



# 经济管理大数据案例及其应用

吴纯杰 教授  
上海财经大学



2022年5月



---

# 目录

---

1. 大数据背景和宏观战略
2. 大数据介绍
3. 大数据与国内外经济发展





---

# PART I

# 大数据背景和战略

---

# 大数据时代来临



**Nature**  
2008年9月封面



**Science**  
2011年2月封面

# 大数据时代来临



《纽约时报》2012年2月11日

[HOME PAGE](#) [TODAY'S PAPER](#) [VIDEO](#) [MOST POPULAR](#) [TIMES TOPICS](#)

Log In | Register Now | Help

**The New York Times**  
**Sunday Review** | The Opinion Pages

Search All NYTimes.com

[WORLD](#) [U.S.](#) [N.Y. / REGION](#) [BUSINESS](#) [TECHNOLOGY](#) [SCIENCE](#) [HEALTH](#) [SPORTS](#) [OPINION](#) [ARTS](#) [STYLE](#) [TRAVEL](#) [JOBS](#) [REAL ESTATE](#) [AUTOS](#)

**NEWS ANALYSIS**  
**The Age of Big Data**  
By STEVE LOHR  
Published: February 11, 2012

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.



Enlarge This Image

Chad Hagen

Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers," says Ms. Zhou, whose job as a data analyst suits her skills.

**Multimedia**

Siera =  $6.145 - 16.986 \times (S - 1.858 \times ((GB - FB - PU) \div PA) \times (((GB - FB - PU) \div PA)^2) + 1 \dots$

[f RECOMMEND](#)  
[TWITTER](#)  
[LINKEDIN](#)  
[COMMENTS \(81\)](#)  
[SIGN IN TO E-MAIL](#)  
[PRINT](#)  
[SINGLE PAGE](#)  
[REPRINTS](#)  
[SHARE](#)

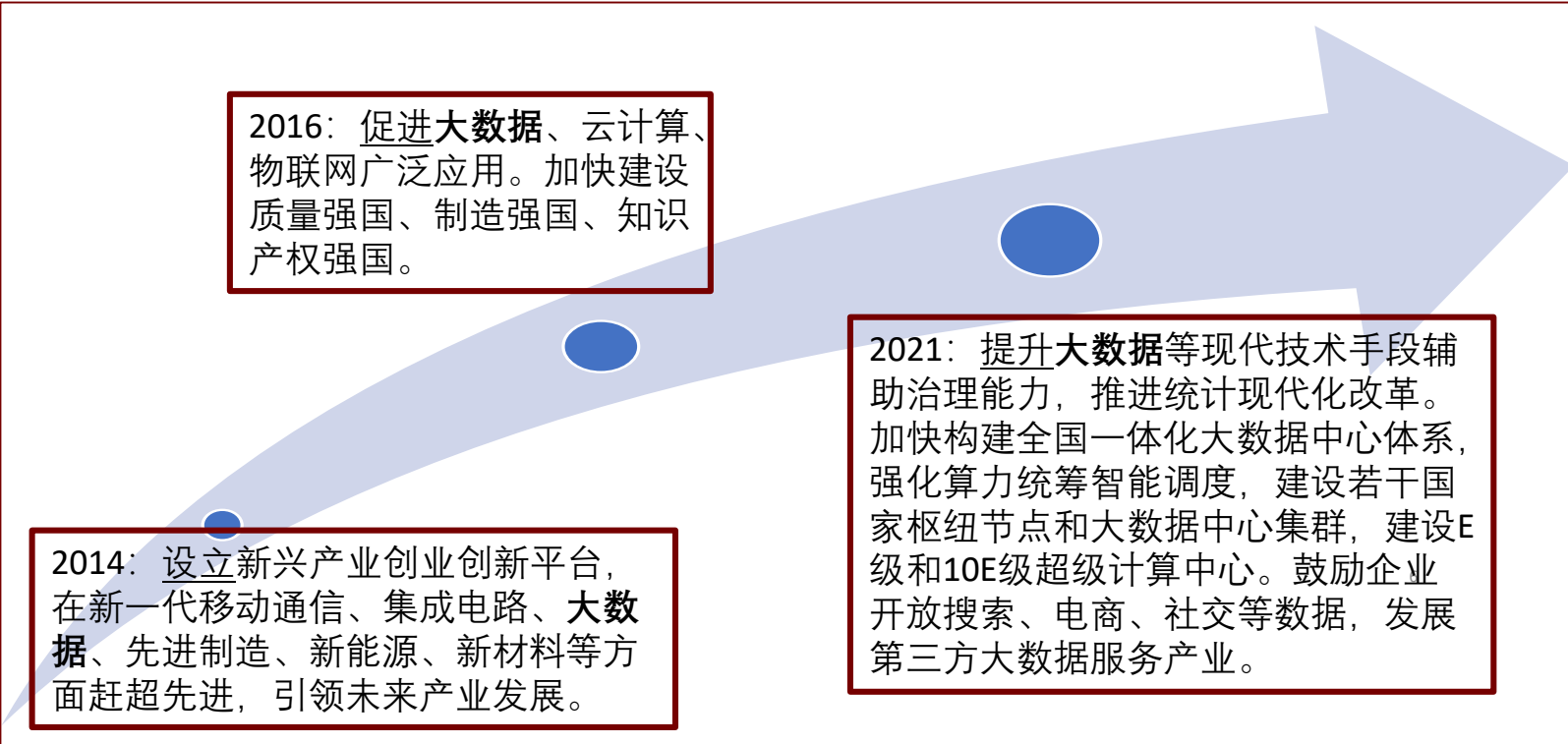
2012年3月29日奥巴马政府投入2亿多美元启动“**大数据发展研究计划**”（美国另一个战略研究计划是1993年宣布的“信息高速公路计划”）

2015年9月5日，国务院正式印发《**促进大数据发展行动纲要**》，意味着大数据发展正式成为国家战略。



# 一、大数据

自2014年以来，“**大数据**”连续八年进入国务院政府工作报告，上升为国家战略。《“十四五”大数据产业发展规划》提出，到2025年我国大数据产业测算规模突破3万亿元，年均复合增长率保持25%左右，创新力强、附加值高、自主可控的现代化大数据产业体系基本形成。



2016: 促进大数据、云计算、物联网广泛应用。加快建设质量强国、制造强国、知识产权强国。

2014: 设立新兴产业创业创新平台，在新一代移动通信、集成电路、**大数据**、先进制造、新能源、新材料等方面赶超先进，引领未来产业发展。

2021: 提升大数据等现代技术手段辅助治理能力，推进统计现代化改革。加快构建全国一体化大数据中心体系，强化算力统筹智能调度，建设若干国家枢纽节点和大数据中心集群，建设E级和10E级超级计算中心。鼓励企业开放搜索、电商、社交等数据，发展第三方大数据服务产业。



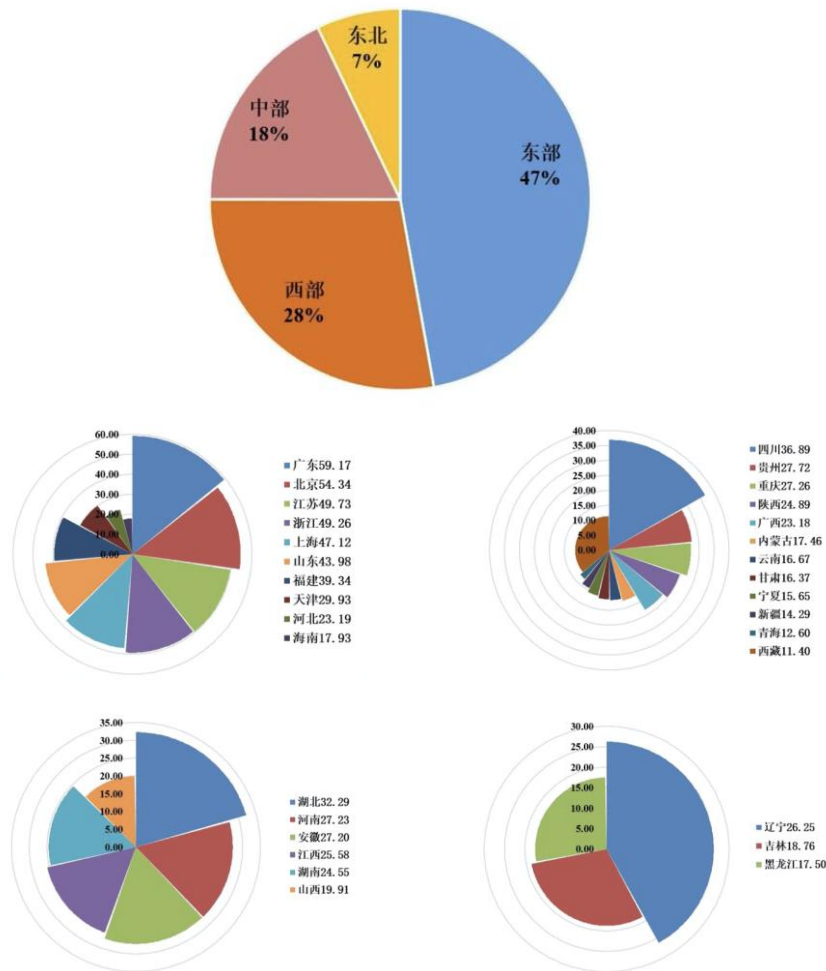
东部地区大数据发展水平较高  
中西部地区赶超势头强劲

# 一、大数据

## 中国大数据产业规模持续增长



数据来源：信通院、中商产业研究院整理





# 大数据时代来临

**战略急需、历史机遇、抢占先机、赢得未来**

- 大数据等新一代互联网技术深刻改变了世界，也让各国站在科技革命的同一起跑线上。中国曾屡次与世界科技革命失之交臂，今天要把握这一历史机遇，抢占先机，赢得未来。



李克强总理  
(2016年5月，贵阳大数据峰会)



# 大数据时代来临



2017年12月8日，习主席在中共中央政治局就实施国家大数据战略第二次集中学习上说：审时度势精心谋划超前布局、力争主动，**实施国家大数据战略，加快建设数字中国**

# 大数据时代来临



人民网 >> 传媒 >> 人民日报新闻研究网 >> 媒介纵横 >> 新媒观察

## 用好大数据须有大智慧

——准确把握、科学应对大数据带来的机遇和挑战

徐宗本

2016年03月15日14:41 来源：人民网-人民日报

《人民日报》，2016年3月15日



徐宗本院士

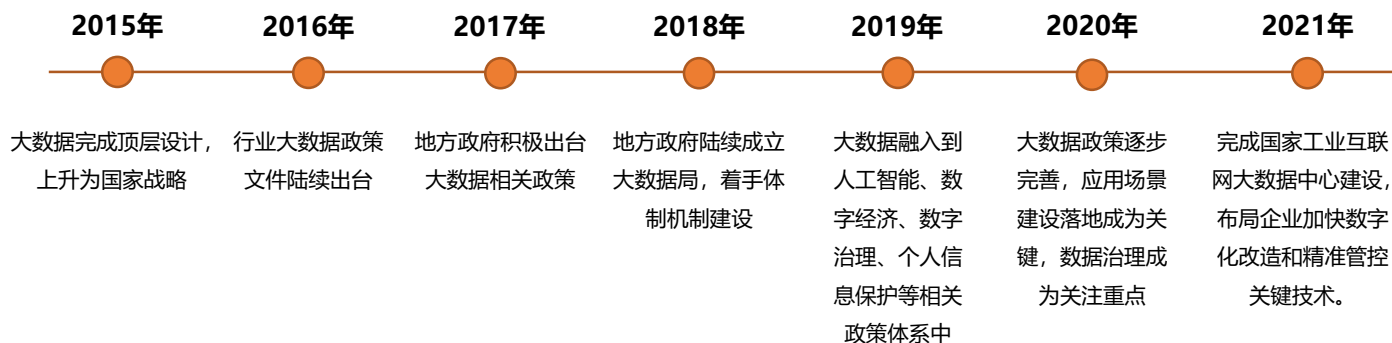


# 一、大数据

## 2012-2021年大数据产业政策关键词



## 2015-2021年大数据产业发展关键节点

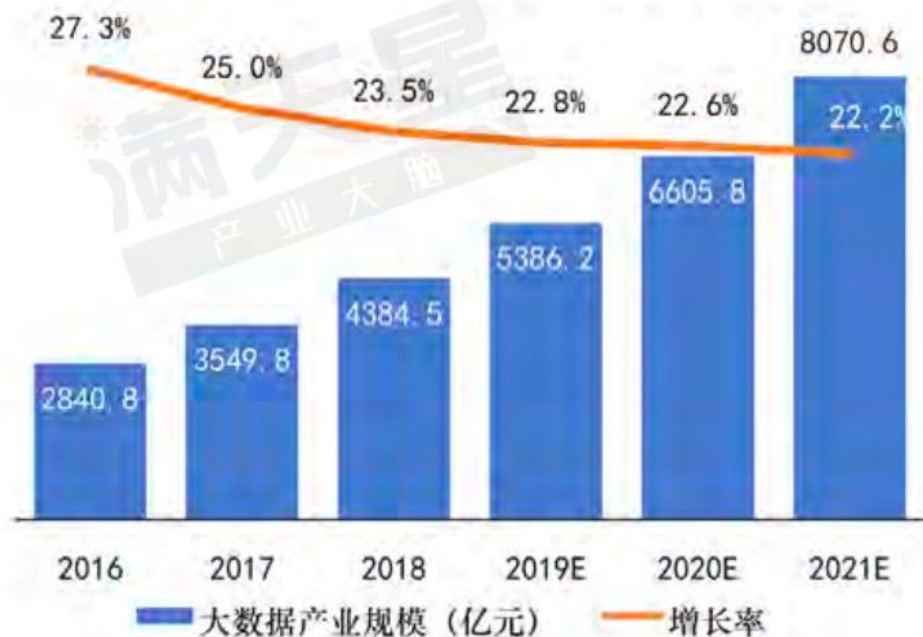




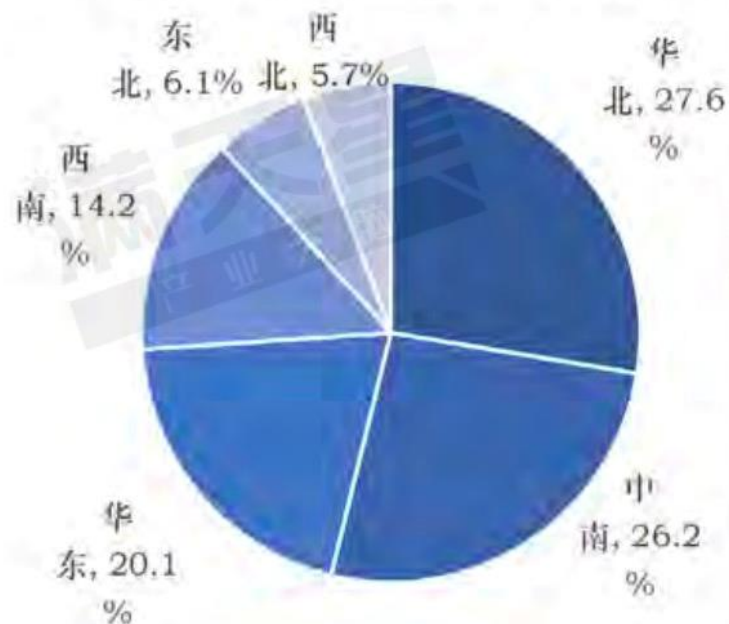
# 大数据时代来临

中国大数据产业规模持续增长，华北、中南和华东地区领跑

2016-2021年中国大数据产业规模及预测



2018年中国大数据企业区域业务布局





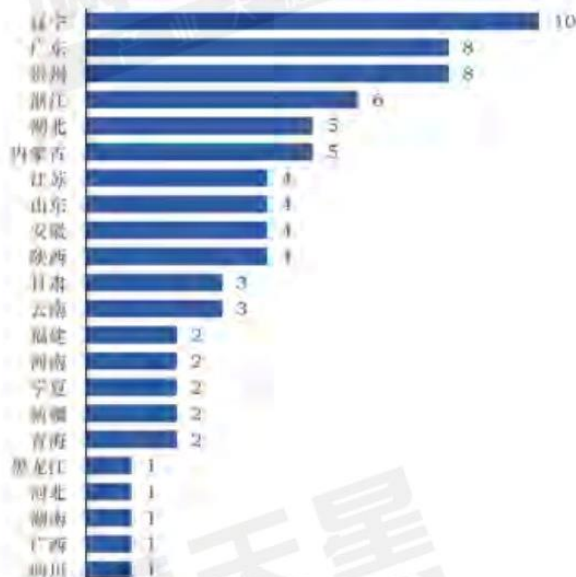
# 大数据时代来临

## 各地政府相继成立大数据管理机构，统筹推进产业发展

组建了政府大数据管理机构

**16个**省级行政区域  
**79个**副省级和地级城市

省内大数据管理机构设置情况



2014-2018年省级大数据管理机构设置情况



数据来源：赛迪顾问，2019



## 贵阳市大数据发展成果

### ➤ 成立时间

2016年9月，**贵阳市大数据发展管理委员会**组建。

### ➤ 主要职责

作为政府工作单独部门设立的大数据发展管理委员会，负责统筹推进全市大数据产业发展和应用，其设立成为贵阳智慧城市建设的有力抓手。

### ➤ “云上方舟”的建立

电商的冲击使得实体零售业遭遇重重压力、面临前所未有的挑战。贵阳市委、市政府提出让大数据成果广泛运用于政商民的要求。贵州首个智慧商业综合体——**云上方舟**的建成，是对大数据的深度融合运用，探索破解实体商业痛点、开启了智慧商业的发展新模式。



## 贵阳市大数据发展成果

### ➤ 大数据和商业有机结合

云上方舟推出了智能导航机器人、多媒体信息大屏、蓝牙导航、亮化信息工程、监控安全报警系统、人体探测客流等智能服务，力求为消费者打造科技感十足的智慧体验购物中心。

### ➤ 智慧运营体系

在云上方舟的智慧运营体系中，客户购买的商品、消费习惯、消费额度、优惠券使用等都将转化为数据收集，形成独一无二的标签，让深层次个性化需求得到满足。云上方舟引进全世界最先进的智能设备和大数据技术，拥有专业的大数据分析团队，在大数据应用上做到更全面、更精准。



# 大数据时代来临



序号	名称	成立时间	主要职能	官网链接（部分机构没有独立网站）
省级大数据局				
1	内蒙古大数据发展管理局	2017年1月	是推动内蒙古国家大数据综合试验区建设的主管部门，将数据交易纳入发展重点，推动自治区大数据流通交易平台建设。	<a href="https://nmgdata.org.cn/">https://nmgdata.org.cn/</a>
2	贵州省大数据发展管理局	2017年2月	贵州省原经济和信息化委员会承担的有关数据资源管理、大数据应用和产业发展（除电子信息制造业外）、信息化（除“两化融合”外）等职责，整合划入省大数据发展管理局。	<a href="http://dsj.guizhou.gov.cn/">http://dsj.guizhou.gov.cn/</a>
3	江西省大数据中心	2018年1月	将承担江西省大数据发展核心技术研发和标准制订、基础设施及应用的建设维护、公共数据汇聚共享和开放等工作。	<a href="http://ixic.jiangxi.gov.cn/">http://ixic.jiangxi.gov.cn/</a>
4	上海市大数据中心	2018年4月	构建全市数据资源共享体系，制定数据资源归集、治理、共享、开放、应用、安全等技术标准及管理办法，实现跨层级、跨部门、跨系统、跨业务的数据共享和交换。	<a href="http://wap.sh.gov.cn/nw2/nw2314/nw2319/nw32905/nw32914/nw32994/nw44523/nw44616/index.html">http://wap.sh.gov.cn/nw2/nw2314/nw2319/nw32905/nw32914/nw32994/nw44523/nw44616/index.html</a>
5	福建省大数据管理局	2018年10月	在新一轮机构改革中，设置省政府部门管理机构数字福建建设领导小组办公室，加挂省大数据管理局牌子。	<a href="http://fgw.fujian.gov.cn/">http://fgw.fujian.gov.cn/</a>
6	山东省大数据局	2018年10月	为主动适应数字信息技术的快速发展，解决部门信息“孤岛”和信息“烟囱”问题，加快推进“互联网+电子政务”，建设“数字山东”，在山东省政府办公厅大数据和电子政务等管理职责的基础上，组建省大数据局，作为省政府直属机构。	<a href="http://bdb.shandong.gov.cn/">http://bdb.shandong.gov.cn/</a>
7	浙江省大数据发展管理局	2018年10月	负责推进政府数字化转型和大数据资源管理等工作。通过组建这一机构，进一步加强互联网与政务服务的深度融合，统筹管理公共数据资源和电子政务，推进政府信息资源整合利用，打破信息孤岛、实现数据共享，进一步助推“最多跑一次”改革和政府数字化转型，加快推进数字浙江建设。	<a href="http://www.zj.gov.cn/">http://www.zj.gov.cn/</a>
8	广东省政务服务数据管理局	2018年10月	统筹推动“数字政府”建设，促进政务信息资源共享协同应用，提升政务服务能力。	<a href="http://zfsg.gd.gov.cn/">http://zfsg.gd.gov.cn/</a>
9	吉林省政务服务和数字化建设管理局	2018年10月	整合电子政务、大数据建设、营商环境优化等职责。该局作为吉林省政府直属机构，加挂吉林省软环境建设办公室牌子。	<a href="http://zsj.jl.gov.cn/">http://zsj.jl.gov.cn/</a>
10	广西壮族自治区大数据发展局	2018年11月	是广西壮族自治区机构改革新成立的直属机构，承担推进数字广西建设，统筹全区电子政务基础设施和重要信息系统建设，负责政府数据和社会数据采集、汇聚、管理，推进战略新兴产业发展等方面重要职责，整合相关部门信息化建设、政务服务监督管理等职能。	<a href="http://dsjfz.gxzf.gov.cn/">http://dsjfz.gxzf.gov.cn/</a>
11	河南省大数据管理局	2018年11月	是将河南省政府相关部门的电子政务规划建设指导、行政审批与便民服务建设、数据挖掘利用等职责整合组建的，作为省政府办公厅的部门管理机构。	<a href="https://dsj.henan.gov.cn/">https://dsj.henan.gov.cn/</a>
12	重庆市大数据应用发展管理局	2018年11月	大数据应用发展管理局是由市经济和信息化委员会的人工智能、大数据、信息化推进职责，市发展和改革委员会的社会公共信息资源整合与应用、智慧城市建设职责等整合。	<a href="http://dsjj.cq.gov.cn/">http://dsjj.cq.gov.cn/</a>
13	海南省大数据管理局	2019年5月	将承担我省大数据建设、管理和服务等职责，推进海南省大数据管理体制机制创新和大数据统一建设、统一管理、高效服务。	<a href="http://dsj.hainan.gov.cn/">http://dsj.hainan.gov.cn/</a>



