python?



- › **代码短小精悍，干净整洁**
没有变量声明，不需要花括号begin/end，也没有分号，比java短80%，比C短98%
- › **解释执行，上手就玩，编程小白福音**
不用焚香沐浴安装GB级别的开发环境compile/build，可以随问秒答，边玩边改
- › **“包装内附带电池”**
自带大量运行库，网络、数据库、图形图像、GUI、压缩加密一应俱全，几行代码建网站
- › **功能无比强大，开发左右逢源，最酷的网络应用都是用它**
Google/Youtube/Instagram/豆瓣……，NASA用它哦
- › **搞大数据和AI的人们也爱它**
有各种面向大数据处理的数据模型、数值分析、机器学习、空间分析等Python工具随时恭候

Python坐稳人工智能时代的头牌语言

- › Google开源的AI系统Tensorflow
- › 支持Python和C++开发
- › 160行Python代码可以让AI从游戏视频中学习玩Flappybird



课程目标和考评

› 课程目标

掌握数据结构和算法的基本概念，特别是数据结构的逻辑结构和物理结构
理解每种数据结构及其相应的运算
能采用“抽象”和“自顶向下”方法分析问题，使之简化，设计算法
能分析给定数据结构和算法的效率，特别是时间效率和空间效率

› 教学方式

课堂授课、课后作业、团队大作业、网络交流

› 评分方式（可能会微调）

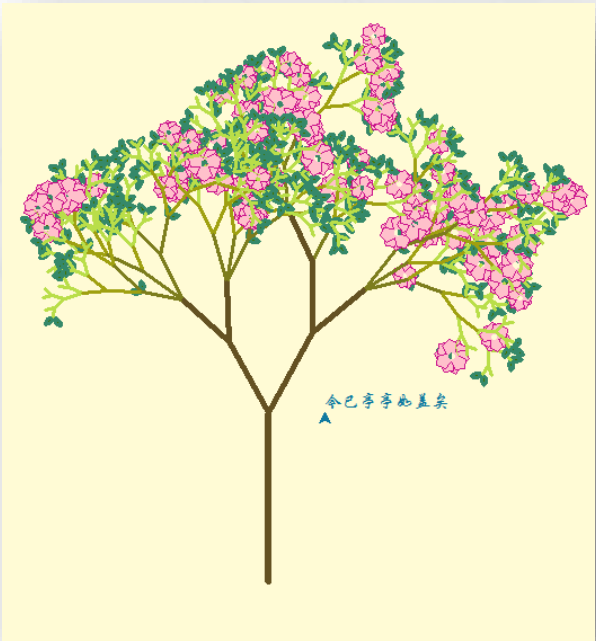
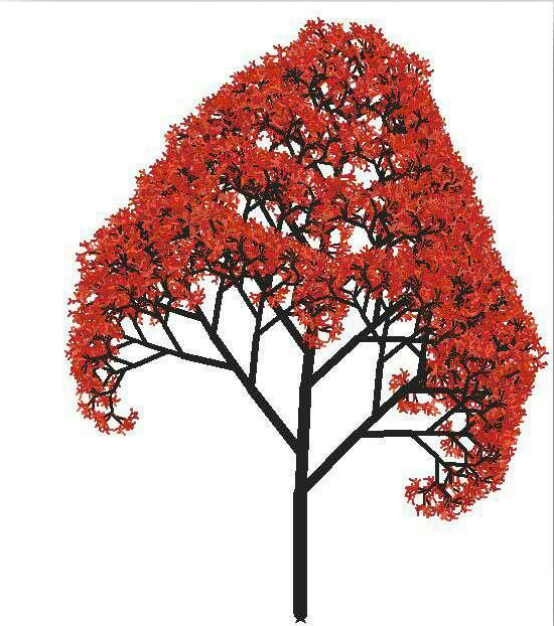
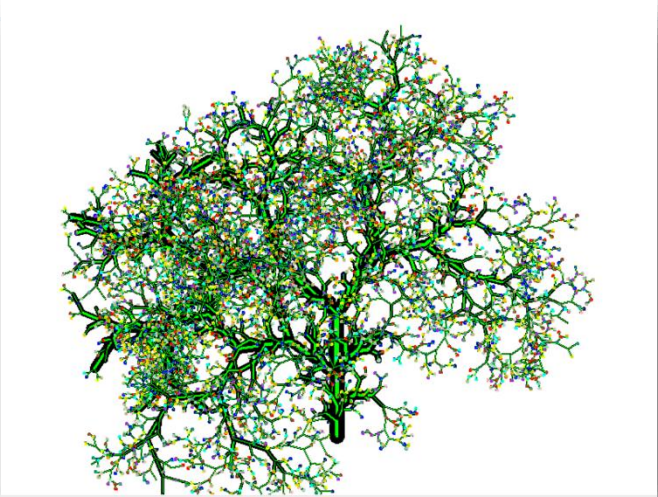
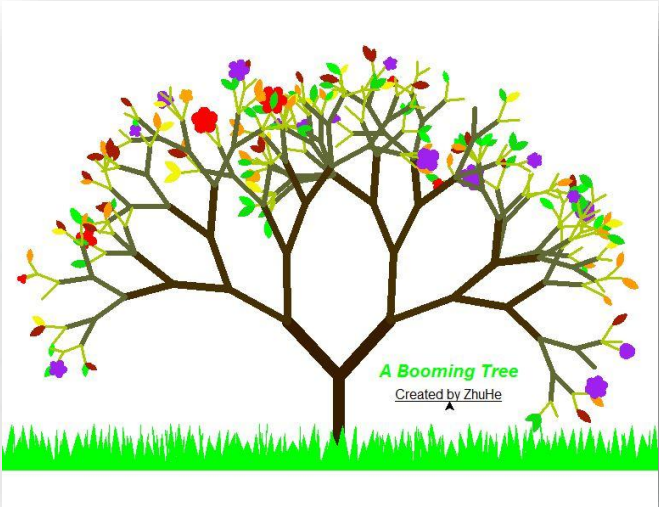
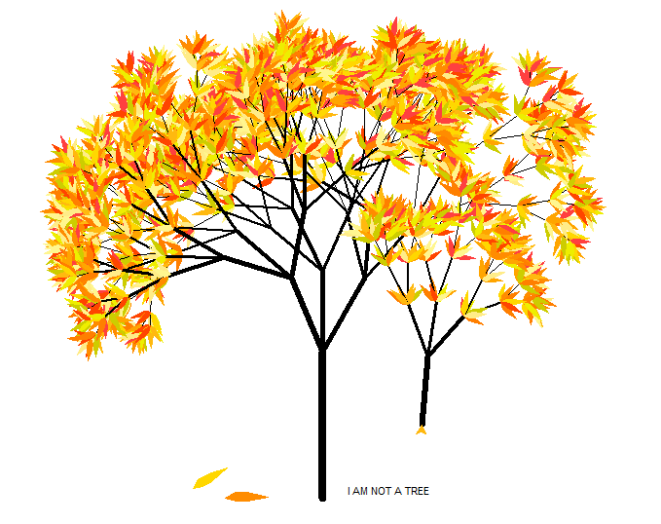
作业（报告/上机之类）30%，一次大作业（团队竞赛）20%，
期末考试（机考）20%，期考考试（闭卷）30%