# 上海市大数据中心

## ➤ 成立时间

2018年4月12日，上海市大数据中心正式揭牌。

## ➤ 主要职责

构建全市数据资源共享体系，制定数据资源归集、治理、共享、开放、应用、安全等技术标准及管理办法，实现跨层级、跨部门、跨系统、跨业务的数据共享和交换。

## ➤ 政校合作案例

2020年7月14日，上海市大数据中心与上海交通大学签署战略合作协议，围绕公共数据治理和一网通办，共同深化政产学研用协同创新。政校强强联合，将充分发挥双方数据平台优势、科研能力优势、跨界融合优势，形成聚合力量，激发数据活力，共同助力经济社会数字化转型。

## 数字经济

党的十九大报告中将推动互联网、大数据、人工智能与实体经济深度融合明确为我国建设现代化经济体系的重要着力点。

习近平总书记在全国网信工作会上要求**加快推动产业数字化，释放数字对经济发展的放大、叠加、倍增作用**。当前，我国数字经济发展基础良好，融合创新活跃，数字经济蓬勃发展为我国实体经济发展注入了新动力。







# 数字经济

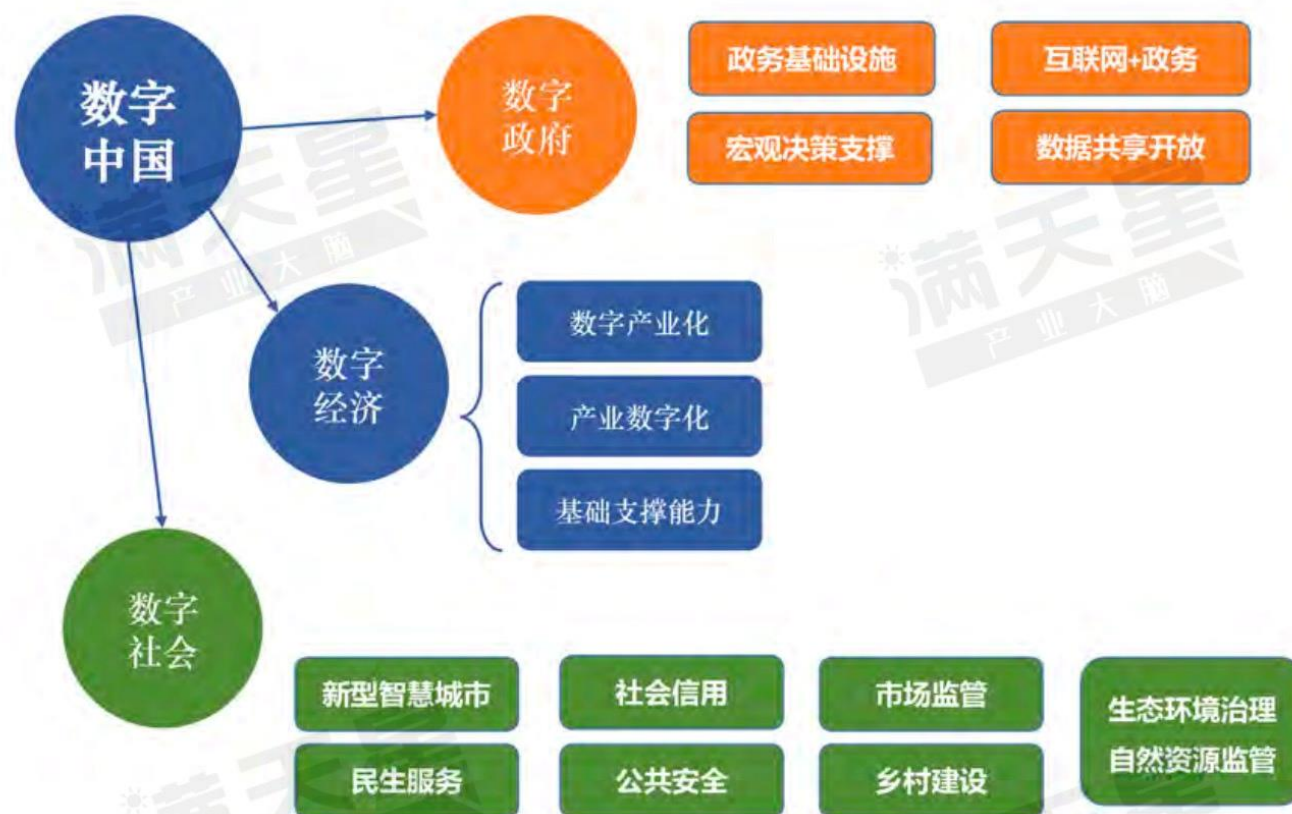
《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》 2020年11月3日新华社

## ➤ 加快数字化发展

发展数字经济，推进数字产业化和产业数字化，推动数字经济和实体经济深度融合，打造具有国际竞争力的数字产业集群。加强数字社会、数字政府建设，提升公共服务、社会治理等数字化智能化水平。建立数据资源产权、交易流通、跨境传输和安全保护等基础制度和标准规范，推动数据资源开发利用。扩大基础公共信息数据有序开放，建设国家数据统一共享开放平台。保障国家数据安全，加强个人信息保护。提升全民数字技能，实现信息服务全覆盖。积极参与数字领域国际规则和标准制定。

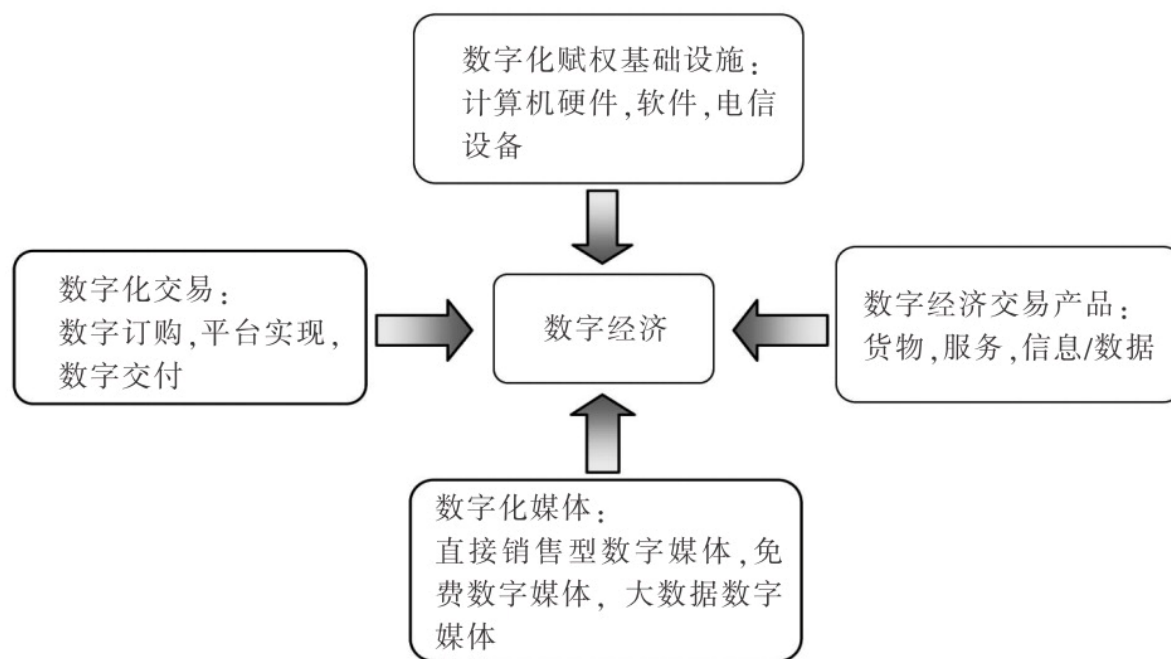
# 数字经济

## “数字中国”建设为大数据产业发展注入新功能



# 数字经济

伴随着信息技术的快速发展及其经济运行方式的不断融合，数字经济已被视为经济增长的“新引擎”，在世界上多数国家的发展战略中占据重要位置。



数字经济形成要素

许宪春、张美慧 《中国数字经济规模测算研究——基于国际比较的视角》



# 新基建

**新基建**一般是指新型基础设施建设。新基建是智慧经济时代贯彻新发展理念，吸收新科技革命成果，实现国家生态化、数字化、智能化、高速化、新旧动能转换与经济结构对称态，建立现代化经济体系的国家基本建设与基础设施建设。

**新基建七大领域**，主要包括5G基站建设、特高压、城际高速铁路和城市轨道交通、新能源汽车充电桩、**大数据中心**、人工智能、工业互联网七大领域，涉及诸多产业链，是以新发展理念为引领，以技术创新为驱动，以信息网络为基础，面向高质量发展需要，提供数字转型、智能升级、融合创新等服务的基础设施体系。



5G基站



特高压



城际高速铁路



城市轨道交通



充电桩



大数据



人工智能



工业互联网

# 数字经济



表 4

中国数字经济增加值及其占 GDP 的比重

年份	数字经济增加值(亿元)			数字经济总增加值 (亿元)	数字经济增加值 占 GDP 比重(%)
	数字化赋权基础设施	数字化交易	数字化媒体		
2007	12715.70	419.68	788.13	13923.52	5.16
2008	15205.89	524.80	942.75	16673.45	5.22
2009	15617.40	580.97	1077.13	17275.51	4.96
2010	17434.68	716.50	1297.01	19448.20	4.72
2011	19522.62	876.54	1506.89	21906.06	4.49
2012	21212.69	998.82	1755.27	23966.79	4.45
2013	25638.17	1399.10	1839.23	28876.50	4.87
2014	29311.81	1552.71	2048.11	32912.63	5.13
2015	33413.24	1646.31	2306.57	37366.13	5.45
2016	38028.17	1773.27	2586.62	42388.06	5.73
2017	47783.48	1931.65	3313.72	53028.85	6.46

中国数字经济增加值及其占GDP的比重

许宪春、张美慧 《中国数字经济规模测算研究——基于国际比较的视角》



## 数字经济

基于国际比较的视角，2017年，中国数字经济增加值约为美国的58.12%；数字经济增加值占GDP比重低于美国0.44个百分点；2016年，中国数字经济增加值约为美国的52.77%，占GDP比重低于美国0.77个百分点，略高于澳大利亚0.03个百分点。

近年来，中国数字经济增加值年均实际增长率明显高于美国和澳大利亚。2008-2017年，中国数字经济增加值年均实际增长率达14.43%，明显高于国内生产总值年均实际增长率8.27%，数字经济推动经济增长的作用明显。

# 数字经济



## 数字经济分类

1. 数字产品制造业
2. 数字产品服务业
3. 数字技术应用业
4. 数字要素驱动业
5. 数字化效率提升业



国家统计局，《数字经济及其核心产业统计分类（2021）》








# 数字经济



## 三大城市群先行先试，区域一体带动产业经济

### 布局 新征程

#### 北京:

-  以创新链带动产业链供应链
-  推进文化产业发展引领区建设
-  加快建设“三城一区”主平台和中关村示范区主阵地
-  加强京雄产业创新协同
-  建设全球数字经济标杆城市
-  打造高精尖产业，培育万亿产业集群
-  优化产业空间布局，强化“两带”联动效应

**天津:** 加快构建现代工业产业体系

**河北:** 共建京津冀协同发展新格局







# 数字经济


## 三大城市群先行先试，区域一体带动产业经济

### 布局 新征程

 建好“长三角科技创新园”，为长三角一体化发展提供动力支撑

 推动科技基础设施集群化发展

 开展关键核心技术协同攻关

 开展产业链补链固链强链行动

 多维度共建长三角国家技术创新中心

 打造全国数字经济创新高地

**上海:** 加快推动经济高质量发展，同长三角区域产业集群加强分工协作

**江苏:** 着眼于在长三角区域的重要地位，大力推进产业创新等“六个一体化”

**浙江:** 共建“数字长三角”，深入实施数字经济“一号工程2.0版”

**安徽:** 合力建设长三角科技创新共同体




# 数字经济



三大城市群先行先试，区域一体带动产业经济

## 布局 新征程

### 广东：

-  广东培育发展战略性新兴产业集群和战略性新兴产业集群
-  推动制造业高质量发展，加快发展现代产业
-  加快构建“一核一带一区”区域发展格局，培育世界级先进制造业集群

**香港：**以粤港澳大湾区发展为切入点，积极成为国内循环的参与者和国际循环的促成者

**澳门：**积极培育新兴产业，推动产业升级转型





---

# PART 二

# 大数据介绍

---

# 大数据定义发展时间轴

## ➤ 2004年

“大数据”这个术语最早期的引用可追溯到apache org的开源项目Nutch。当时，大数据用来描述为更新网络搜索索引需要同时进行批量处理或分析的大量数据集。

## ➤ 2010年

Big Data is not a precise term; rather, it is a characterization of the never-ending accumulation of all kinds of data, most of it unstructured. It describes.

大数据不是一个精确的术语，它是对**无穷尽的各种类型数据积累的描述**。（参考：Making sense of Big Data [M]. Technology forecast, A quarterly journal, 2010）



# 大数据定义发展时间轴

## ➤ 2011年

美国咨询公司麦肯锡在其研究大数据的报告中定义：大小超过常规的数据库工具获取、存储、管理和分析能力的数据集。（参考：Big data: The next frontier for innovation, competition, and productivity）

## ➤ 2012年

美国研究机构定义到：大数据是基于新的处理模式而产生的具有强大的决策力、洞察力以及流程优化能力的多样性的、海量的且增长率高的**信息资产**。（参考：Hype Cycle for Big Data, 2012）



McKinsey Global Institute

大数据：下一个创新、竞争、生产的先驱

May 2011

Big data: The next frontier for innovation, competition, and productivity

James Manyika  
Michael Chui  
Brad Brown  
Jacques Bughin  
Richard Dobbs  
Charles Finkelstein  
Angela Hung Berts

## 大数据定义发展时间轴

### ➤ 2012年

在维克托·迈尔·舍恩伯格及肯尼斯·库克耶的《大数据时代》中，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。（参考：Big Data: A Revolution That Will Transform How We Live, Work, and Think）

### ➤ 2014年

维基百科中定义到：“巨量资料（Big Data），或称大数据，指所涉及的资料量**规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理**成为帮助企业经营决策更积极的资讯”。



## 基于大数据特征的定义

对于大数据尚未有一个公认的定义，然而大数据不同的定义基本是从大数据的特征出发，通过这些特征的阐述和归纳试图给出其定义。

### ➤ 3V 定义

✓ **规模性** Volume

✓ **多样性** Variety

✓ **高速性** Velocity

### ➤ 4V 定义

✓ **价值性** Value/ **真实性** Veracity





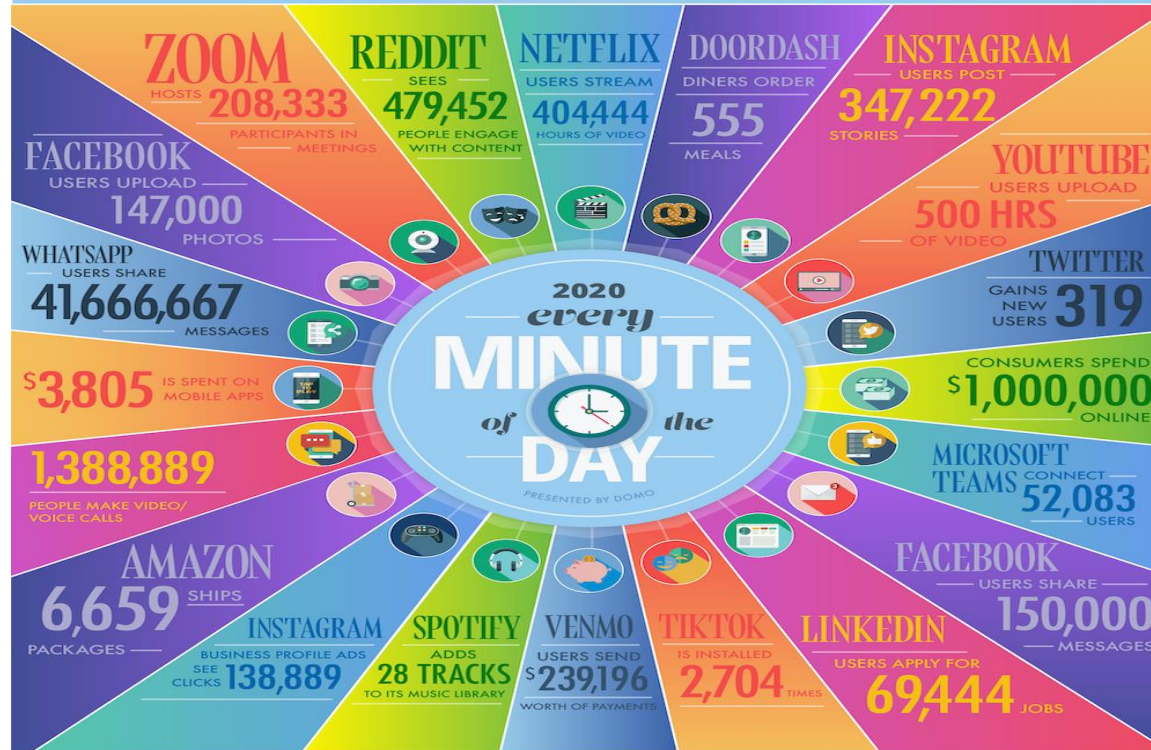
# 数字时代: DATA, DATA, DATA...



## DATA NEVER SLEEPS 8.0

How much data is generated *every minute*?

In 2020, the world changed fundamentally—and so did the data that makes the world go round. As COVID-19 swept the globe, nearly every aspect of life—from work to working out—moved online, and people depended more and more on apps and the Internet to socialize, educate and entertain ourselves. Before quarantine, just 15% of Americans worked from home. Now over half do. And that's not the only big shift. In our 8th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute—a trend that shows no sign of stopping.



The world's Internet population is growing significantly year over year. As of April 2020, the internet reaches 59% of the world's population and now represents 4.57 billion people — a 6% increase from January 2019.



GLOBAL INTERNET POPULATION GROWTH 2014–2020  
(IN BILLIONS)

As the world changes, businesses need to change with the times—and that requires data. Every click, swipe, share or like tells you something about your customers and what they want, and Domo is here to help your business make sense of all of it. Domo gives you the power to make data-driven decisions at any moment, on any device, so you can make smart choices in a rapidly changing world.

Learn more at [domo.com](https://domo.com)

SOURCES: STATISTA, VISUAL CAPITALIST, BUSINESS INSIDER, GAMESPOT, TECHCRUNCH, OMNICORE AGENCY, DOORDASH, BUSINESS OF APPS, NEW YORK TIMES, MUSIC BUSINESS WORLDWIDE, INC., THE VERGE, INC., HOOTSUITE, DUSTIN STOUT, REDDIT, USER, AMAZON, VIX



## 大数据已经渗透生活

- 数据科学时代改变了人们对数据价值的认识，不再认为数据是无价值的，也不再简单认为数据是死的、被动的东西，而是更加重视数据的积极作用。大数据在各个领域得到了充分的应用。
- “数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。” ——麦肯锡





# 大数据——最具价值资源

## 大数据革命 “Big Data Revolution”

### ➤ 数字经济时代

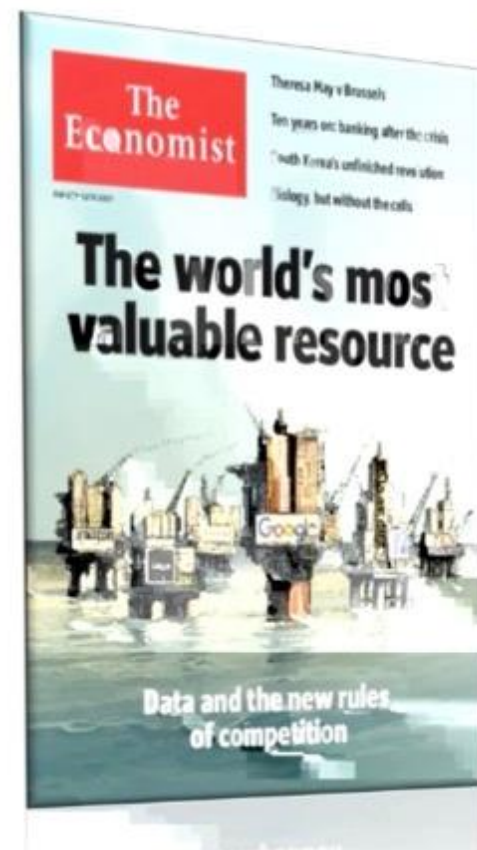
- ✓ 数据是关键的生产要素
- ✓ 新的GDP概念: Gross Data Product

大数据和人工智能（尤其是机器学习）正在改变人类生产与生活方式，也在改变经济学研究范式。

### ➤ 万物皆可互联

- ✓ 所有经济现象均可用数据来描述
- ✓ 一切皆可编程

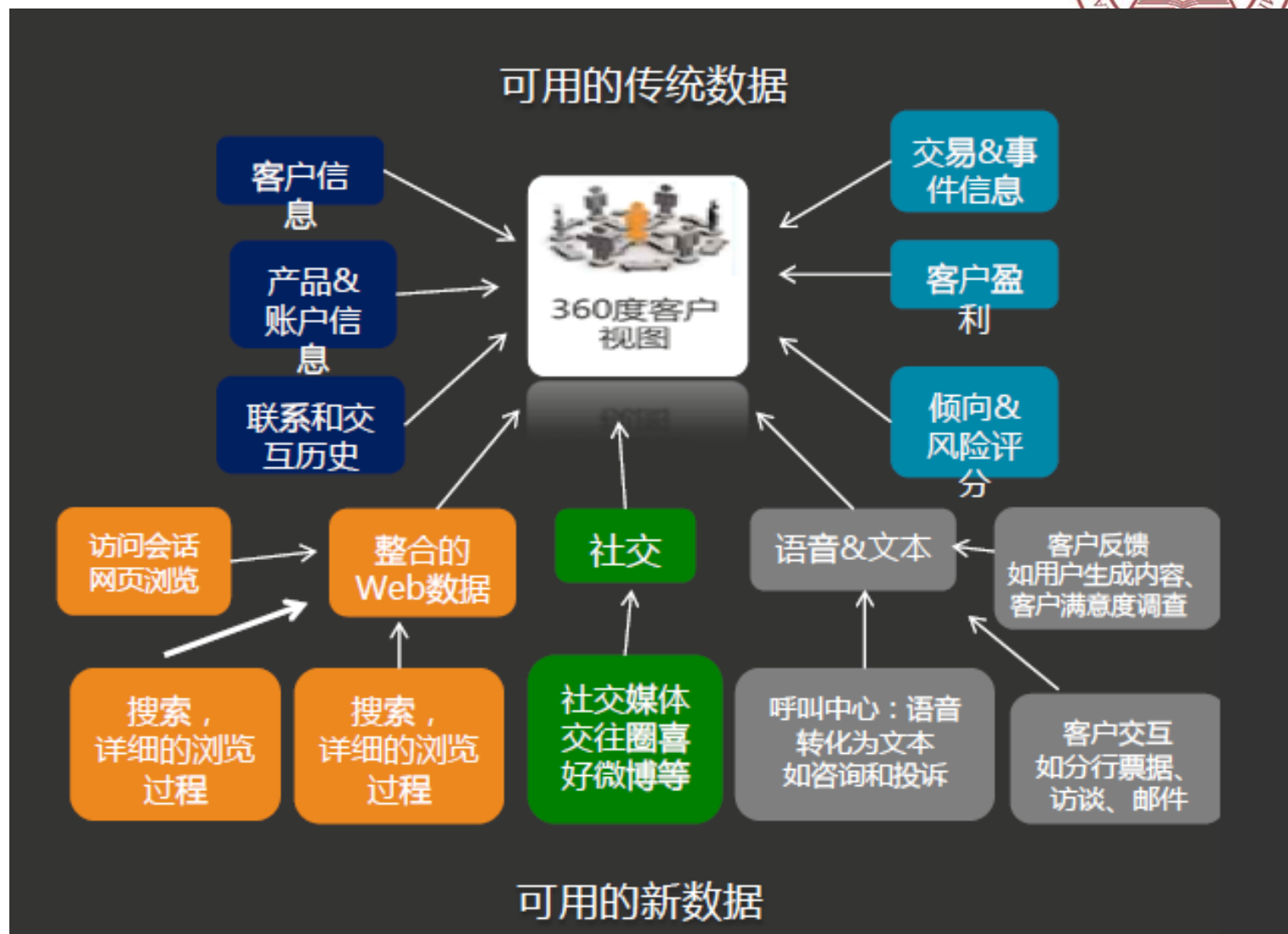
大数据为实证研究提供丰富的素材，大数据正在改变经济学、社会科学的研究范式。



# 金融业的大数据



“不要将大数据作为独立的课题来考虑，多考虑一下目前企业的核心业务问题是什么，如何结合大数据分析来解决这些问题”  
- Facebook 大数据团队前主管

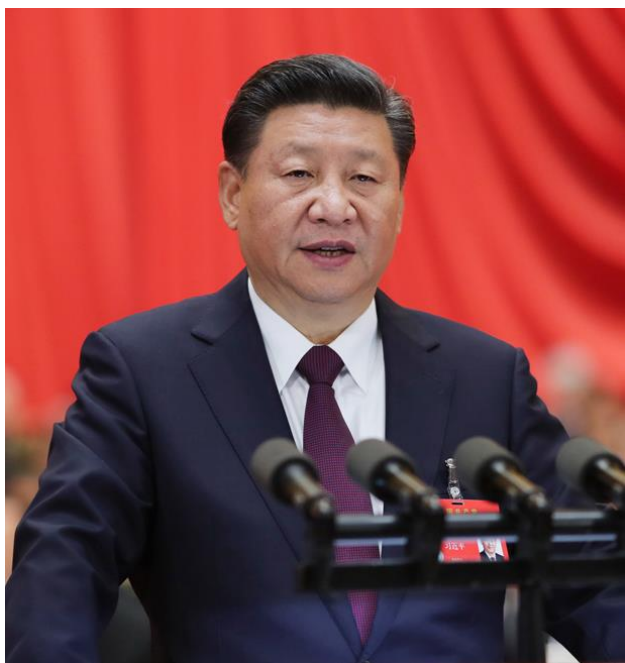


# 数据量和复杂度不断增加，如何挖掘数据价值？



## “5·17” 讲话

### 习近平：哲学社会科学工作座谈会上的“5·17”讲话



“对现代社会科学积累的有益知识体系，运用的**模型推演**、**数量分析**等有效手段，我们也可以用，而且应该好好用。需要注意的是，在采用这些知识和方法时不要忘了老祖宗，不要失去了科学判断力。马克思写的《资本论》、列宁写的《帝国主义论》、毛泽东同志写的系统调查报告等著作，都运用了大量**统计数字**和**田野调查**材料。”

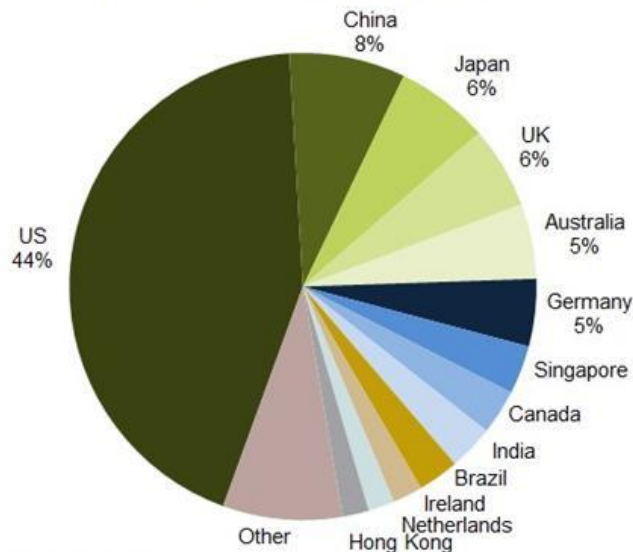
# 大数据运用的紧迫性

## 大数据分析处理的紧迫性



据IDC统计数据(2015年)显示, 中国目前拥有的数据量占全球的**14%**, 但数据利用率不到**0.4%**, 大量的数据“沉睡”在各个角落。

Hyperscale Data Center Operators  
Data Center Locations by Country - December 2017



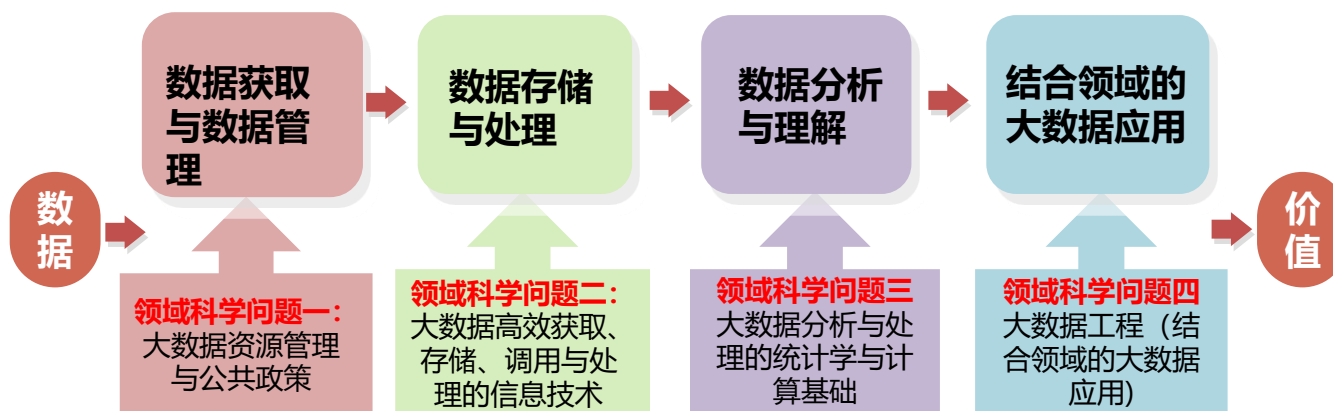
Source: Synergy Research Group



# 大数据的核心

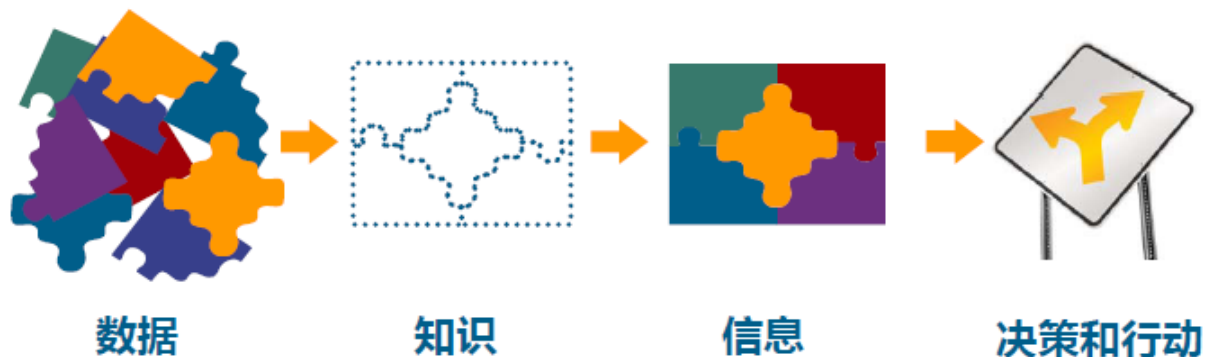
在大数据技术中，分析与处理是核心  
(MIT Technology Review, 2015)

- 大数据技术需要多学科综合研究



数据是基础、平台是支撑、**分析是核心**、效益是根本

# 大数据和大数据挖掘



分析数据  
建立模型  
分析建模人员

模型  
已发现的知识

把模型用在实际数据上  
预测未来的行为  
业务用户

**发现:**

找出隐藏在数据背后的模式，  
这些模式能把数据转化为知识

**部署:**

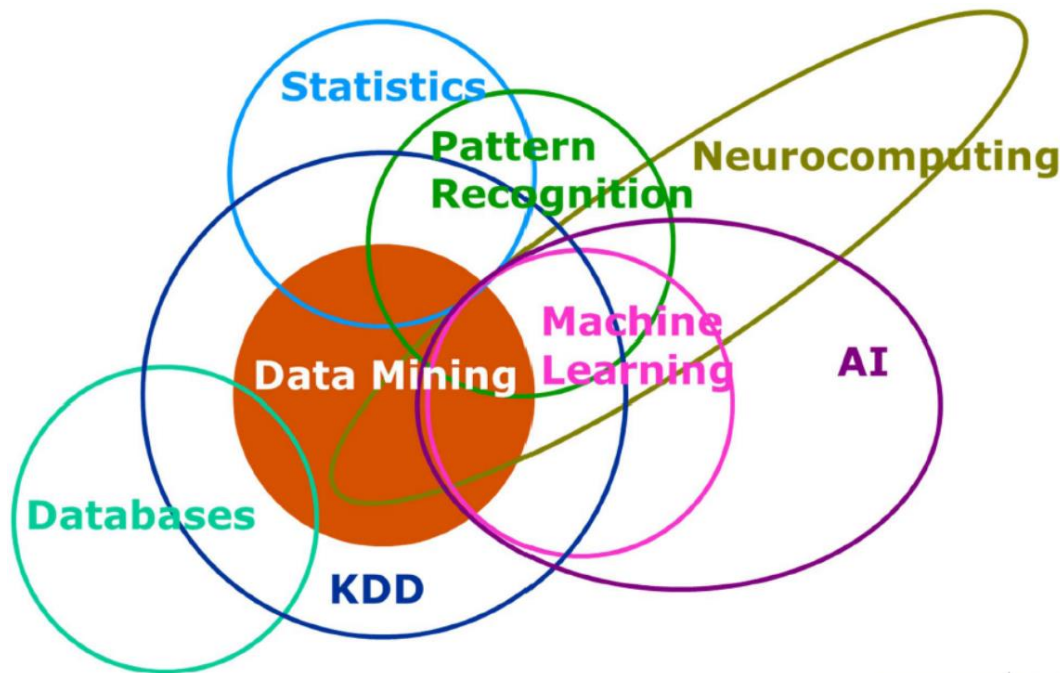
应用已发现的知识达成实用的目的—  
例如：预测

Data Mining is a process of discovering and interpreting patterns in data to solve problems

数据挖掘是一个发现和解释数据中的模式，并用于解决问题的过程

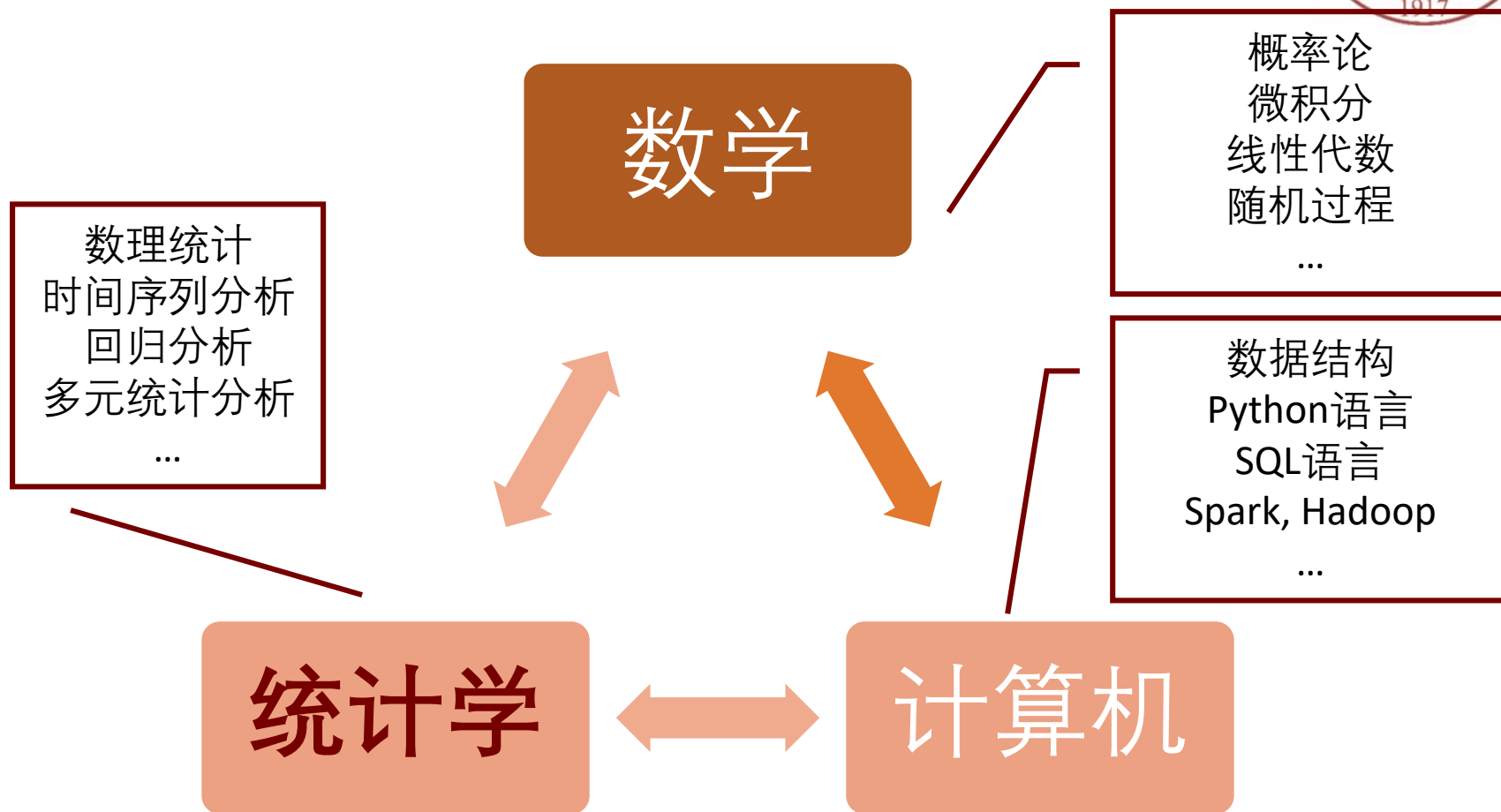
# 大数据和数据挖掘

数据挖掘融合了数据库、人工智能、机器学习、统计学、高性能计算、模式识别、神经网络、数据可视化、信息检索和空间数据分析等多个领域的和技术，是21世纪初期对人类产生重大影响的十大新兴技术之一。





# 大数据和数据挖掘



## 萨金特：人工智能其实就是统计学



托马斯·萨金特 (Thomas J. Sargent)

诺贝尔经济学奖得主、纽约大学教授、北京大学名誉教授、北京大学汇丰商学院“萨金特数量经济与金融研究所”所长

2011年诺贝尔经济学奖获得者Thomas J. Sargent在“共享全球智慧 引领未来科技”的世界科技创新论坛上发言表示，**人工智能其实就是统计学**，只不过用了一个很华丽的辞藻。



# 大数据人才培养



## 大数据相关专业人才培养现状

### 2015年8月国发[2015]50号文《促进大数据发展行动纲要》

- 提出“……建立健全多层次、多类型的大数据人才培养体系。鼓励高校设立数据科学和数据工程相关专业，重点培养专业化数据工程师等大数据专业人才。鼓励采取跨校联合培养等方式开展跨学科大数据综合型人才培养，大力培养具有统计分析、计算机技术、经济管理等多学科知识的跨界复合型人才。”



# 大数据人才培养



## 大数据专业迅猛发展

全国474所高校共设立了488个“数据科学与大数据技术”专业分布地图







## 上财数据科学人才培养

2017年8月25日 星期五 上海 34°C~26°C 雷阵雨 详细»



上海教育新闻网  
www.shedunews.com

新闻 图片 视频 学校 资料 百度

搜索

【手机报】【电子报】

新闻热线: 021-52372392

投稿邮箱: shedunews@126.com

首页

资讯

人物

时评

招考

视频

专题

活动

名片

微博

### 上财数据科学人才为何成就 业市场“香饽饽”？

作为大数据人才培养的先行尝试，上财第一届“数据科学与商务统计”应用统计专业硕士今年6月正式毕业，该专业毕业生在毕业前几个月既已实现百分百就业，平均签约月薪高达五位数。[详细]

«

1 2 3 4 5

»