DSA'17~22都做了什么？

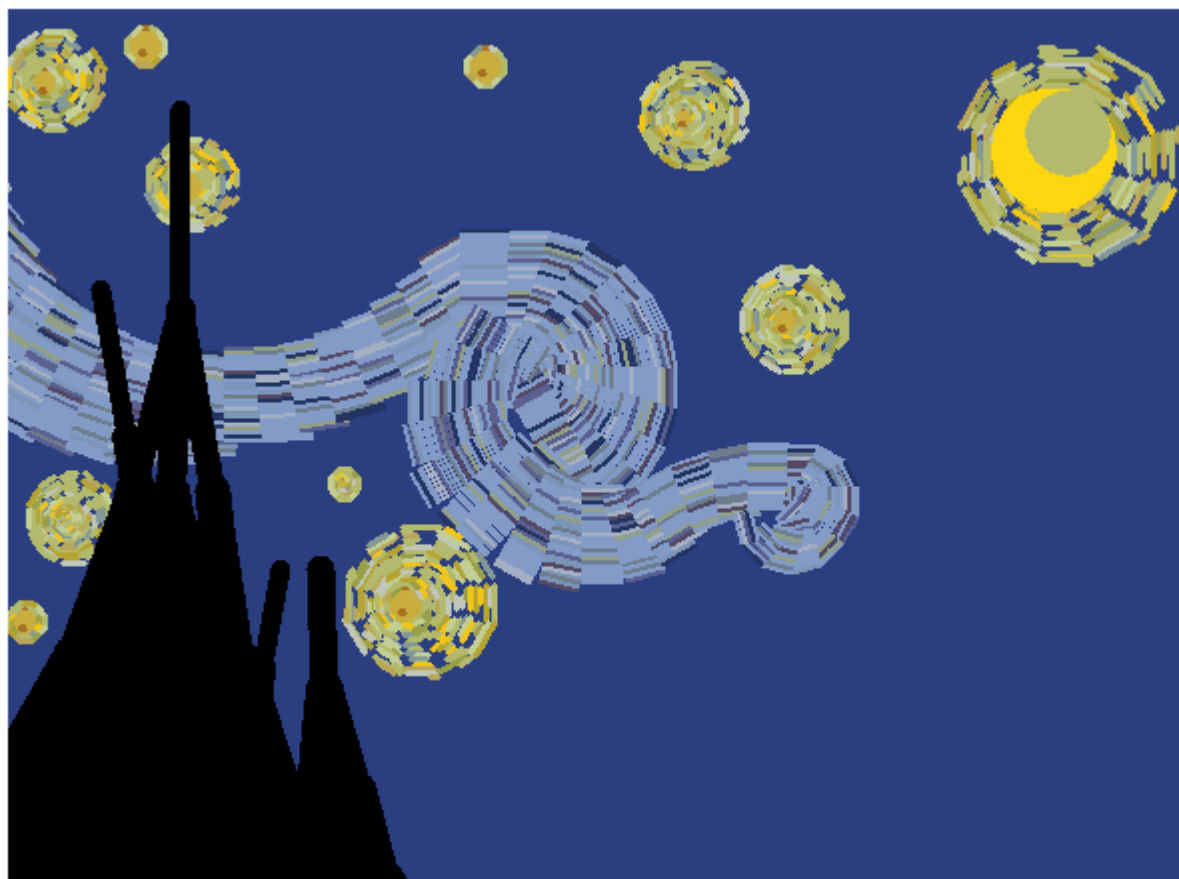
	人数	平时	作业 / 报告	递归作业	分组大作业	闭卷考试
2018	112	20	30分	分形树	纸带圈地 20分	30分
2019	88		30分	分形树	星际吞噬 30分	40分
2020	89		30分	汉诺塔	贰零肆捌 30分	40分
2021	110		30分	汉诺塔	星际群落 30分	40分
2022	128		30分	汉诺塔	方块大战 30分	40分