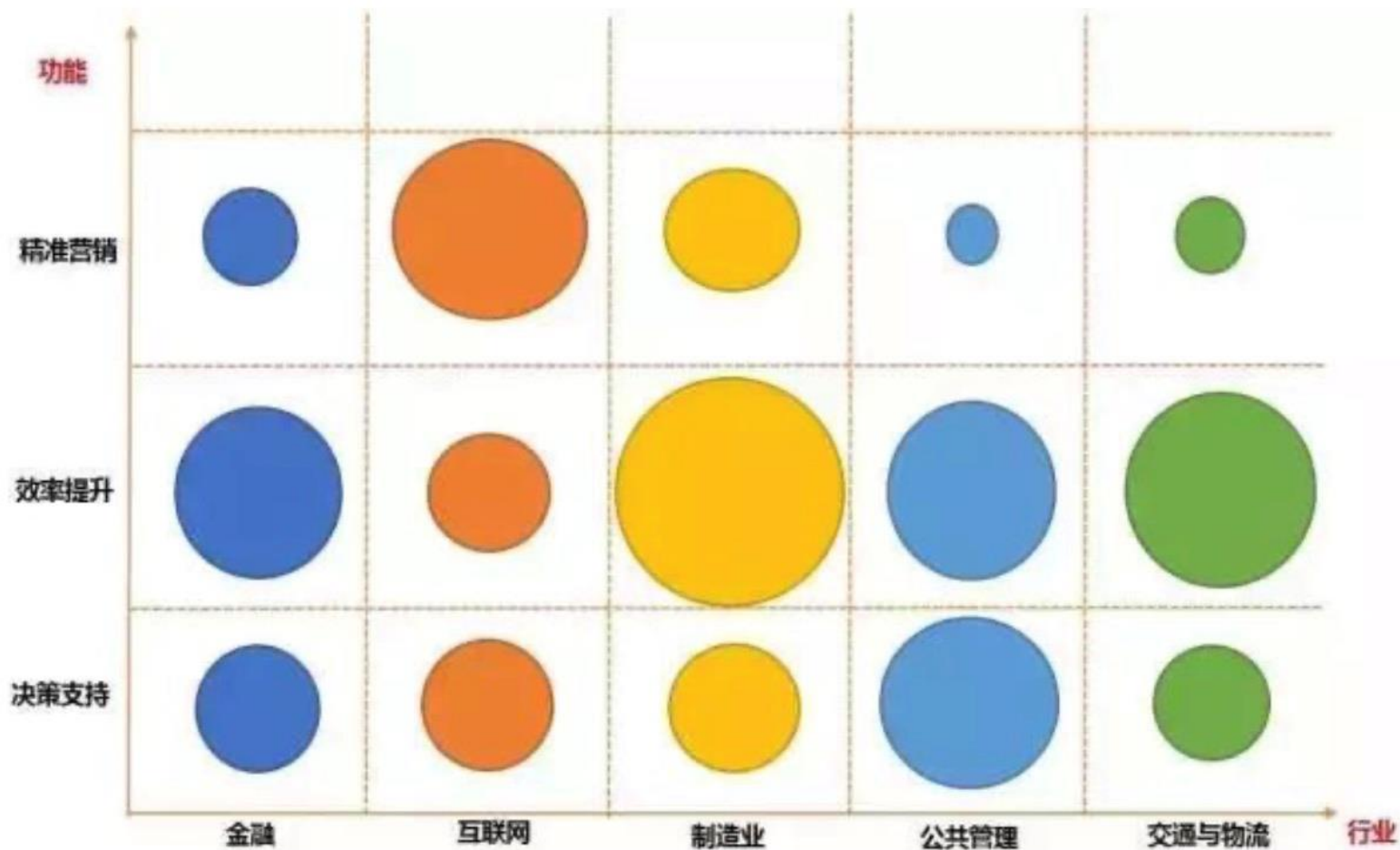
图说教育



# 2017上海新高考



# 大数据的行业应用





# 大数据的行业应用

大数据无处不在，大数据应用于各个行业，包括金融、汽车、餐饮、电信、能源、体育和娱乐等在内的社会各行各业都已融入大数据的印迹。

制造  
业

利用工业大数据提升制造业水平，包括产品故障诊断与预测、分析工艺流程、改进生产工艺，优化生产过程能耗、工业供应链分析与优化等。

金融  
行业

大数据在高频交易、社交情绪分析和信贷风险分析三大金融创新领域发挥重大作用。

互联  
网

借助大数据技术，可以分析客户行为，进行商品推荐和针对性广告投放。





# 大数据的行业应用

大数据无处不在，大数据应用于各个行业，包括金融、汽车、餐饮、电信、能源、体育和娱乐等在内的社会各行各业都已融入大数据的印迹。

## 电信行业

- 利用大数据技术实现客户离网分析，及时掌握流失倾向、开展干预。

## 物流行业

- 利用大数据优化物流网络，提高物流效率，降低物流成本。

## 安全领域

- 政府可以利用大数据技术构建起强大的国家安全保障体系，企业可以利用大数据抵御网络攻击，警察可以借助大数据来预防犯罪。

# 大数据和机器学习

## 机器学习的本质

- 机器学习主要用**预测** (prediction) 和**分类** (classification)
- 利用计算机算法程序自动分析大数据，基于大数据进行样本外预测（包括分类预测）
- 可表示为一个**数学优化**问题，通过设计计算机算法程序来自动实施
- 关键：防止**过度拟合** (overfitting)





# 大数据和机器学习

## 机器学习的应用

- 在实际应用中，很多情况下机器学习预测是比较准确的，这种预测**并不是基于经济因果关系**，而是通过挖掘**数据的系统特征与变量之间的统计关系**，比如说相关性或预测关系。
- 经济中的应用实例
  - ✓ 算法交易（80%，外汇市场）
  - ✓ 商务智能（Business Intelligence, BI）
  - ✓ 机器学习改变商业模式（如信用卡、小额贷款审批）
  - ✓ 自动驾驶（Autonomous Vehicle）



---

**PART**  
**三**

# 大数据与国内外经济发展

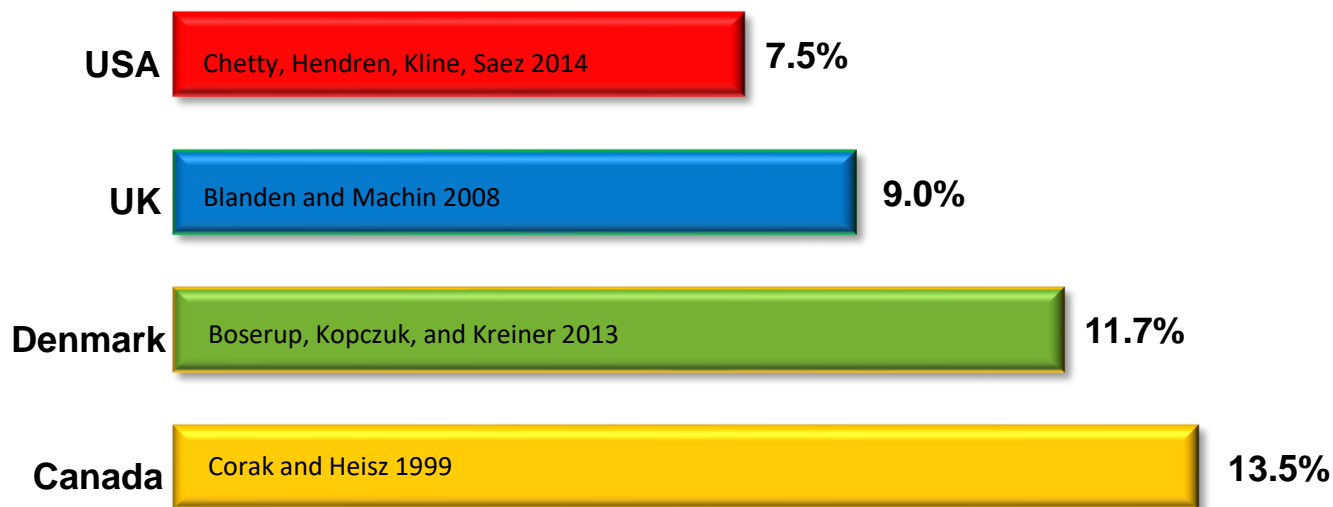
---



## 案例1

## 税收大数据与消逝的“美国梦”

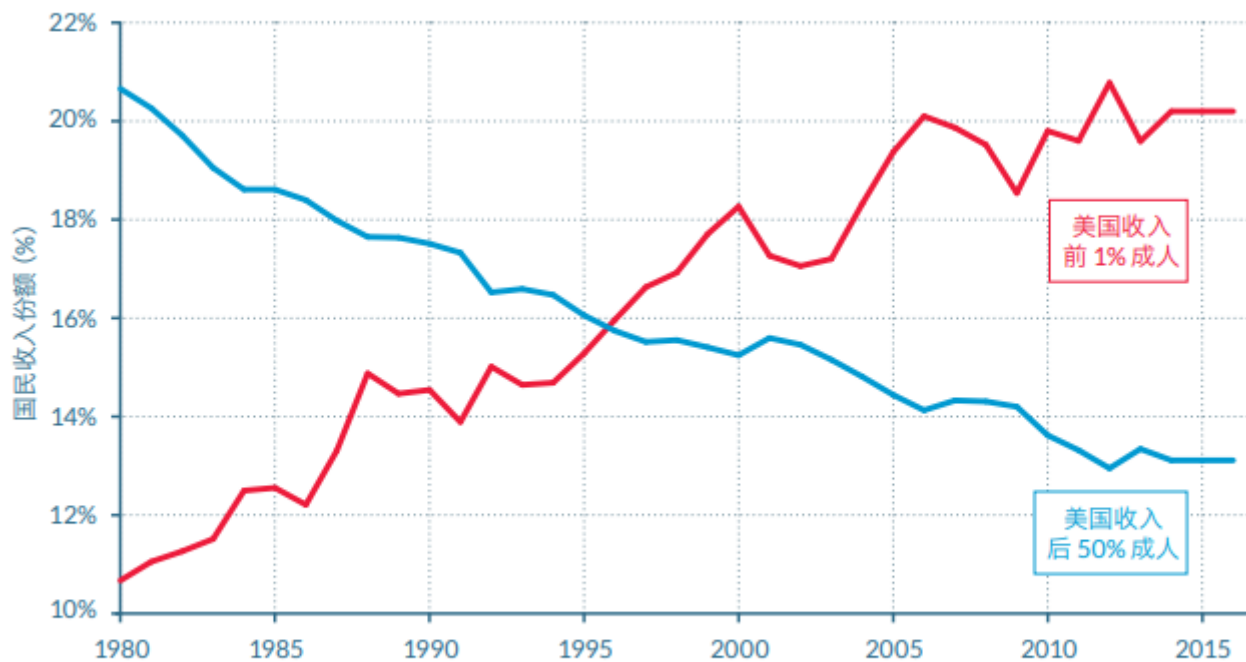
- 生于低收入家庭（处于收入分布底层的20%家庭）的孩子成年后进入高收入阶层（收入分布顶层的20%家庭）的可能性:



- 加拿大阶层向上流动的可能性是美国的2倍
- 美国贫穷人家的孩子，摆脱贫困的机会为何渺茫？

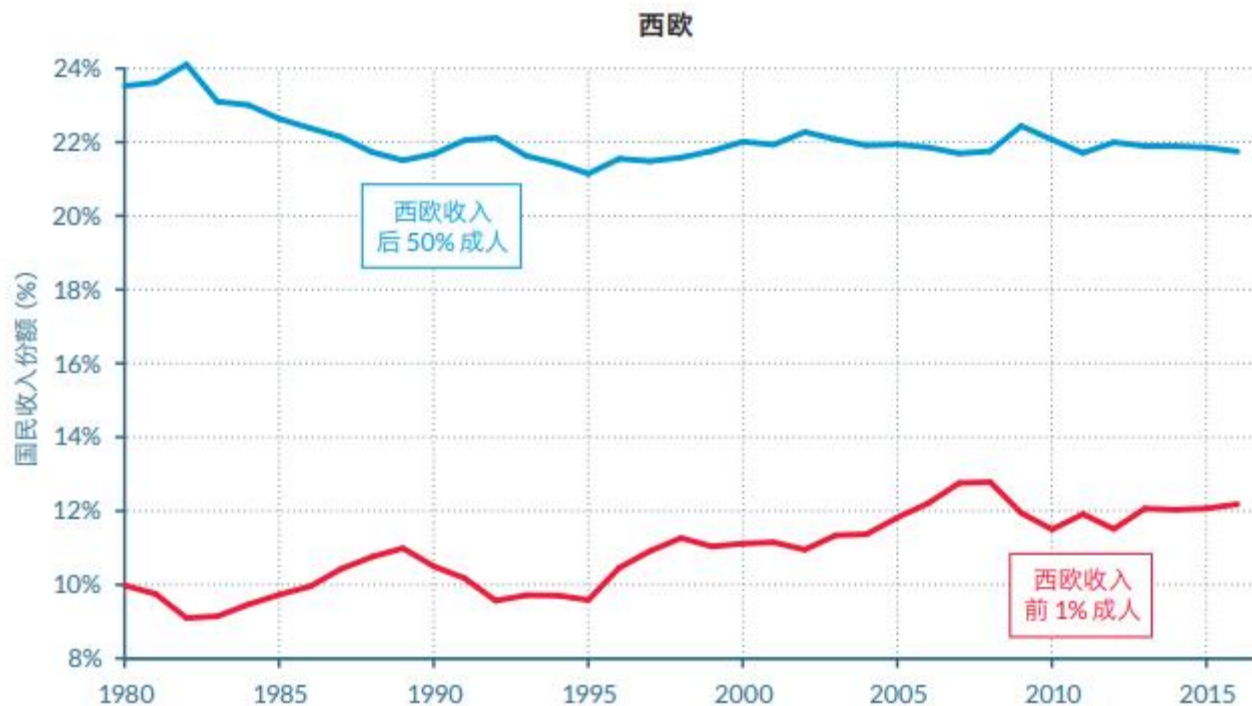


美国



时间财富与收入不平均数据库 (WID.world), 世界不平均报告 2018 (wir2018.wid.world)

2016 年, 收入前 1% 成人的收入份额在欧洲为 12%, 而在美国则超过 20%。

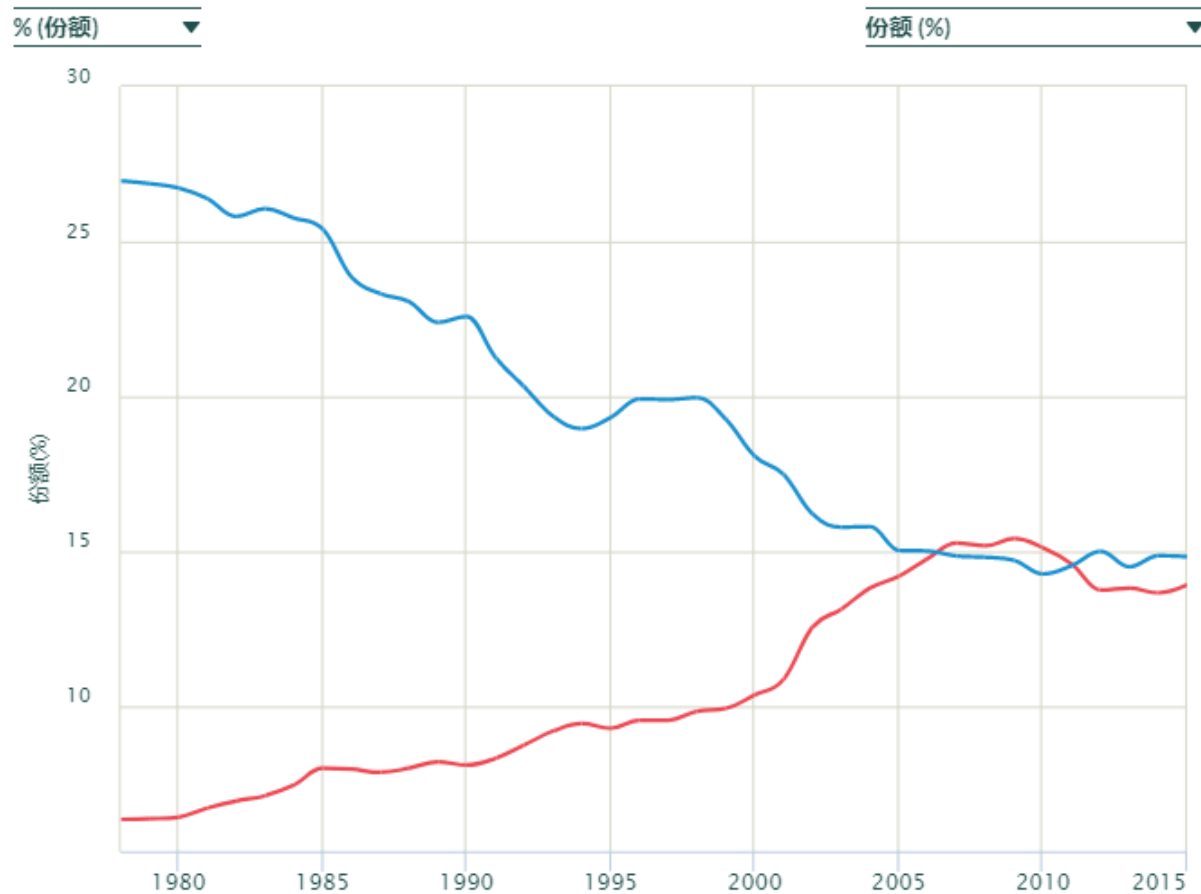


时间财富与收入不平均数据库 (WID.world), 世界不平均报告 2018 (wir2018.wid.world)

2016 年, 收入前 1% 成人的收入份额在欧洲为 12%, 而在美国则超过 20%。

1980 年, 收入前 1% 成人的收入份额在欧洲为 10%, 而在美国则超过 11%。

## 收入不平等, 中国, 1978-2015





## ➤ 中国社会阶层流动性有多大？

✓ 莫衷一是，不同数据不同答案

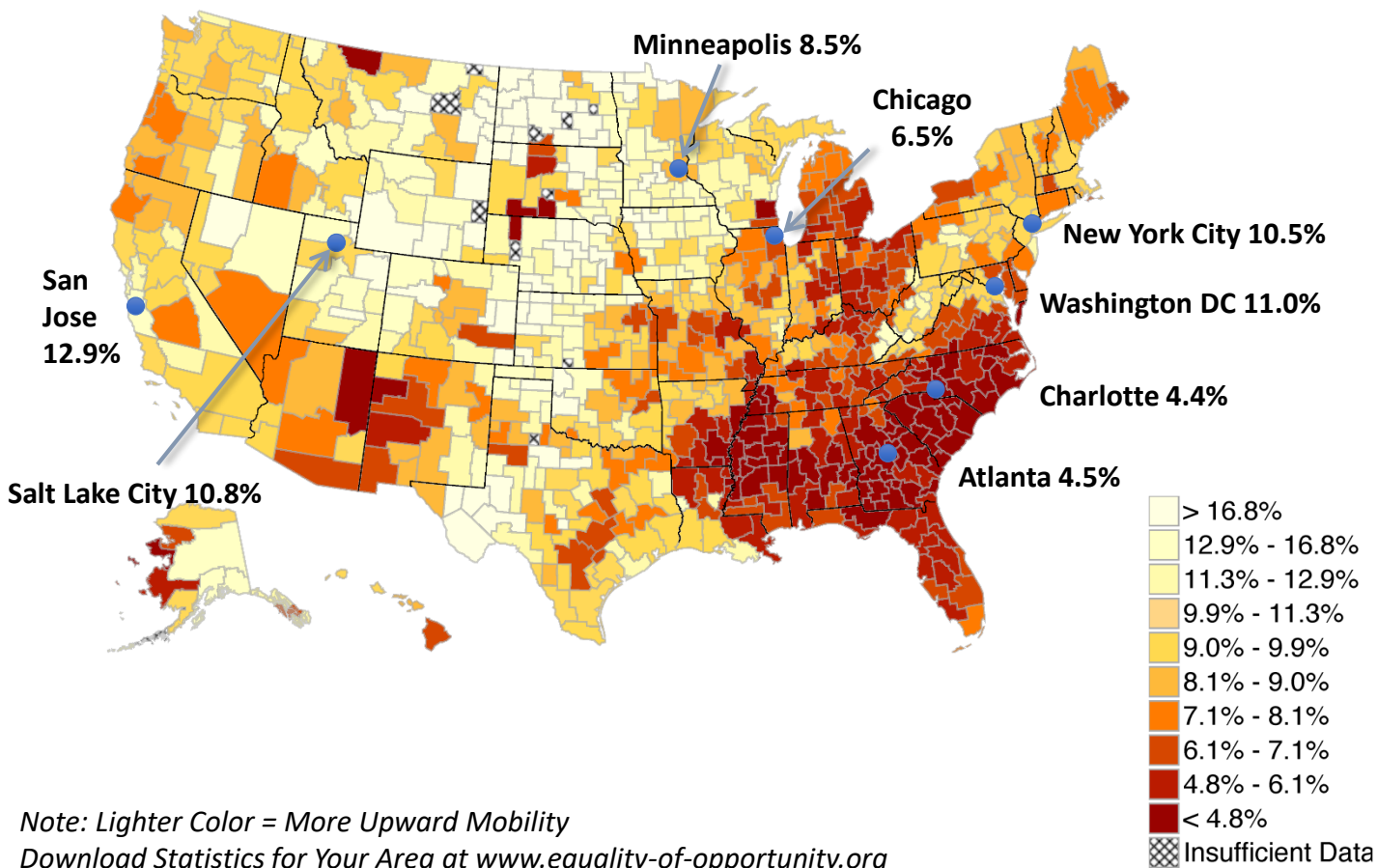
## ➤ 美国1980-1982年出生的1000万儿童所在家庭的脱敏 个税税收大数据

✓ 计算：（1）出生家庭在全国收入分布中的位置

✓ （2）儿童成年后自身所处的收入分布位置

## The Geography of Upward Mobility in the United States

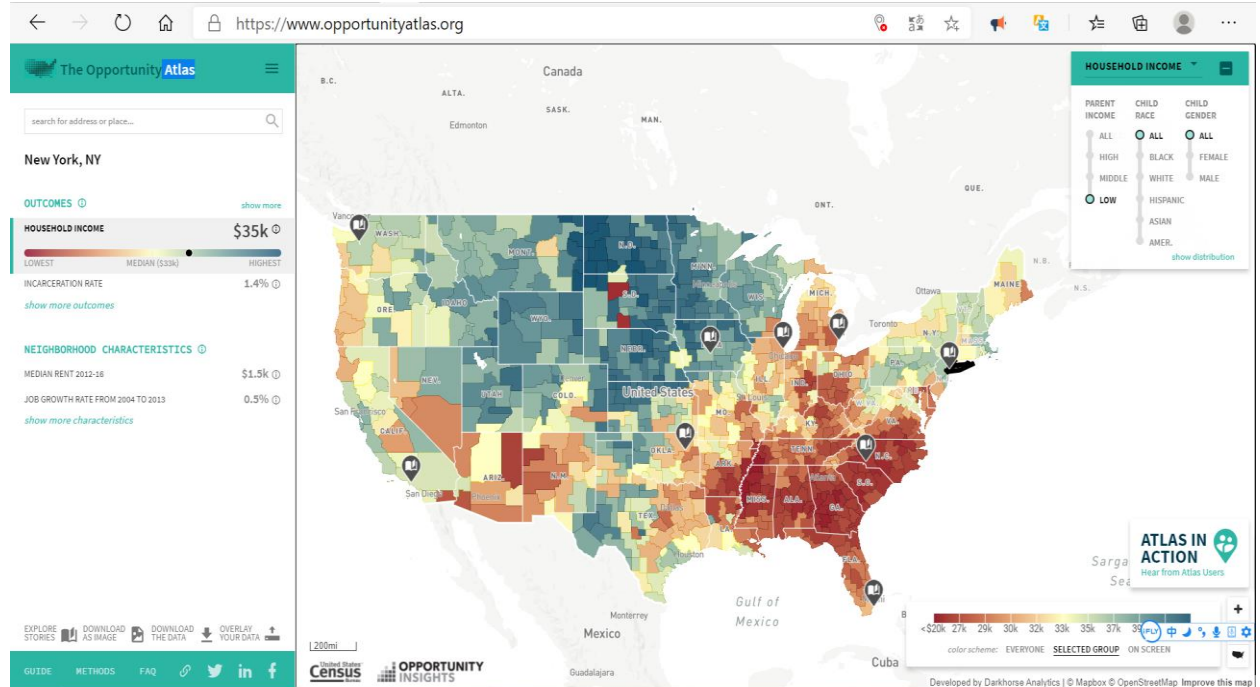
Chances of Reaching the Top Fifth Starting from the Bottom Fifth by Metro Area



## ➤ 社会流动性相关因素:

- ✓ 种族隔离
- ✓ 中小学教育质量
- ✓ 收入不平等
- ✓ 家庭结构
- ✓ 社会资本

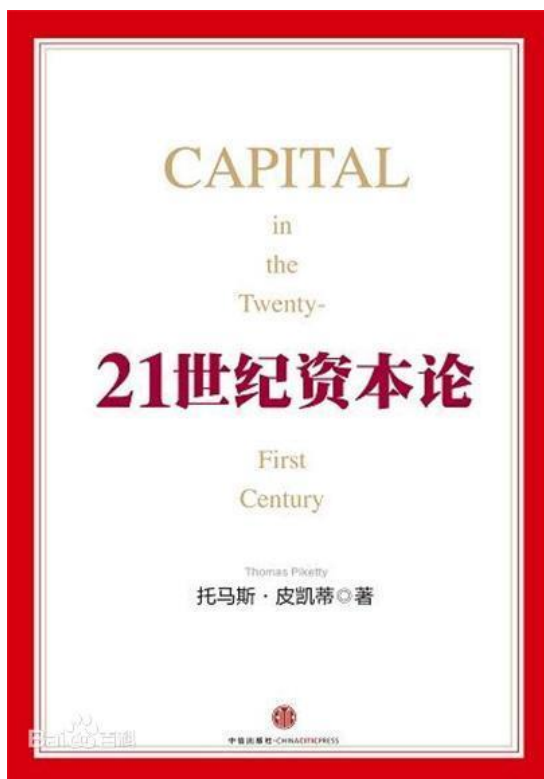
[opportunityatlas.org/](https://www.opportunityatlas.org/)





## 皮凯蒂 (2014) 《21世纪资本论》

- 习近平总书记评价皮凯蒂 (2014) 《21世纪资本论》：



“他用**翔实的数据**证明，美国等西方国家的不平等程度已经达到或超过了历史最高水平，认为不加制约的资本主义加剧了财富不平衡等现象，而且将继续恶化下去，他的分析主要是从分配领域进行的，没有过多涉及更根本的所有制问题，但得出的结论值得我们深思。”



## 皮凯蒂 (2014) 《21世纪资本论》

### ➤ 皮凯蒂 (2014) 《21世纪资本论》研究方法: **实证研究**

- ✓ 通过整合包括国民账户、微观调查数据、税收数据、商业银行报告在内的各种数据，对收入与财富分配进行跨国比较分析。
- ✓ 针对**基尼系数**的缺陷，他主张用**分配表**分析收入不平等的问题，因为分配表能够显示出收入与财富金字塔各自最顶端的10%和1%群体在总收入与总财富的比值。
- ✓ 估算了近300年来多个国家的**资本劳动收入比**，发现进入21世纪以来，收入与分配不平等呈现出与19世纪相似的两极分化情形。
- ✓ 其结论冲击了西方主流经济学关于社会收入和财富分配将随着经济的充分发展而逐步趋于平稳的观点，特别是诺贝尔经济学奖得主西蒙·库兹涅茨 (Simon Kuznets) 的“**倒U型曲线理论**”。



## 皮凯蒂 (2014) 《21世纪资本论》

- 皮凯蒂 (2014) 《21世纪资本论》研究方法: **实证研究**
  - ✓ 英国记者Simon Kuper在2020年对皮凯蒂的采访
  - 皮凯蒂: “我思考的**主要独到之处**是, 我的分析可以**基于历史直到今天的时间序列数据**。”
  - 皮凯蒂的数据主要源自他与同时共同创建的**世界不平等数据库** (World Inequality Database, WID) ,这个数据库推动了不平等经济学整个领域的发展。





## 案例2

## 交易大数据与疫情后经济复苏

- 2014年10月，蚂蚁金服正式成立；
- 致力于打造开放的生态系统，为小微企业和个人消费者提供普惠金融服务；
- 蚂蚁金服旗下有支付宝、余额宝、招财宝、蚂蚁聚宝、网商银行、蚂蚁花呗、芝麻信用、蚂蚁金融云、蚂蚁达客等子业务板块；
- **“码商”**——小微商家只要一张“二维码”就能做生意，除基本的收钱功能，还可以享受到经营分析、财务管理、理财保险、贷款等多维的服务。

## 案例2

### 交易大数据与疫情后经济复苏

- ✓ 基于蚂蚁金服集团旗下的支付产品码商的数据
- ✓ 测算了中国线下**个体经营户的整体规模**、新冠疫情对于这一群体的**短期冲击**以及复苏情况

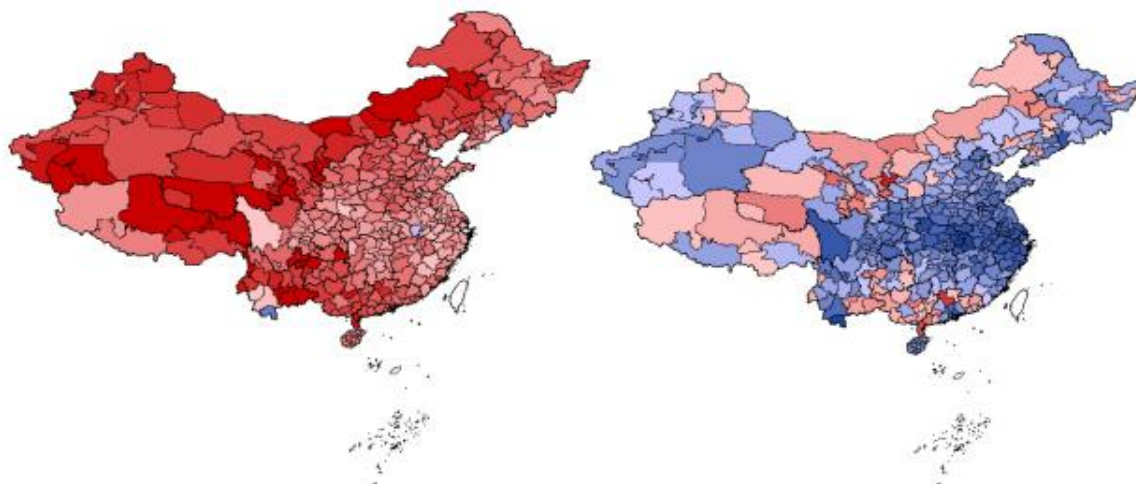


图 1：2020 年春节前后“码商”营业额同比增幅对比

注：①左图为 2020 年腊月廿六前两周码商营业额与 2019 年农历同期的同比变化，右图为 2020 年正月初七后两周的同比变化；②红色代表增长，蓝色代表下降，颜色越深，增长（下降）幅度越大；③港澳台地区数据缺失。

北京大学数字金融研究中心报告《“烟火重聚”：中国个体经营户新冠疫情下的复苏》

## 案例2

## 交易大数据与疫情后经济复苏

- ✓ 随着中国本地疫情传播得到基本控制，中央开始大力推动复工复产，各地也陆续开始解除一级响应。
- ✓ 个体经营的恢复速度非常快，到3月中就已经恢复到估算的反事实结果的80%。

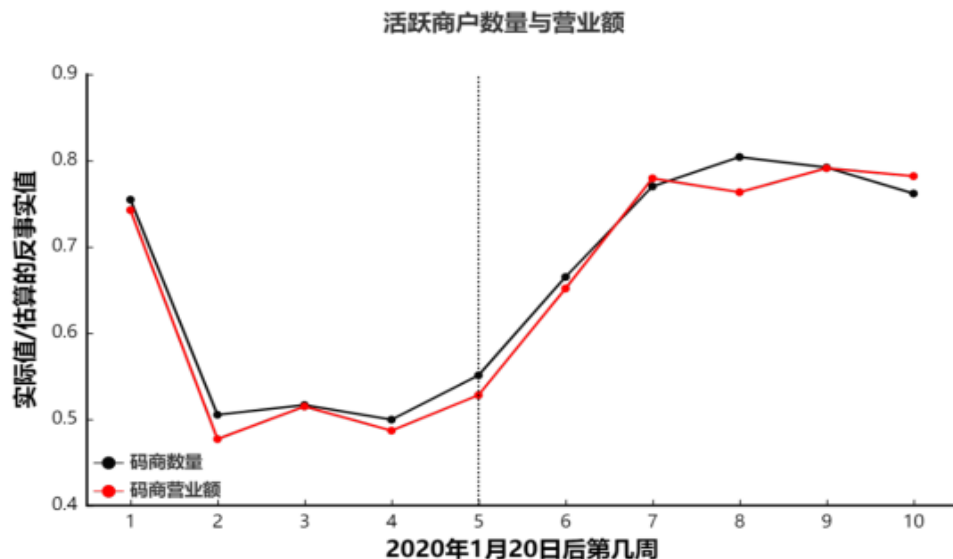


图 4:码商活跃商户量与营业额的受冲击程度和恢复进度

注:①横坐标为1月20日后的第N周,其中第1周包含了1月21日至1月30日10天,其他均为7天的窗口;②纵坐标的数字2020年码商活跃商户量(交易额)与机器学习算法预测出的反事实结果的比值;③虚线代表1月21-2月27日,这一周开始,各地逐渐解除一级响应。

北京大学数字金融研究中心报告《“烟火重聚”:中国个体经营户新冠疫情下的复苏》



## 案例2

## 交易大数据与疫情后经济复苏

运用爬虫技术从“大众点评”网站获取9个城市、1075家餐馆，从2019年1月到2020年7月的89万多条点评数据。通过文本挖掘和机器学习模型，测算餐馆存活率、营业恢复时间、客流量等信息



2月词云



3月词云



4月词云



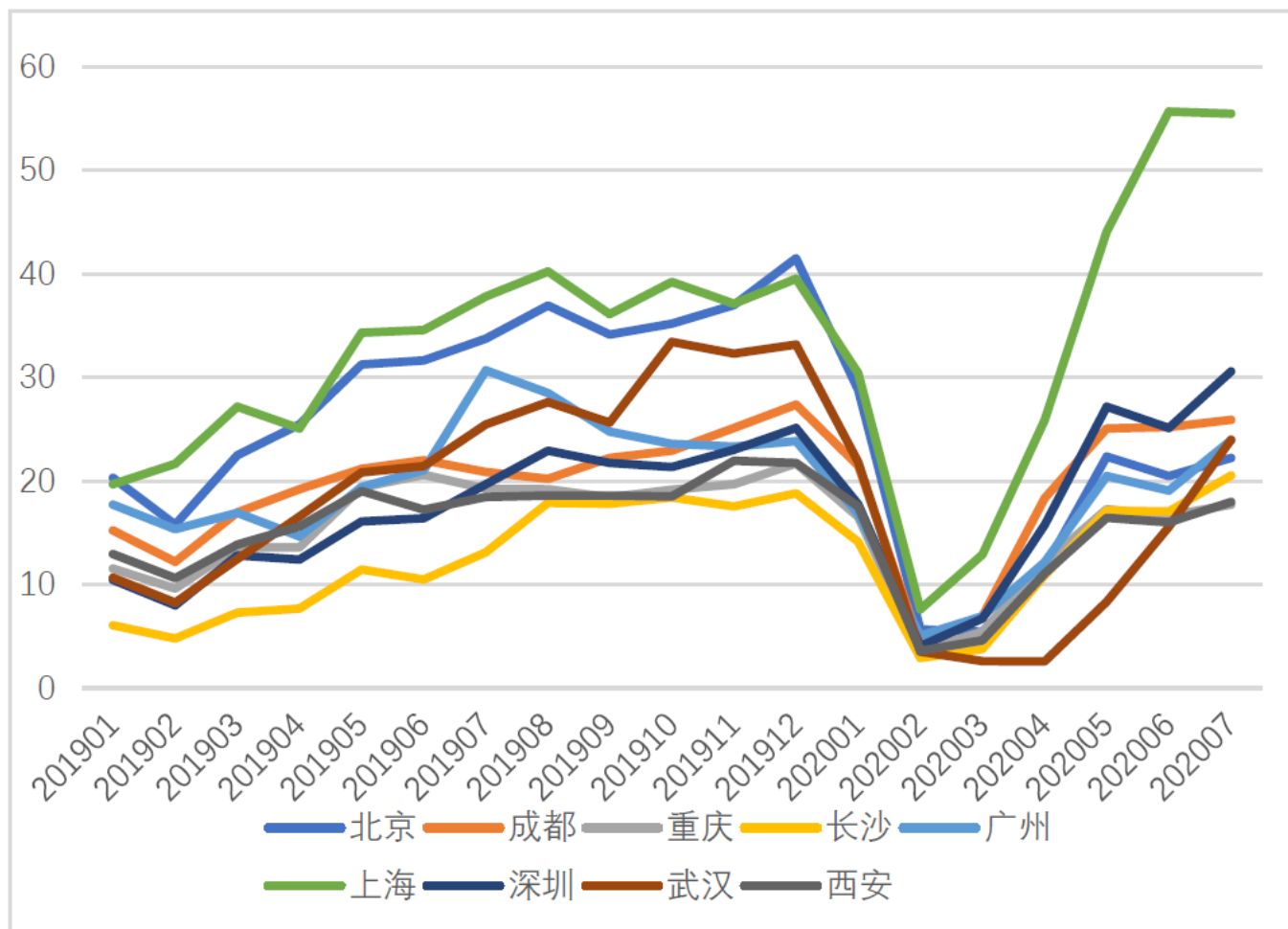
5月词云



6月词云

## 案例2

## 交易大数据与疫情后经济复苏



九城市餐馆月均点评数量

## 案例2

## 交易大数据与疫情后经济复苏



### 各地消费券核销率统计

城市	发放日期	统计时间段	发放总额 (万元)	核销金额 (万元)	核销率 (%)	关联消费 (万元)	关联消费/ 核销金额
(1)	(2)	(3)	(4)	(5)	(5) / (4)	(7)	(7) / (5)
晋城	03.22	03.22-04.19	1500	1442	96.1	17300	.
杭州	03.27	03.27-04.16	37500	32100	85.6	342200	10.7
嘉兴	03.31	03.31-04.16	10600	.	80.0	.	.
佛山	04.01	04.01-04.11	5000	3870	77.4	13600	3.5
内江	04.10	04.10-04.24	.	243	.	821	3.4
沈阳	04.10	04.10-04.24	2000	944	47.2	6459	6.8
衢州	04.03	04.03-04.13	3390	1846	54.5	10147	5.5
郑州	04.03	04.03-04.16	5000	3925	78.5	55200	14.1
赤峰	04.05	04.05-04.20	500	364	72.9	3692	10.1
呼和浩特	04.05	04.05-04.14	1000	86	8.5	719	8.4
绍兴	04.05	04.05-04.12	7500	2333	31.1	10900	4.7
宁波	04.17	04.17-04.23	1070	515	48.1	5321	10.3
乌鲁木齐	04.18	04.18-04.24	4786	330	6.9	5498	16.7
温州	04.18	04.18-04.25	19720	13000	66.2	98200	7.6
武汉**	04.20	04.20-04.27	62	19	30.7	.	.

\* 数据由研究团队收集整理自各地新闻，具体来源参见资料来源中的相关新闻。

\*\*武汉消费券核销数据按照张数（万张）而不是金额（万元）进行统计和计算，数据来源为微信支付团队。

北京大学新结构经济学研究院报告《消费券的中国实践--我国消费券发放的现状、效果和展望研究》



## 案例3

## 专利大数据与创新战略

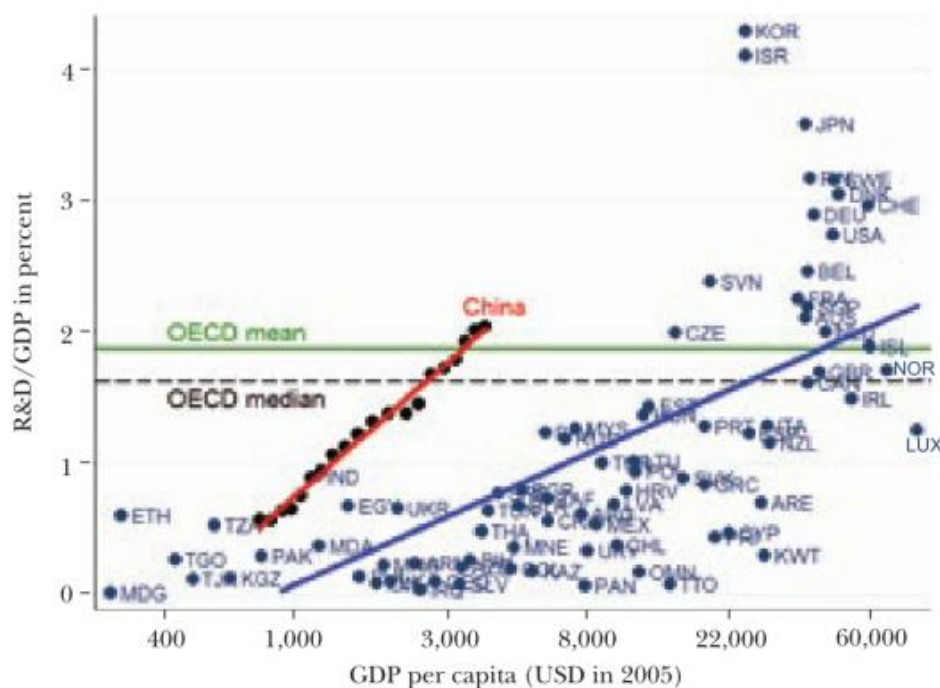
- ✓ 十九届五中全会提出“坚持**创新**在我国现代化建设全局中的**核心地位**，把**科技自立自强**作为国家**发展的战略**支撑”
- ✓ 从“**中国制造**”向“**中国创造**”的伟大转型
- ✓ 如何衡量国家或企业的创新能力？
- ✓ 专利基本信息、技术文本、法律状态、引用状况等
- ✓ 海量，公开，及时更新（每周1-2次）