2019年起，不再进行课堂点名或者签到

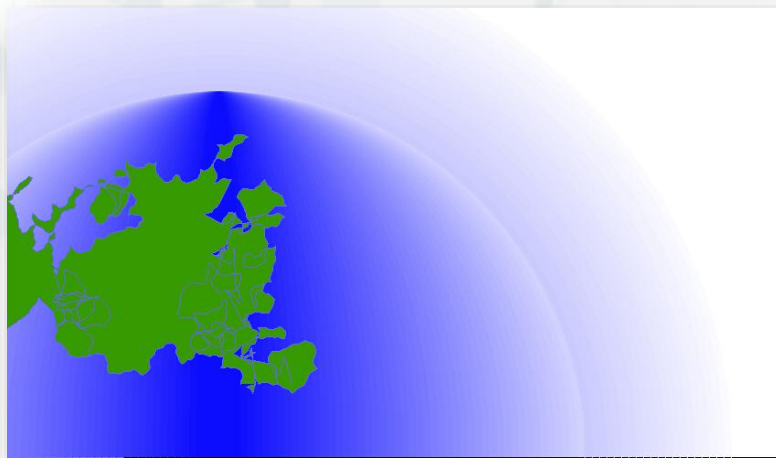
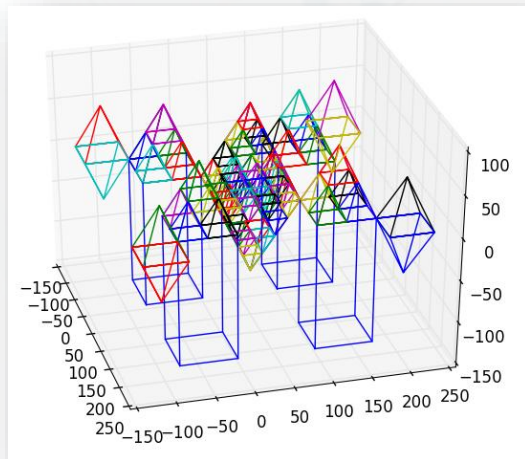
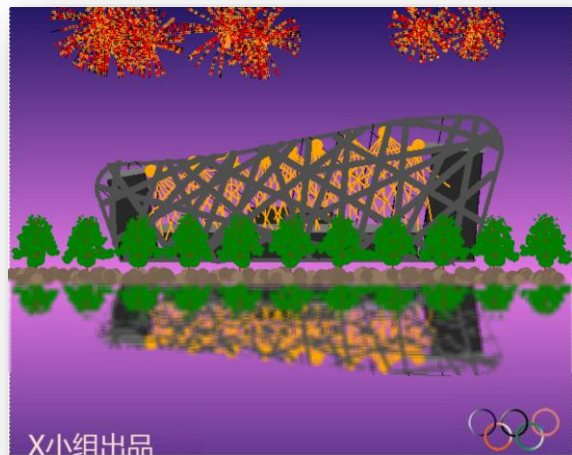
dsa'17: 二叉树的艺术