## 案例3

## 专利大数据与创新战略



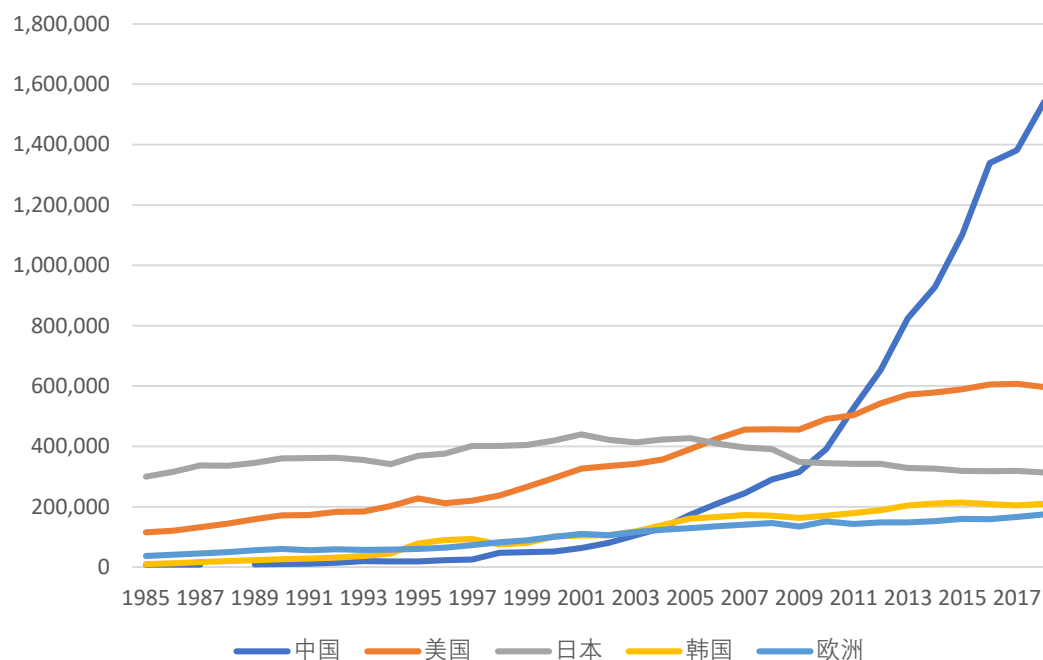
Figure 2  
Research and Development Intensity Comparison



- ✓ 高收入国家的研发支出占国内生产总值的比例往往更高
- ✓ 2010年，中国研发支出占GDP的比重达到了OECD的中位数。
- ✓ 2012年，中国研发支出占GDP的比重达到了OECD的平均值

统计年份：中国：1995-2014；其他国家：最近一年（2014）

全球五大知识产权局专利申请数量，1985-2018

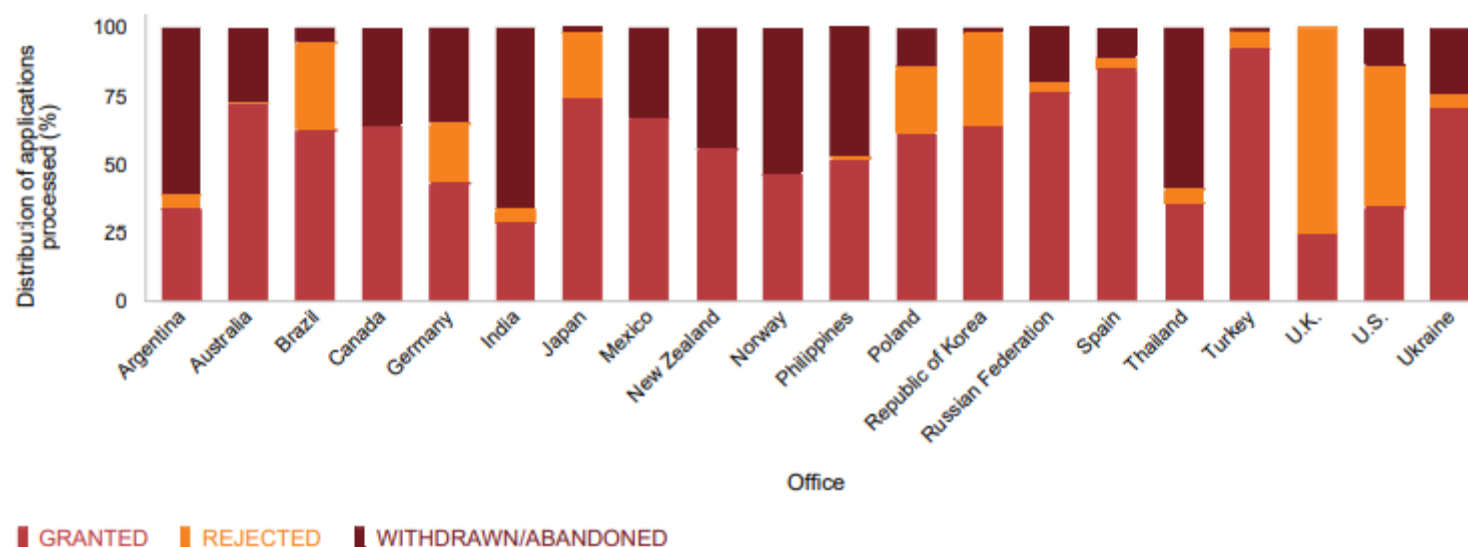


中国专利申请数量占全球专利申请数量的比例从2008年的15%，  
增长至2018年的46.4（154万）

Source: WIPO Statistics Database, August 2019.



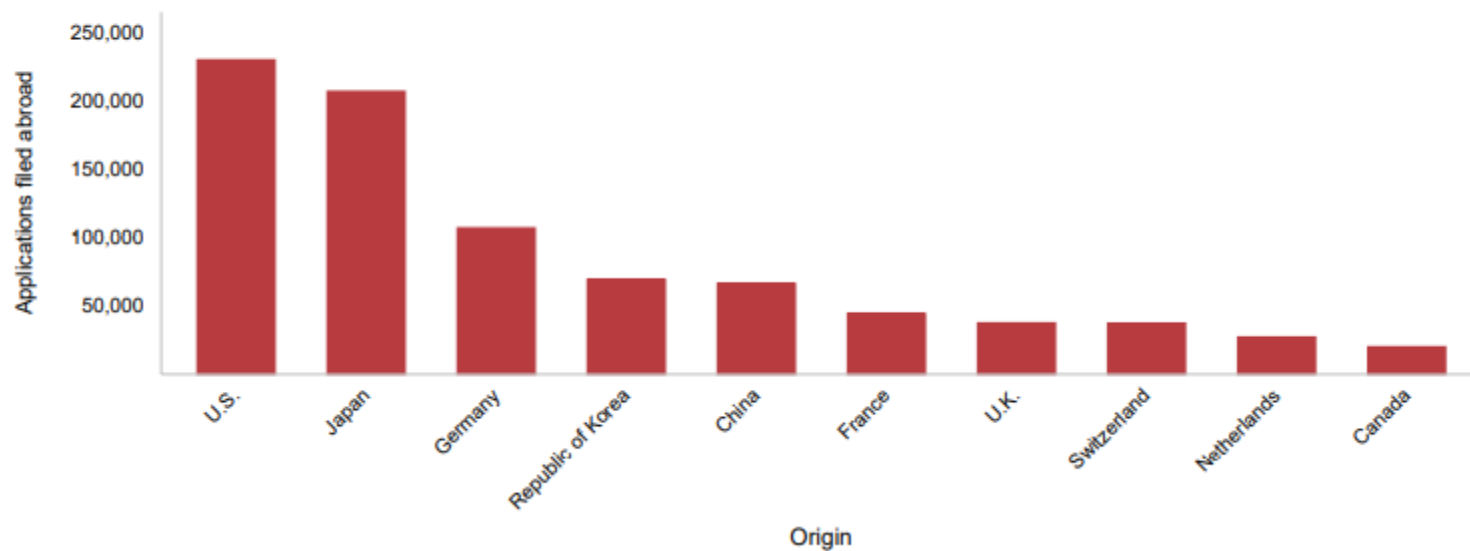
#### A43. Distribution of patent examination outcomes for selected offices, 2018



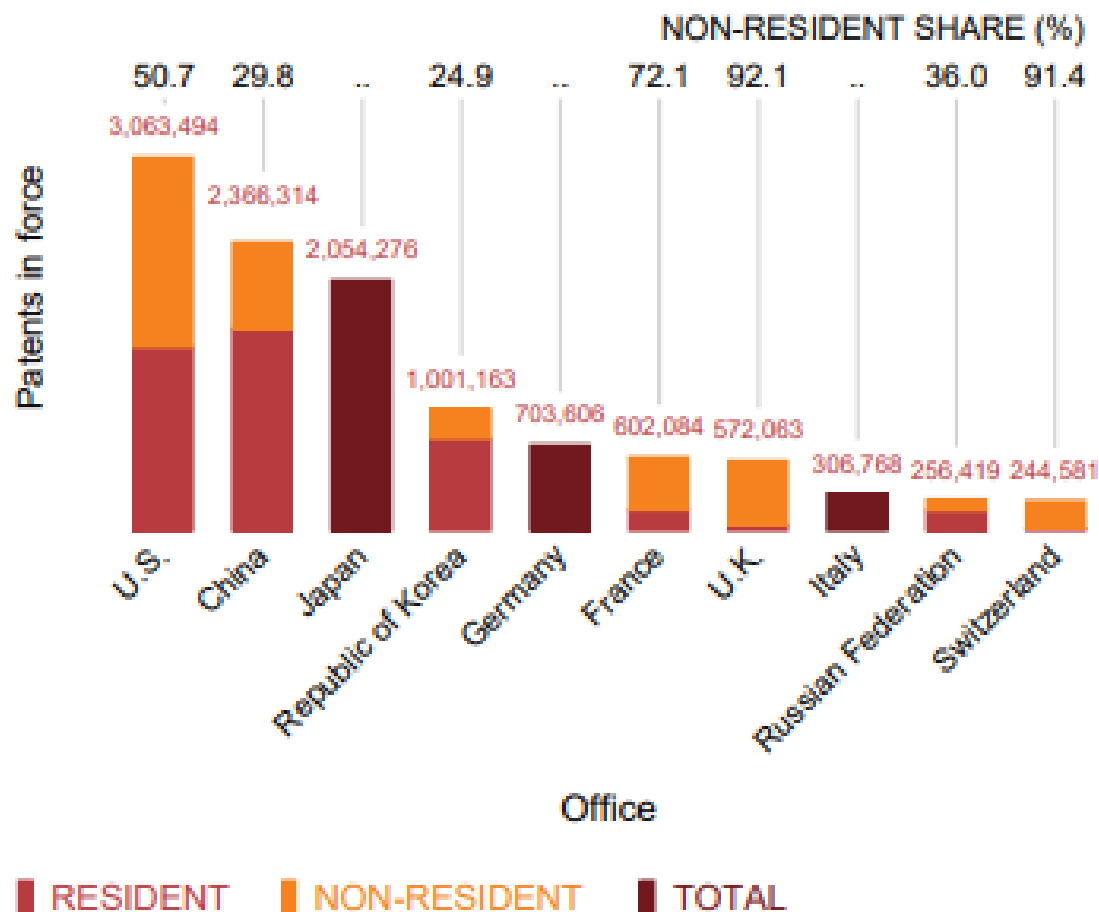


## U.S. applicants filed the largest number of applications abroad

### 1.4. Patent applications filed abroad by the top 10 origins, 2018



## A40. Patents in force at the top 20 offices, 2018



# 案例3

## 专利大数据与创新战略



技术进步是建立在自身的基础上的，一个领域发明的扩展推动了相关领域的未来工作。

使用了180万项美国专利及其引用属性来描绘创新网络及其实力。过去的创新网络对1995年后的未来创新有很强的预测能力。

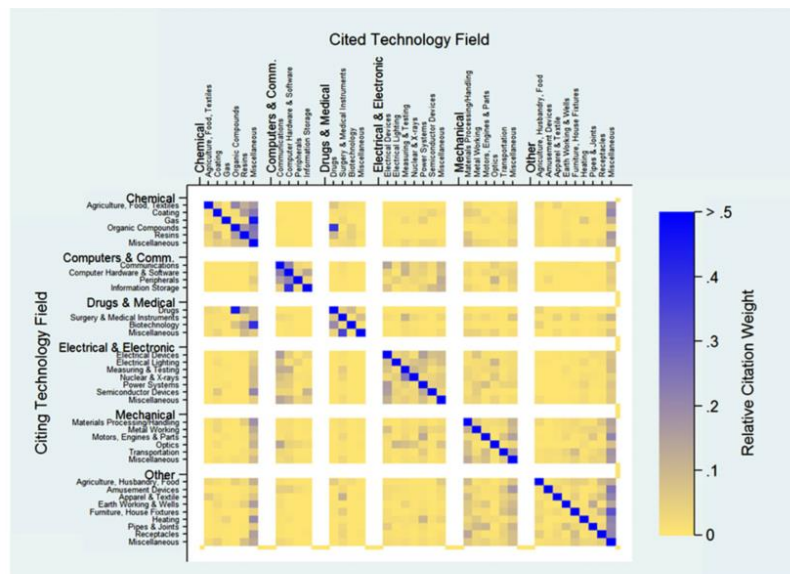


Fig. 1. Citation matrix 1975-1984. Each row describes the field composition of citations made by the technology subcategory indicated on the left-hand side. Entries across cited technology fields for each citing technology subcategory sum to 100%. The diagonals—citations of one's own field, the majority of citations—are excluded from the calculation but given dark shading for reference. SI Appendix, Fig. 1 shows the 1975-2004 network and additional subperiods.

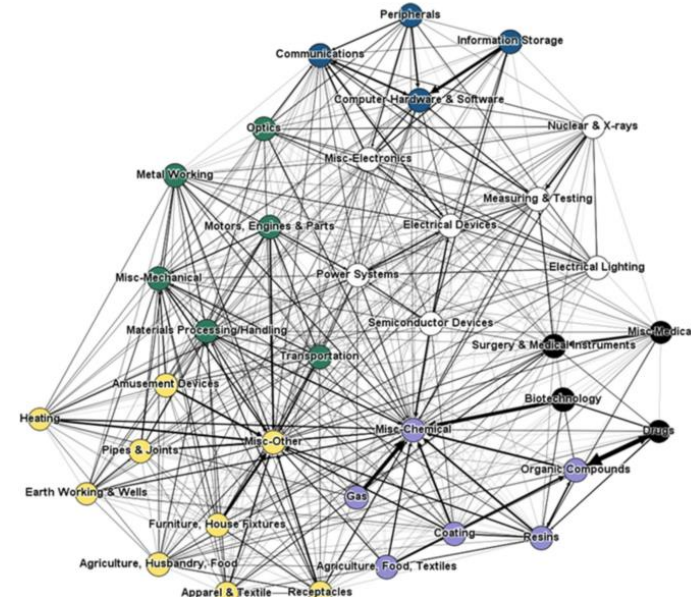
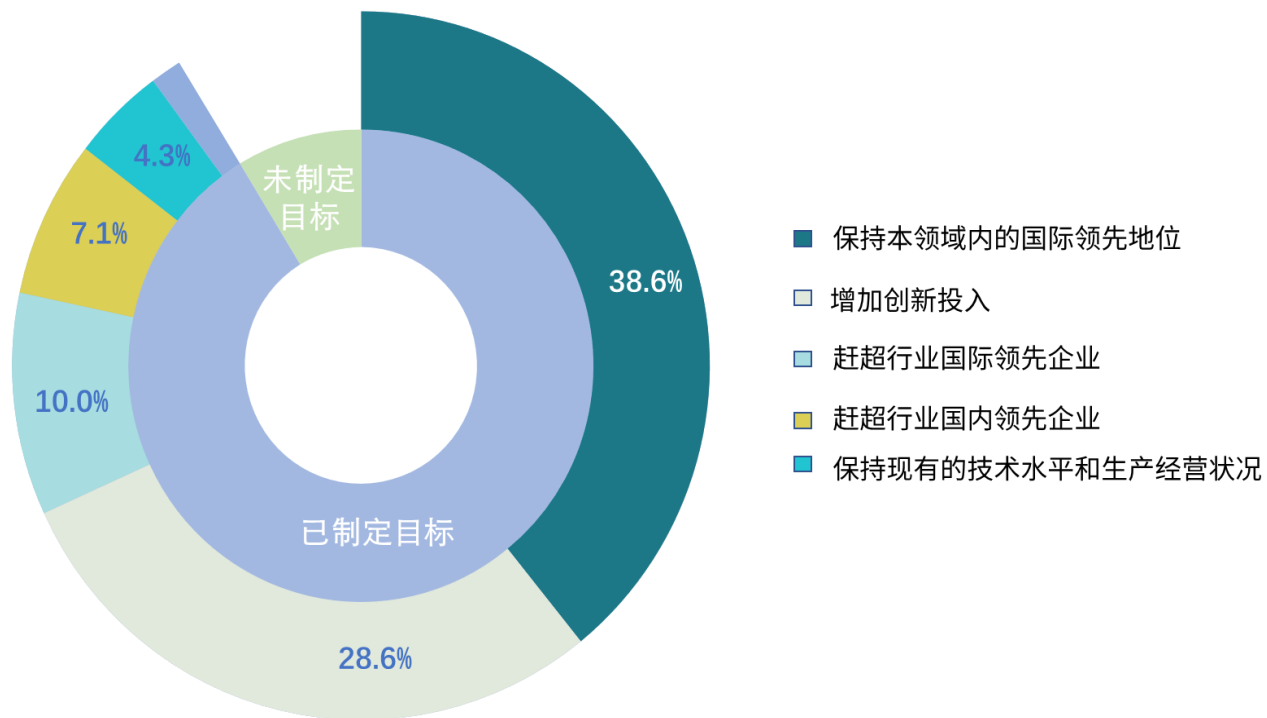


Fig. 2. Innovation network 1975-1984. Network mapping of patent system using technology subcategories. Nodes of similar color are pulled from the same category of the USPTO system. The width of connecting lines indicates the strength of technological flows, with arrows being used in cases of strong asymmetry. Connections must account for at least 0.5% of out-bound citations made by a technological subcategory. SI Appendix, Figs. 2-6 show variations and network properties.

## 案例3

## 专利大数据与创新战略

增强制造业核心竞争力，科技创新推动制造业高质量发展

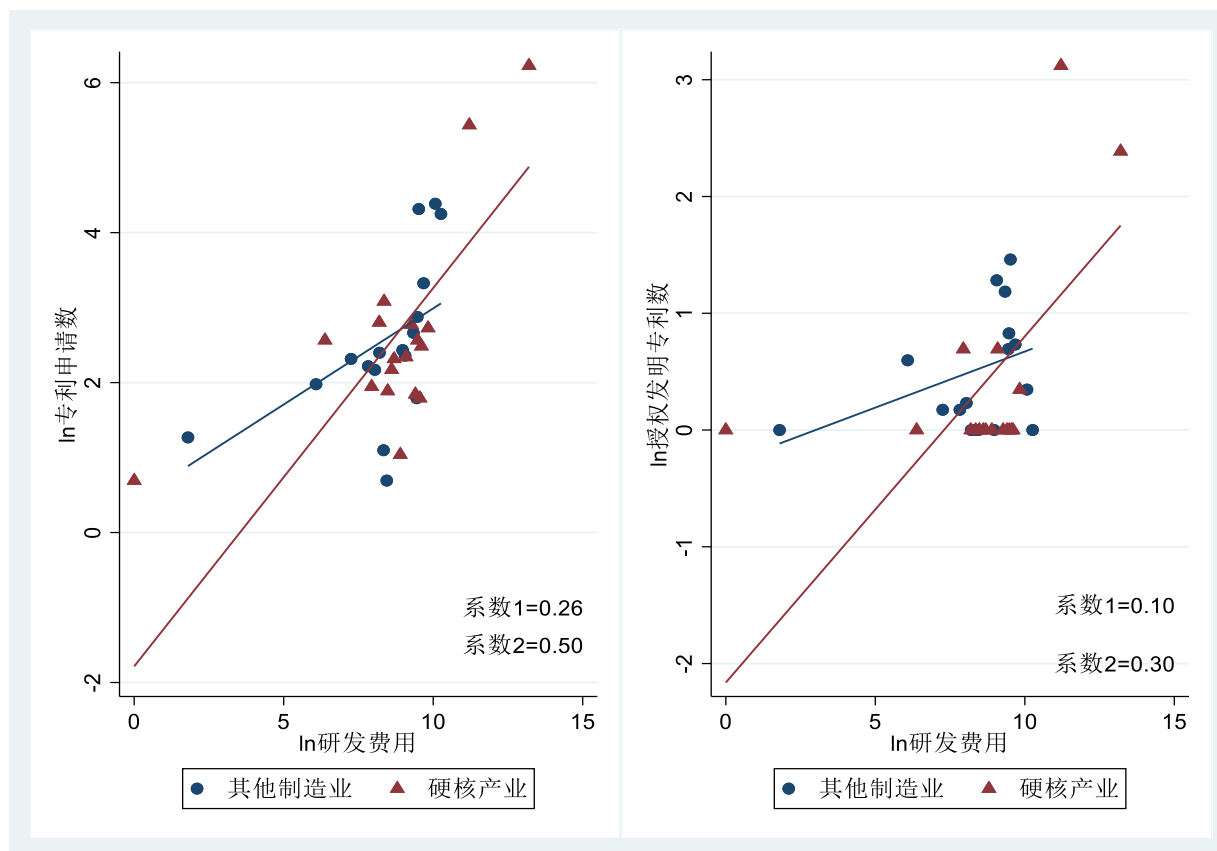




## 案例3

## 专利大数据与创新战略

**创新生产效率：**即创新生产的研发投入弹性，研发费用每增加1%，企业专利申请和授权数增加的百分比

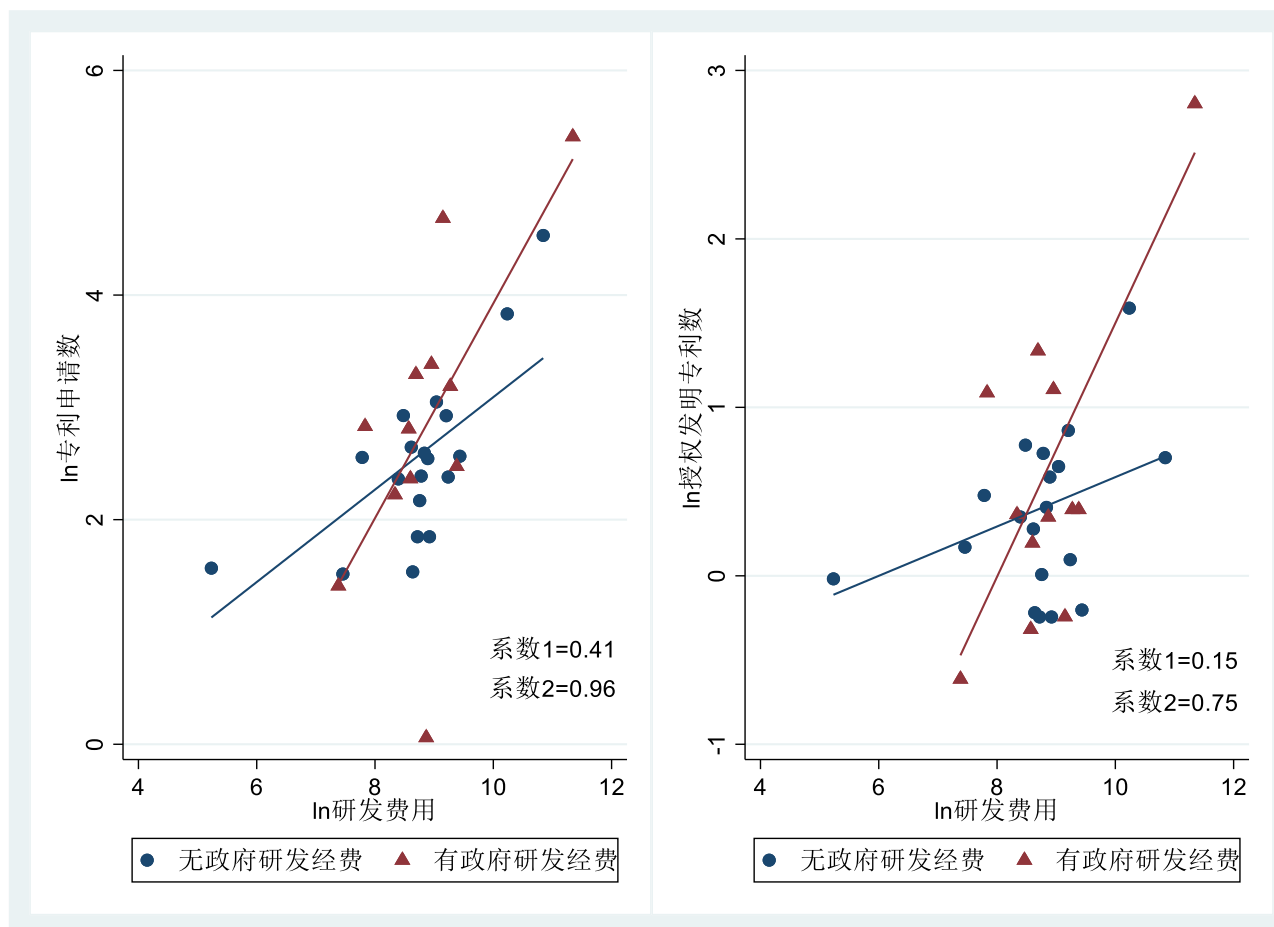


### 案例3

## 专利大数据与创新战略



获得政府研发经费的企业创新生产效率更高



## 案例4

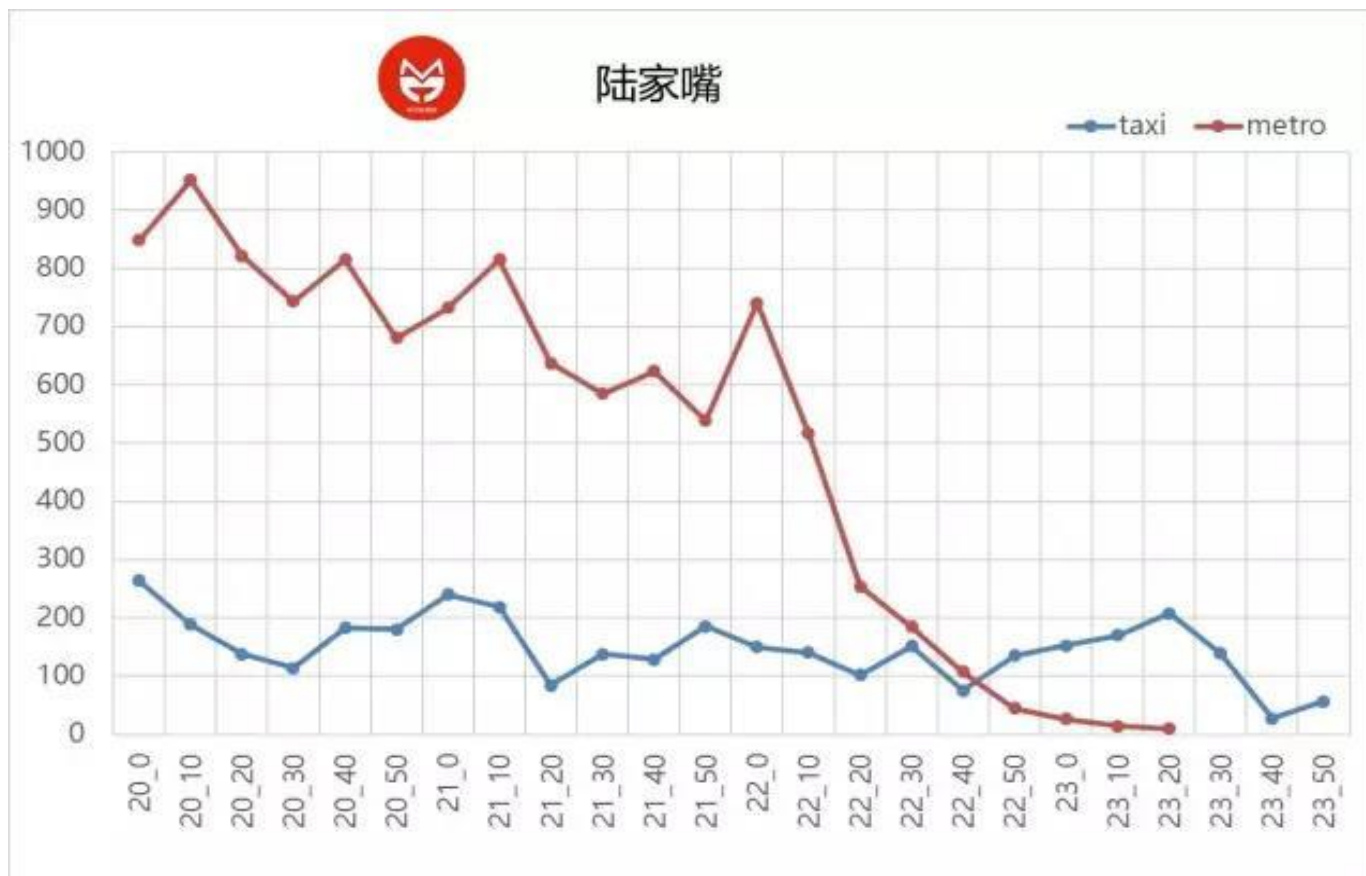
## 交通大数据与上海地铁运营



在晚上20点至22点之间，地铁与出租车的客流波动情况有明显的同步性。但是在22:00-22:10之间，由于地铁的停运，导致出租车的客流需求有个明显的小高峰。

## 案例4

## 交通大数据与上海地铁运营



在停运前的半小时之内（22：50-23：20），陆家嘴地区的出租车需求有了非常明显的上升，而且其上升的顶峰正是停运的23点20分。

## 案例4

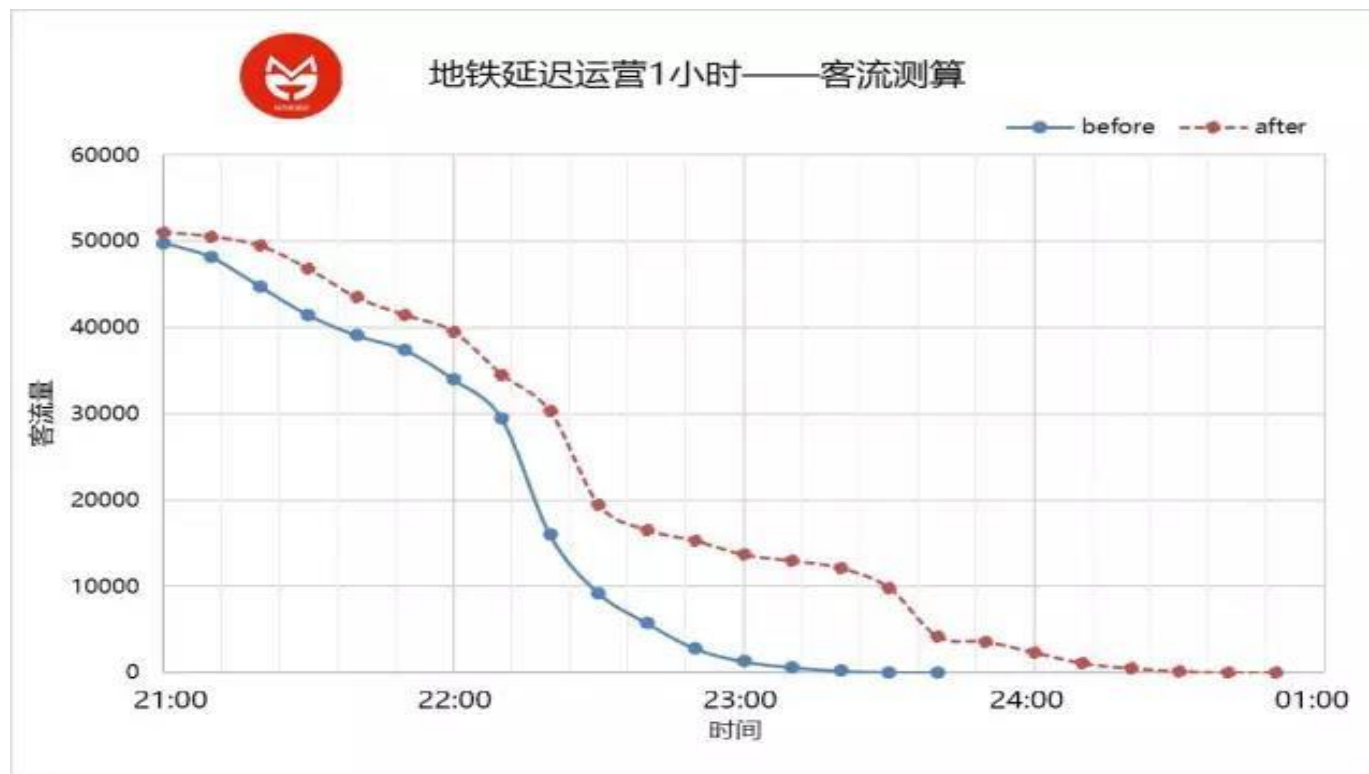
## 交通大数据与上海地铁运营



轨交站点所服务地区的市民，在地铁停运以后，仍然有对公共交通的需求。随着地铁停运，地铁站点附近地区会有大量的公共交通需求转移到了出租车等非公共交通方式上。这种现象存在于商业区、办公区以及交通枢纽区。

## 案例4

## 交通大数据与上海地铁运营



若上海地铁系统延运1小时，上海轨道交通的客流量约增加13万人/次。





## 案例4

## 交通大数据与上海地铁运营

24:00

### 4月28日起上海地铁多条线路延时运营

20:



环球网

17-04-25 14:16

18:00

16

央广网上海4月25日消息（记者韩晓余傅闻捷）为进一步提升上海地铁网络服务水平，特别是进一步满足市民乘客周末、节假日夜间出行需求，上海地铁全网络多条线路自2017年4月28日起延长运营时间：10号线和16号线全线，分别常态延时运营25分钟和30分钟；1、2、7、8、9、10号线六条线路在周五、周六延时运营至零点，“五一”、“十一”及元旦节前最后一个工作日也将延时运营。届时，上海中心城区（外环线内）延时线路的地铁车站周末夜间将运行至零点。

10、16号线常态延时运营

12:00

## 案例5

## 文本大数据与情感分析

### ➤ 经济学中的情感因素

- ✓ 市场情绪（悲观、恐慌）
- ✓ 泡沫、金融传染病、羊群效应
- ✓ 诺贝尔经济学奖得主Robert Shiller(2019) *Narrative Economics*  
《叙事经济学》——重要经济故事传播对经济的影响





## 案例5

# 文本大数据与情感分析

### ➤ 经济学中人文元素的定量分析

- ✓ 用**大数据（文本数据）** 测度社会心理变量（大数据经济统计学）
  - 投资者**情感**指数
  - 消费者**幸福感**指数
  - **经济政策不确定性**指数
  - **经济政策变化**指数
  - **社会舆情**指数
- ✓ 用**文本回归分析（textual regression）** 研究社会或群体心理、情感、文化因素与经济之间的相关关系

## ➤ 人文社会科学的定量分析

- ✓ 用文本数据、概率论与统计学研究《红楼梦》

通过统计《红楼梦》前80回和后40回常用副词出现的频率，用统计学两样本假设检验方法判断是否为统一作者所写。

- ✓ 计算机社会科学 (Computational Social Science, CSS)