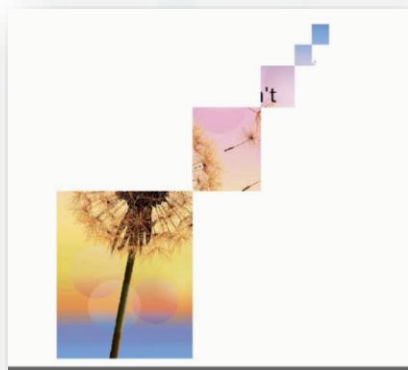
sessdsa'17: 二叉树的艺术



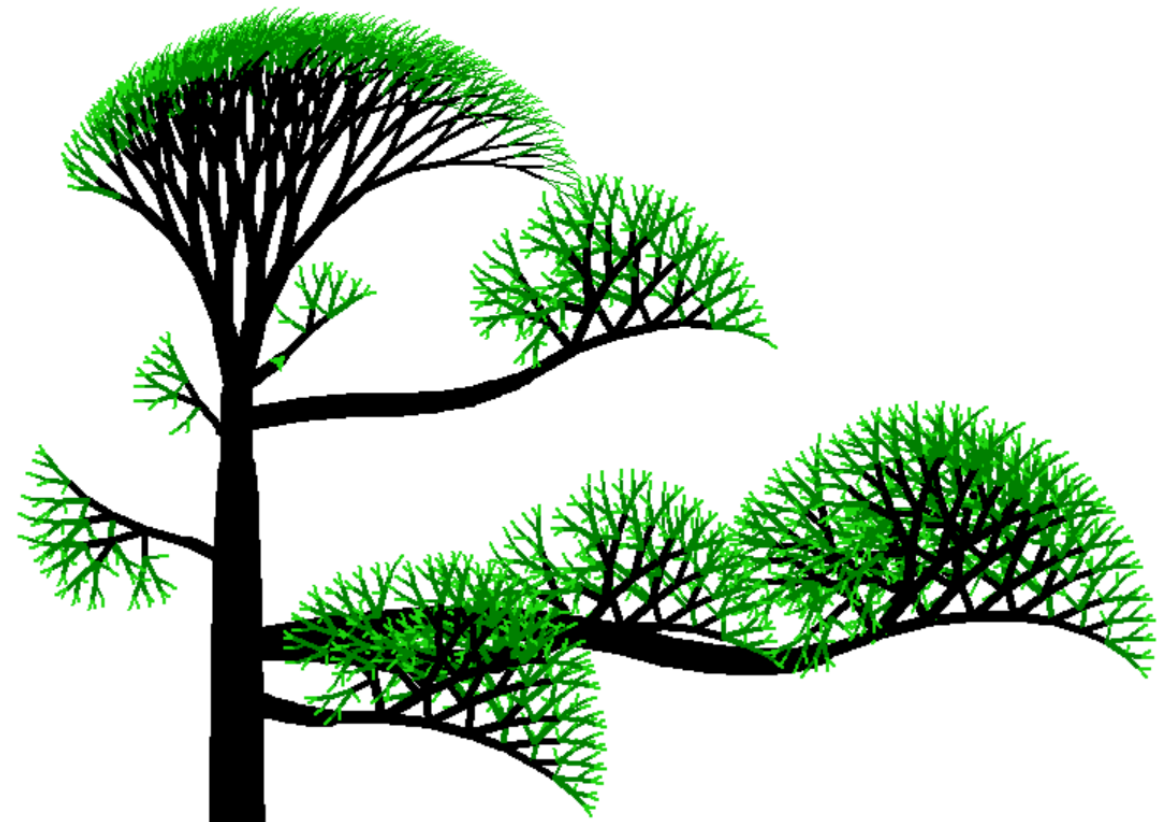
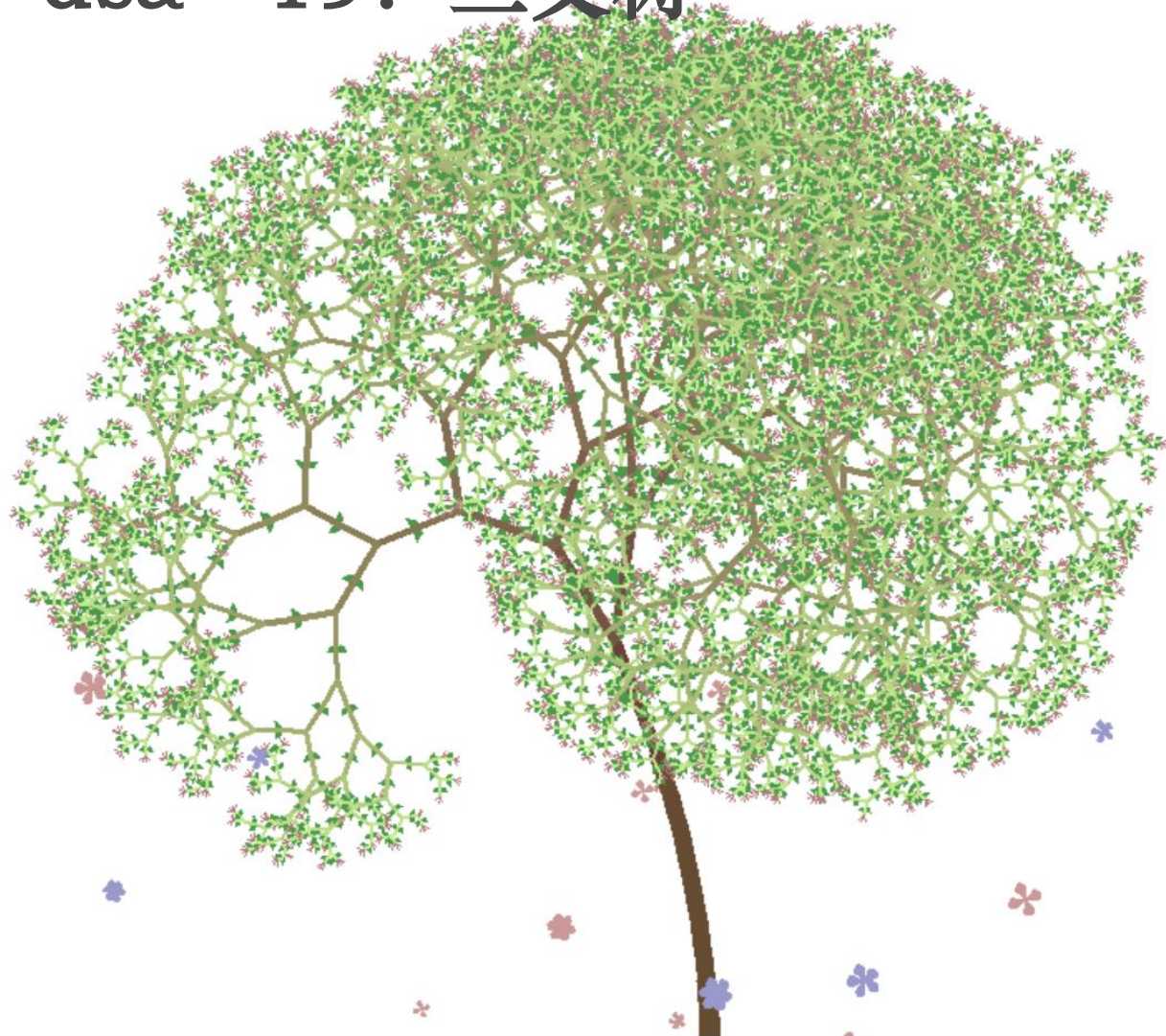
dsa'18: 递归视觉艺术