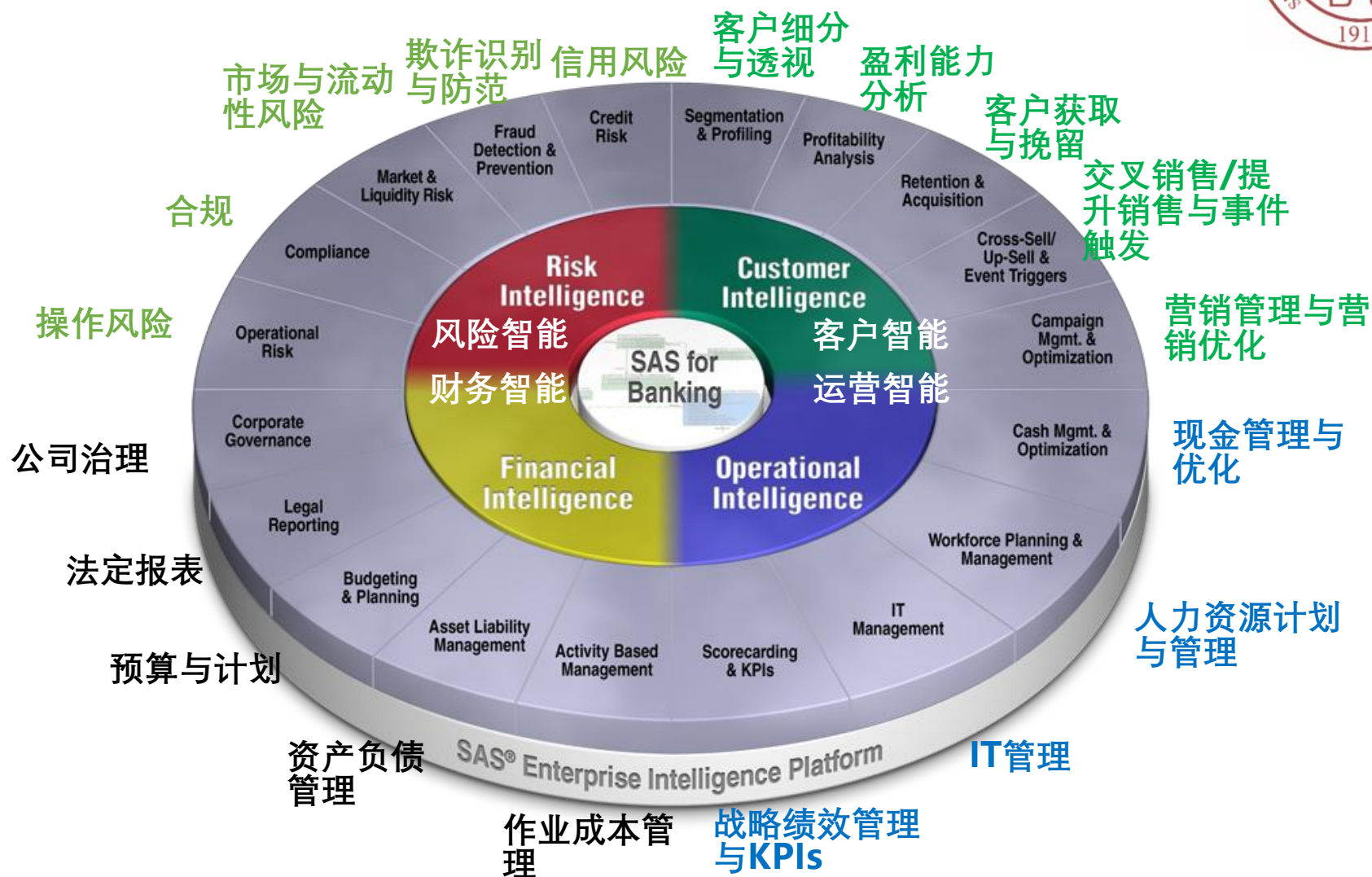
## ➤ 数字经济中的人文元素

- ✓ 外卖与算法
- ✓ 机器学习、信贷风险与人文关怀：普惠金融
- ✓ 大数据、健康码与个人隐私



## 案例6

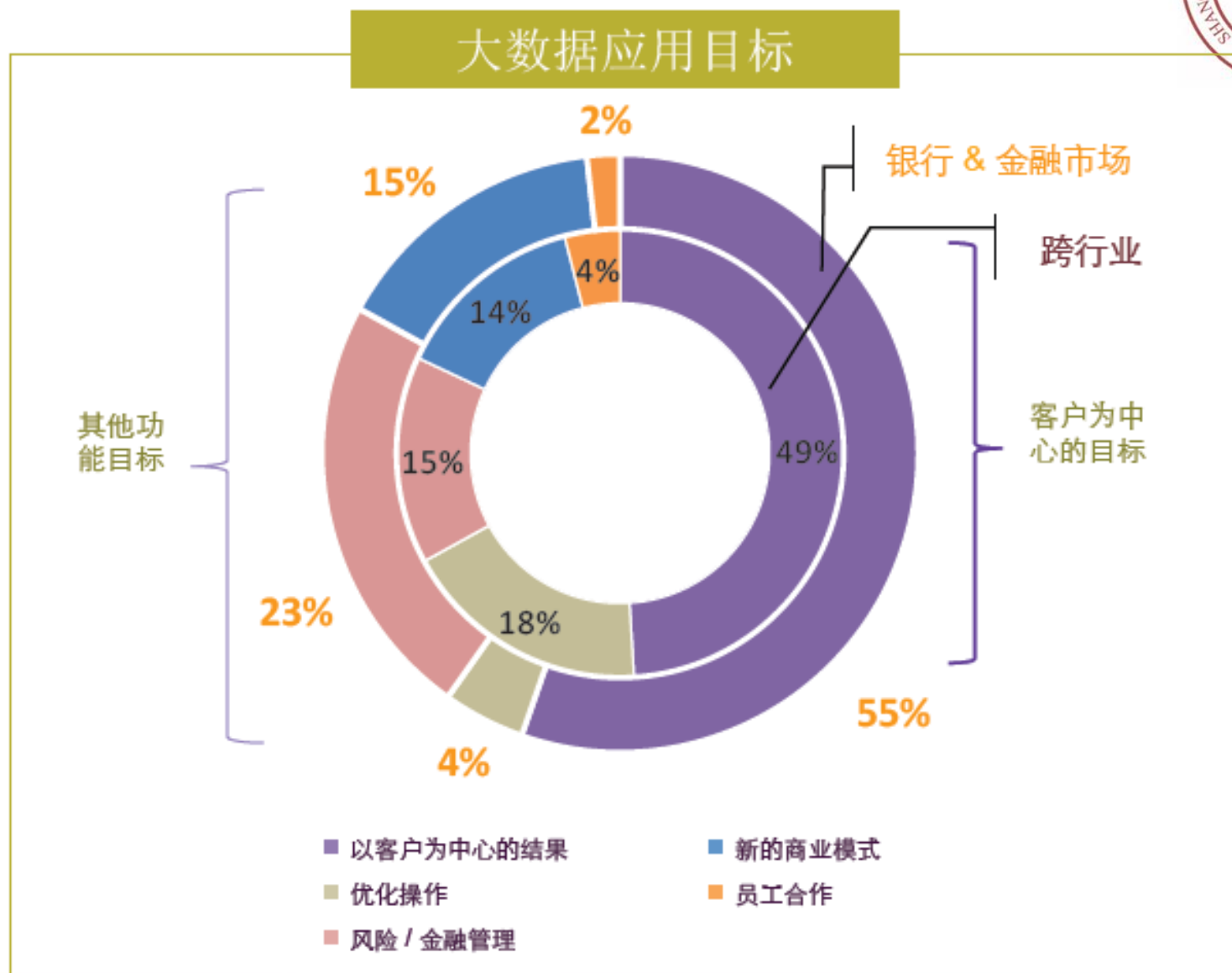
# 大数据在银行业的应用





## 案例6

# 大数据在银行业的应用



## 案例6

# 大数据在银行业的应用

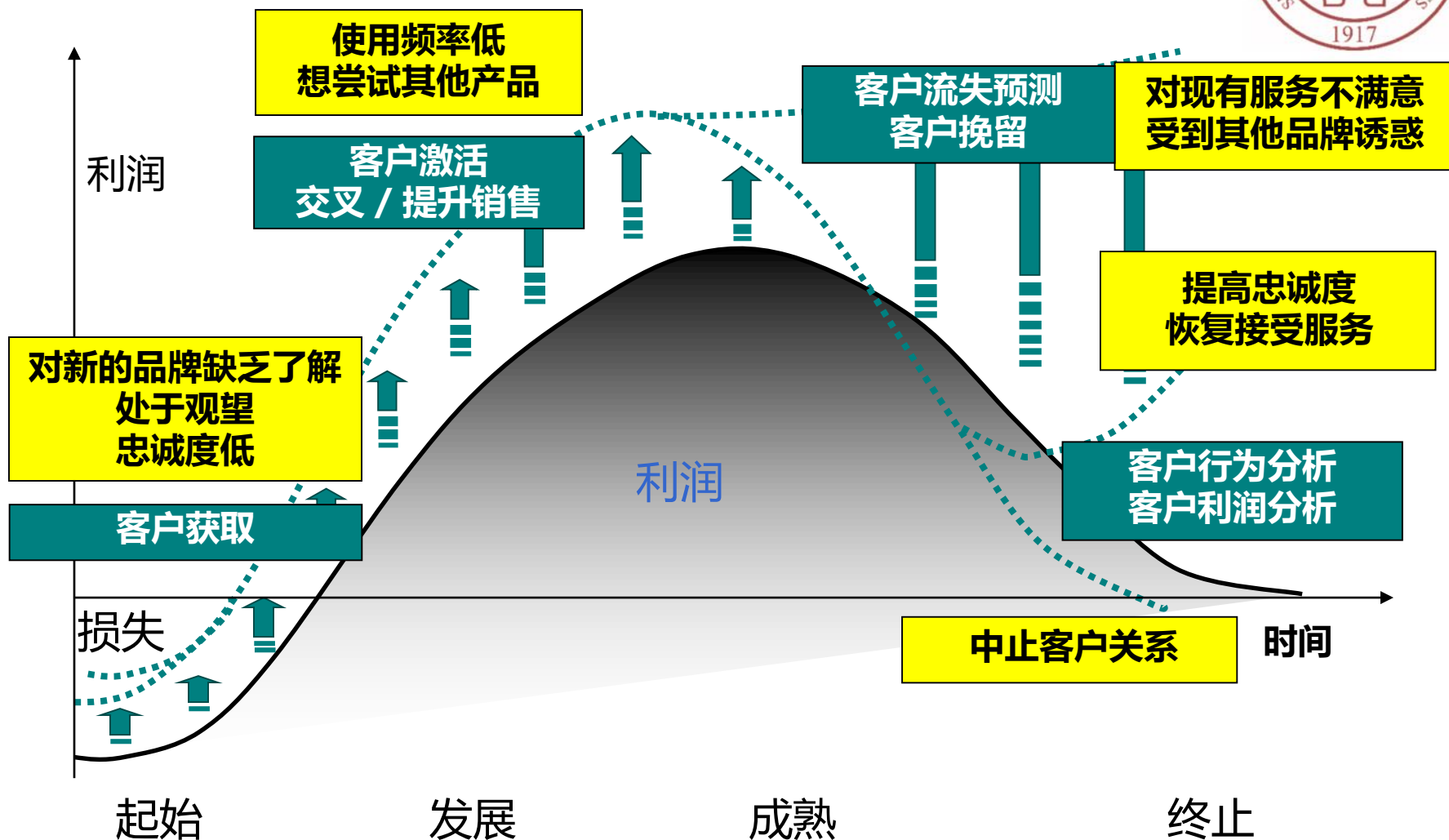




## 案例6

## 大数据在银行业的应用

客户生命周期价值



## 案例6

# 大数据在银行业的应用



## 客户画像

交互设计之父Alan Cooper最早提出客户画像概念：即从真实的客户行为中抽象出客户的特征，目的是为企业活动中与客户相关的决策提供信息基础，指导产品服务研发和市场营销。

客户标签即对客户特征进行符号化表示和刻画；

对客户进行画像的过程也就是客户信息的标签化的过程；



## 案例6

# 大数据在银行业的应用



## 客户画像

### 挑战

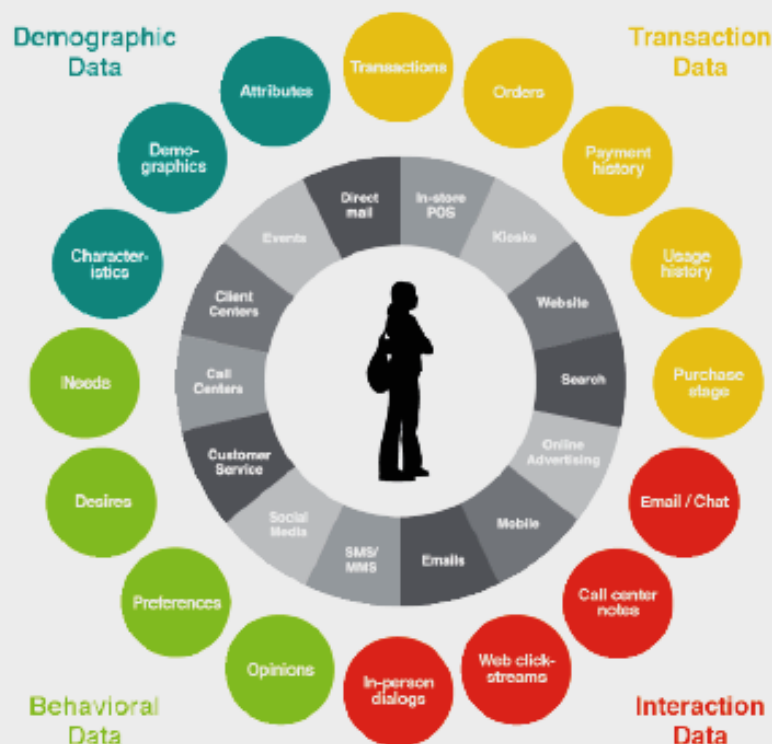
由于越来越多的客户通过数字渠道互动，所以金融机构需要进行许多实时分析并快速响应客户需求

### 数据与分析

客户的人口统计信息、交易数据、点击流和网络日志，使用路径分析确定客户模式或倾向

### 成果

提高了客户洞察力... 经验，交叉销售和盈利

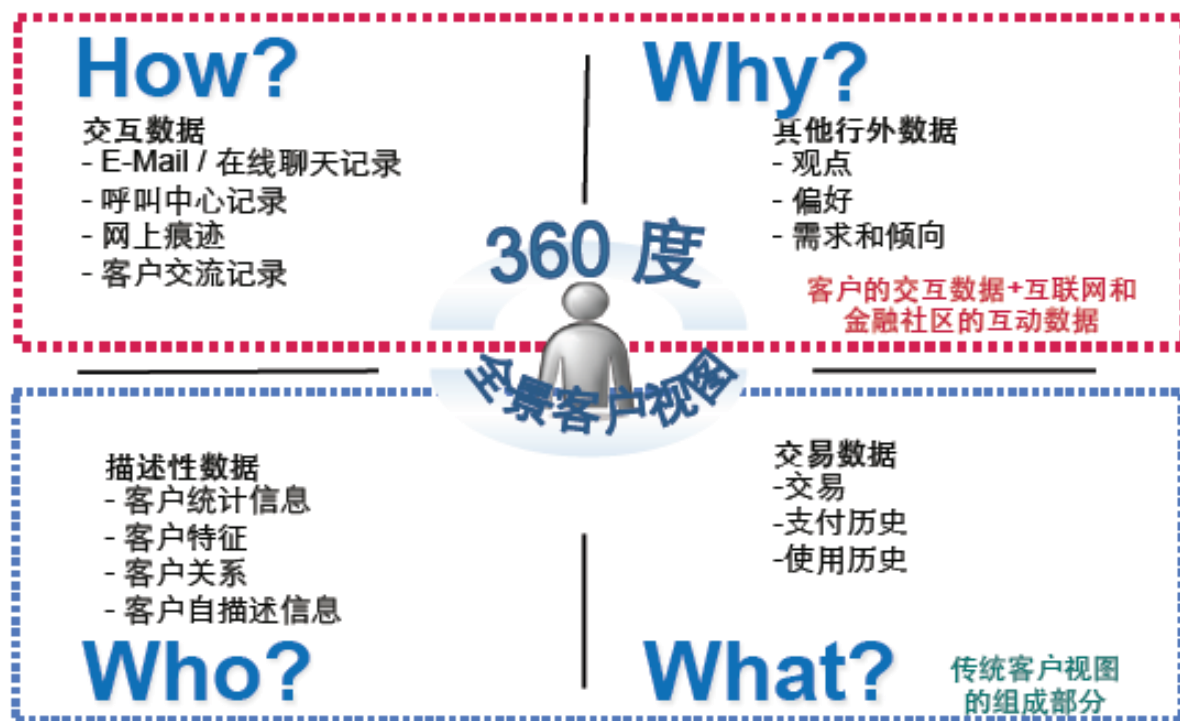


## 案例6

# 大数据在银行业的应用

## 数据是客户画像的核心

基于行内优质的结构化数据和非结构化数据、结合行外跨界合作和交换获得的数据，丰富并完善客户的基本信息、金融信息、行为信息、交互信息等，清晰地描述客户的特征和轮廓。



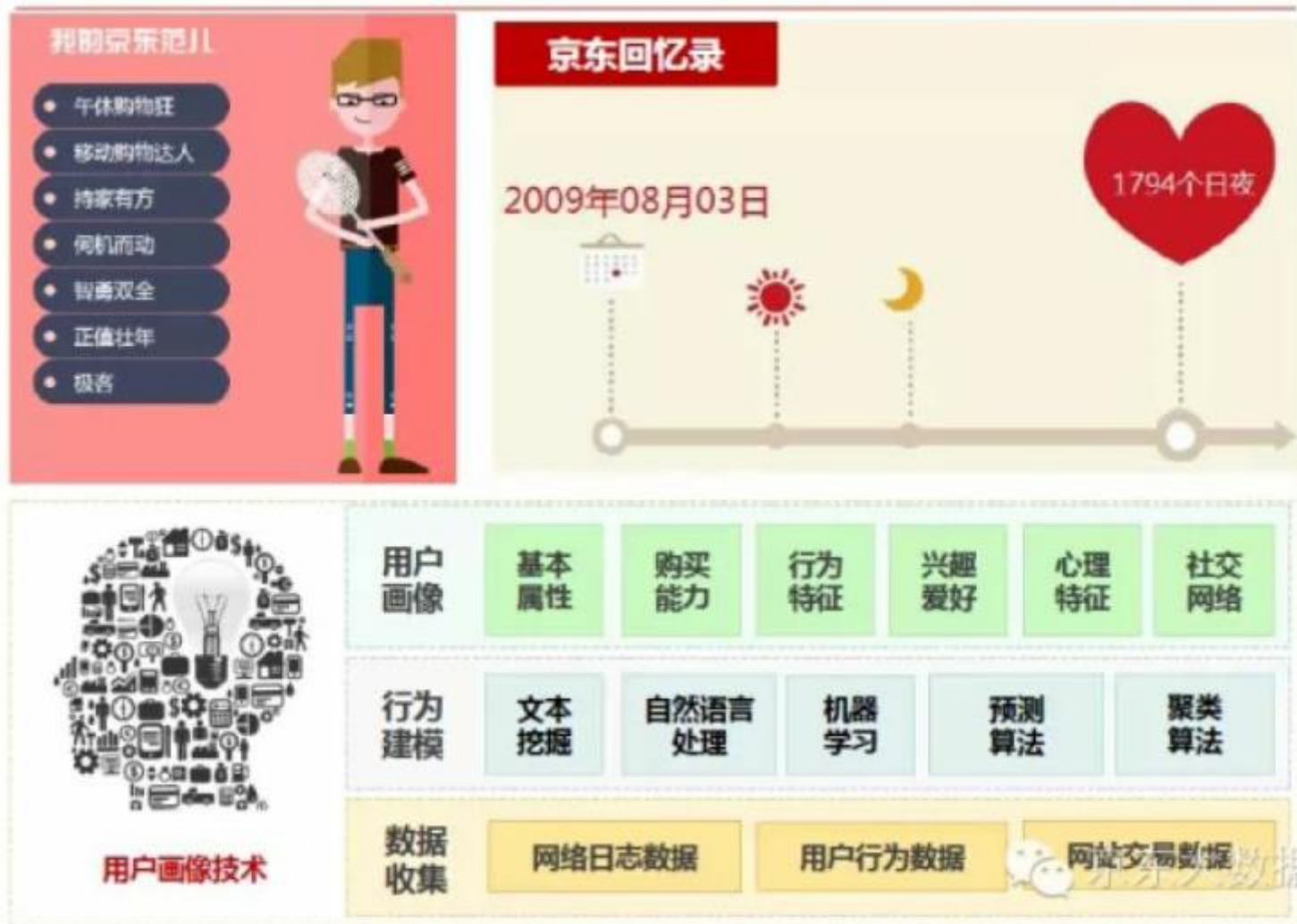


# 京东客户画像案例



京东大数据

# 京东客户画像应用案例

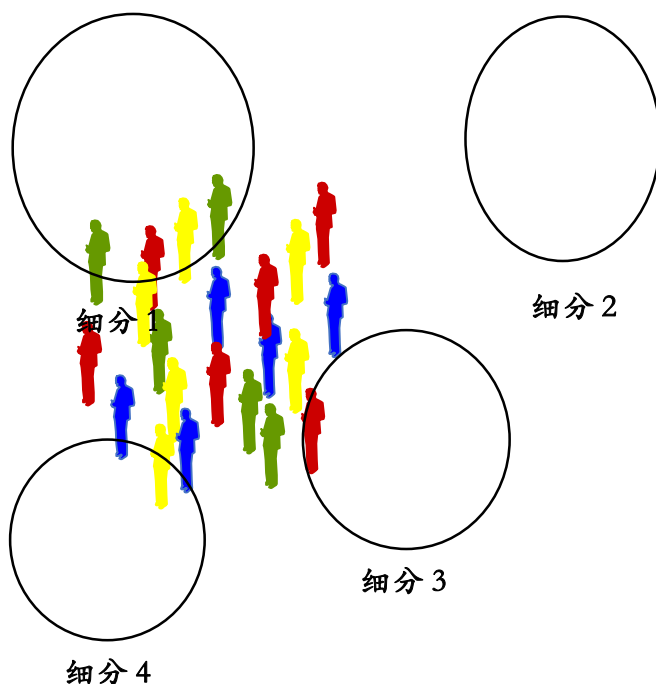


## 案例6

# 大数据在银行业的应用



## 客户细分

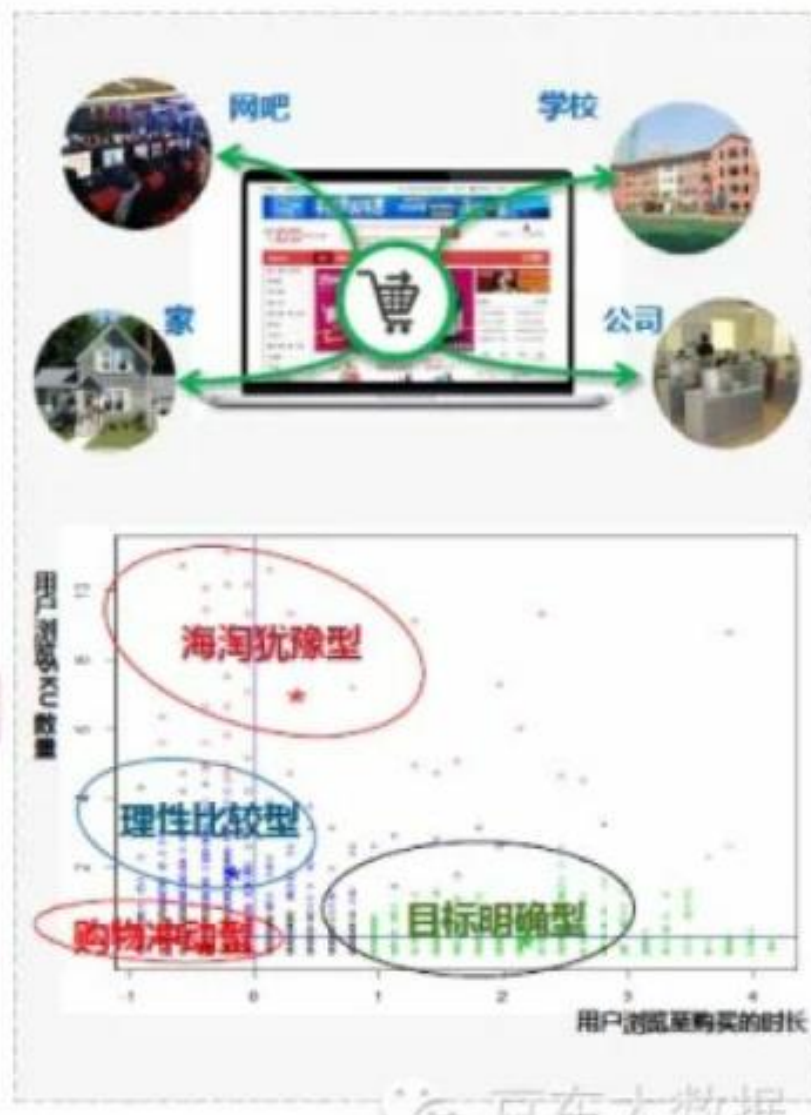


客户细分是将全体客户划分为多个分组并刻画特征的过程，使得组内客户高度相似，组间客户差异明显。

### 目的

- 有的放矢—精细化营销活动，生成可控的目标客户群
- 量体裁衣—发现各个细分的客户特性和需求，有针对性地设计营销计划
- 高瞻远瞩—发现战略焦点和业务发展方向

# 京东客户分群案例





## 案例6

# 大数据在银行业的应用



## 交叉销售

### 挑战

在对客户了解不够时如何向他们推荐产品（如数字银行的新用户）

### 数据与分析

通过客户、产品、渠道、服务、社交数据分析、购物篮分析/Cfilter 进行‘people like me’推荐，NBO 算法

### 成果

改善了客户体验，增加了销售额，提高了客户保持率



## 案例6

# 大数据在银行业的应用



操作风险、市场风  
险与信用风险





## 案例6

# 大数据在银行业的应用



## 资金流分析

### 挑战

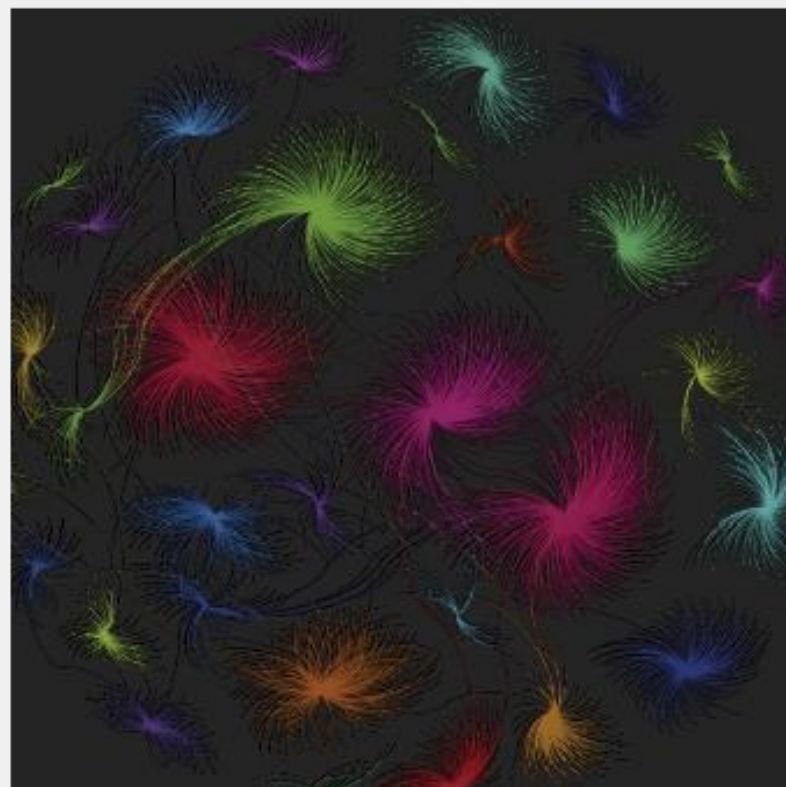
理解企业和银行间价值链上的资金流向，管理风险，核实贷款目的和商业活动

### 数据与分析

客户层面的贷款组合数据，单一客户视图。Teradata Aster & Aster Lens

### 成果

卓越客户洞察，改进风险信贷管理，改进客户获取和交叉销售





# 谢谢!

Thank You