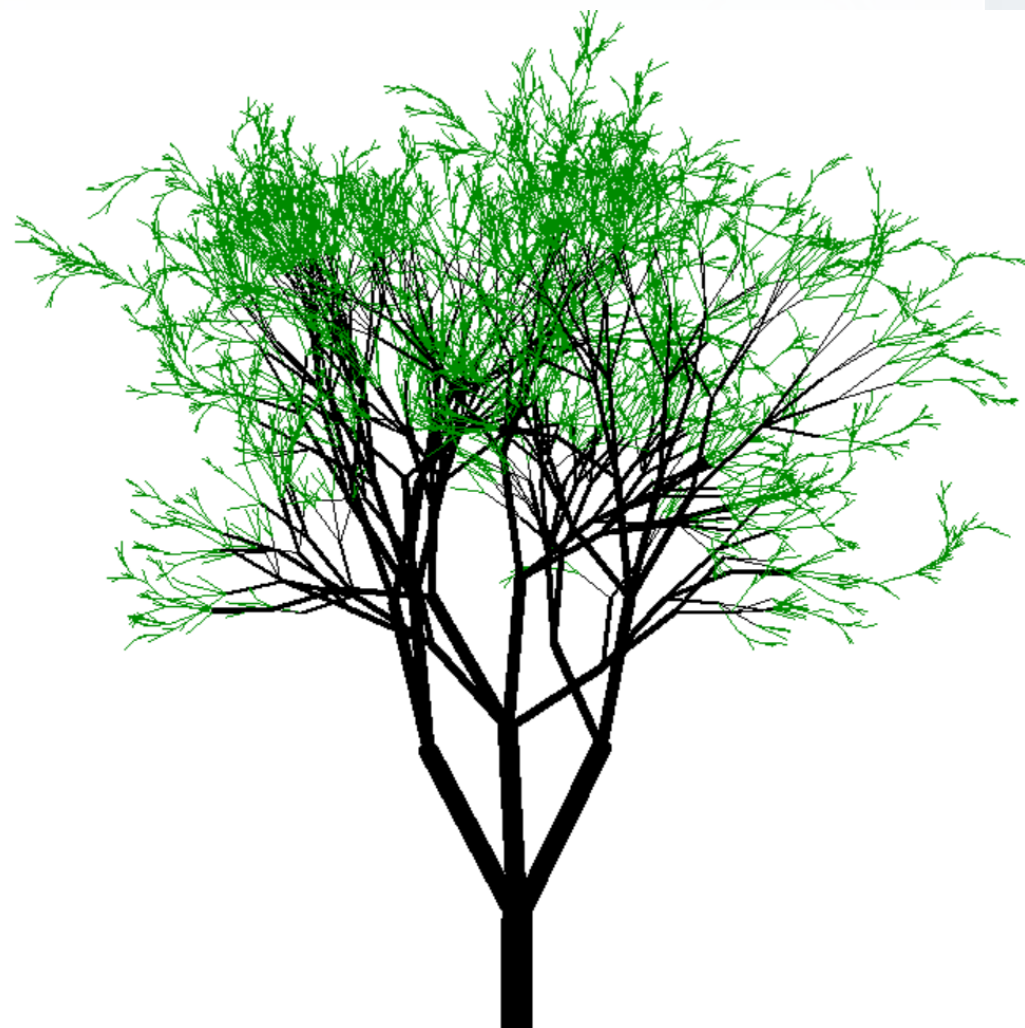
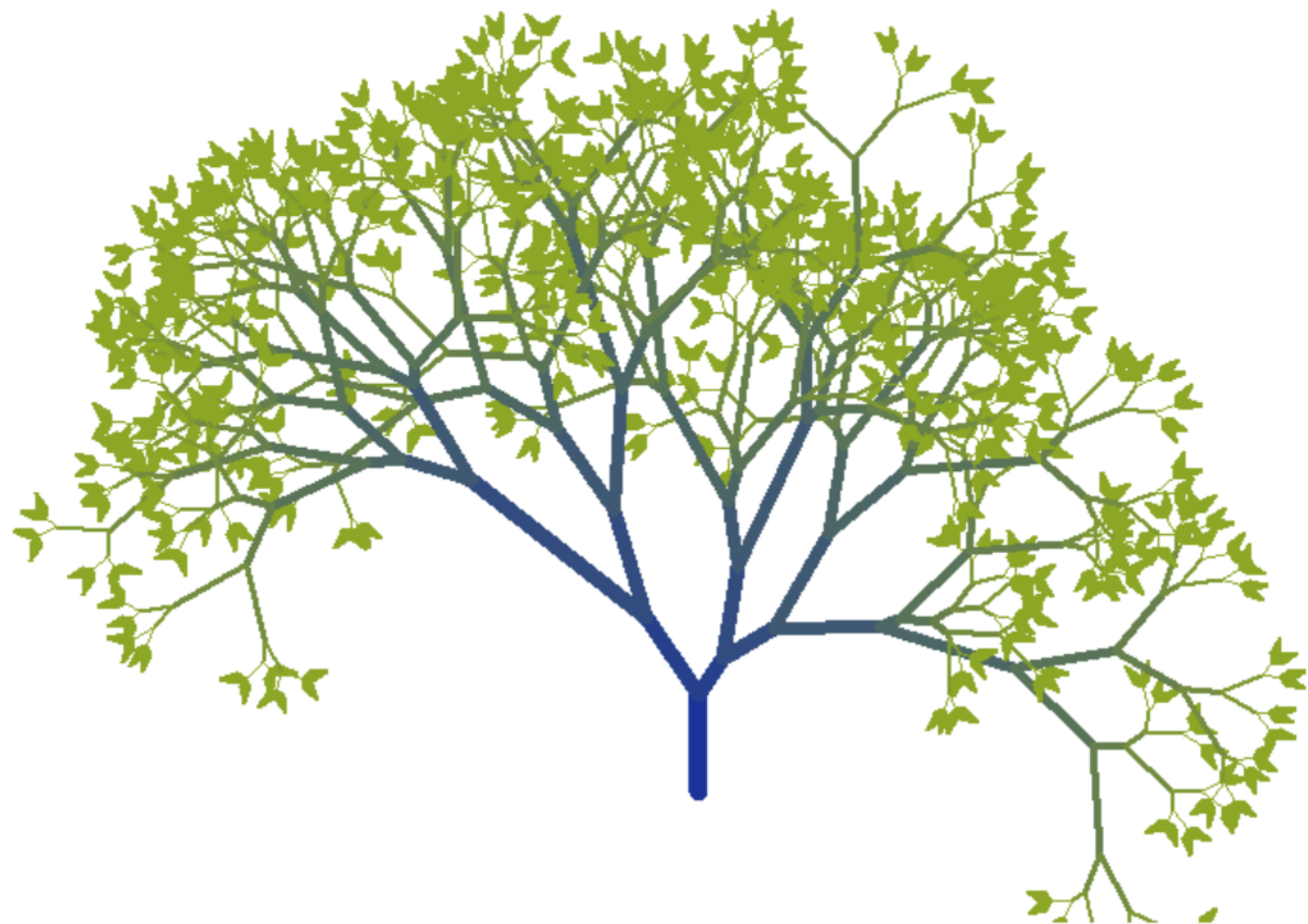
dsa'18: 递归视觉艺术



dsa '19: 二叉树

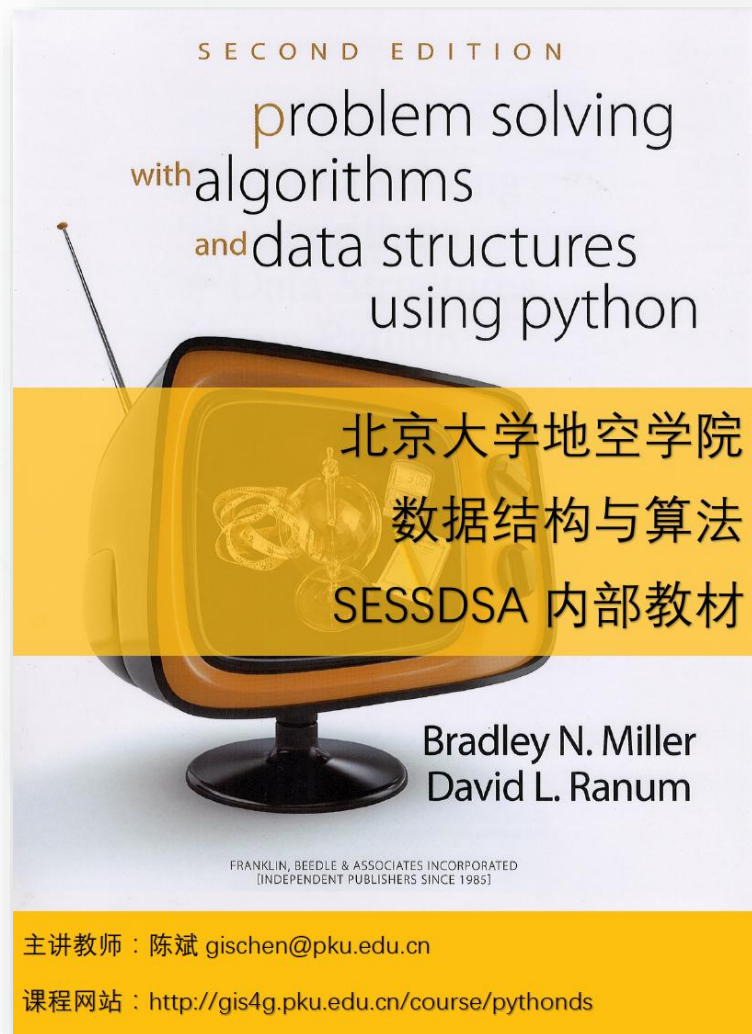


dsa '19: 二叉树





DSA'15: 教材众包翻译／助教之选优秀作业



DSA'15：黑白棋大战



```
SOUTH: GOLF >- ROMEO 36,28 (END, time: 21.925933:10.703209)
SOUTH: ALPHA <- GOLF 20,44 (END, time: 12.404031:56.17389)
SOUTH: GOLF >- ALPHA 36,28 (END, time: 29.095909:17.664121)
SOUTH: GOLF >- BRAVO 39,25 (END, time: 0.006144999999995:41.687809)
SOUTH: ROMEO <- ALPHA 12,52 (END, time: 0.006144999999995:41.687809)
WEST: INDIA <- KILO 27,37 (END, time: 73.501144:53.690597)
WEST: KILO <- INDIA 28,36 (END, time: 66.200688:62.920071)
SOUTH: ALPHA >- ROMEO 43,21 (END, time: 51.511866:0.005240000000001)
SOUTH: BRAVO >- ROMEO 36,28 (END, time: 23.439517:0.005974000000021)
SOUTH: ROMEO <- BRAVO 14,50 (END, time: 0.007491000000002:39.960571)
SOUTH: GOLF <- LIMA 9,55 (END, time: 18.68421:110.74275)
SOUTH: BRAVO <- GOLF 27,37 (END, time: 26.581381:84.9053)
```

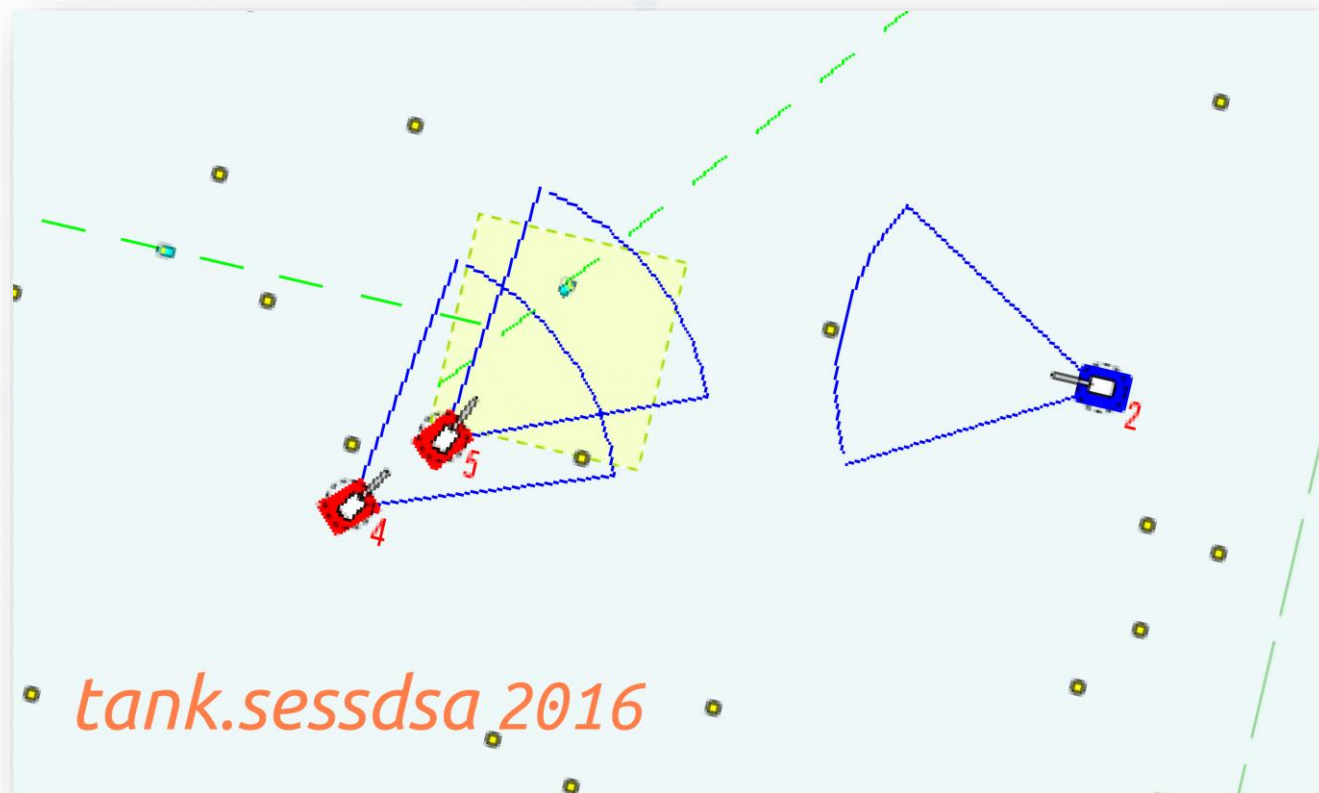
DSA'15：黑白棋大战



DSA'15 : 黑白棋大战



DSA' 16 : 坦克大战

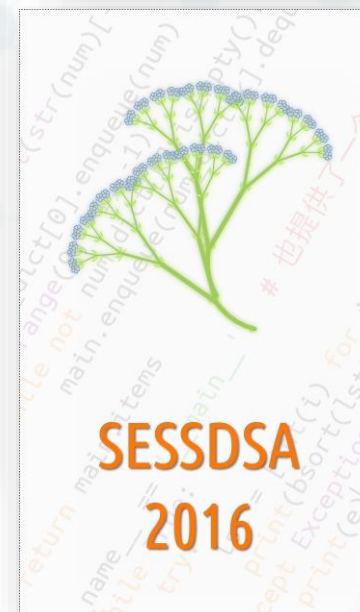


DSA' 16 : 坦克大战

数据结构与算法 (Python)

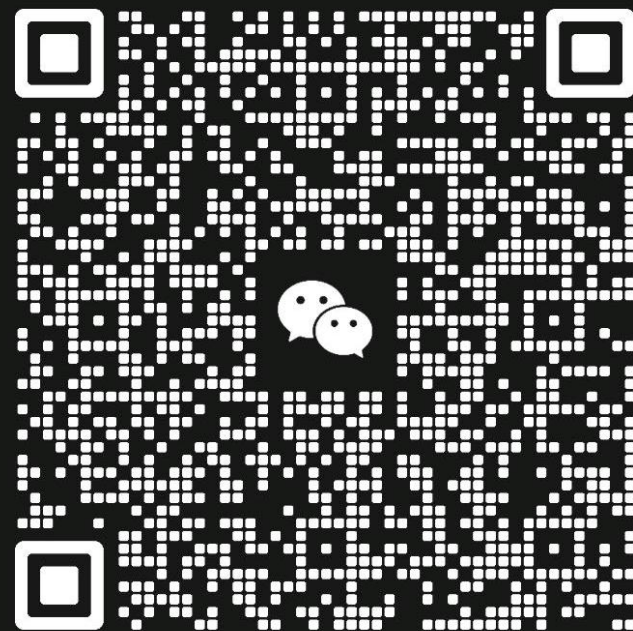


DSA' 16 : 坦克大战



联系方式与课程表

- › 微信群：2025数算(B)
教师：刘云淮（微信:yunhuailiu）
助教：陈嘉、唐健恒
- › 上课地点：二教307
- › 上课时间：
每周二，7-9节；15:10-19:00
- › 上机地点：计算中心7-8#机房
上机时间：每周五(第二周)，9-10节



该二维码7天内(3月4日前)有效，重新进入将更新

我们的教材

- › Problem Solving with Algorithms and Data Structures

在线教材: <http://interactivepython.org/runestone/static/pythonds/index.html>

- › Python数据结构与算法-京东有售

- › 参考资料

数据结构与算法 (Python程序实现) -郭炜, 京东有售

数据结构与算法可视化

- <http://visualgo.net/>

- › 课程网站

<http://www.yunhuai.net/DSA2025/CoursePage/DSA2025.html>

参考书

Python数据结构与算法分析

- [美] [布拉德利·米勒](#)，[戴维·拉努姆](#)著，[吕能](#)，[刁寿钧](#) 译
数据结构与算法（Python语言实现）

- 迈克尔·古德里奇 等
数据结构与算法

- 张铭，北京大学
算法与数据结构-C语言描述（第3版）

- 张乃孝主编，高等教育出版社，2011，6
数据结构-C语言版

- 严蔚敏等，清华大学出版社

有用的软件和网站

- › 在浏览器里运行Python

<http://pythonfiddle.com/>

<http://pythontutor.com/visualize.html>

<https://www.python.org/shell/>

- › 集成开发环境Geany

<https://www.geany.org/Download/Releases>

- › 更高级的集成开发环境PyCharm

<https://www.jetbrains.com/pycharm/download/>

- › 有用的地空学院陈斌老师的网站

<http://gis4g.pku.edu.cn/course/pythonnds/>



python



powered

