



中國人民大學
RENMIN UNIVERSITY OF CHINA

计算传播理论与实务

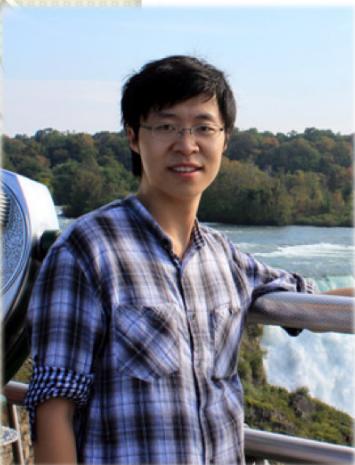
2019-2020 秋季学期

授课教师：范举副教授、塔娜讲师

上课地点：公共教学二楼2111

上课时间：每周一 10:00 – 11:30

教学小组



- 任课教师：范举
 - 信息学院副教授
 - 清华大学计算机系博士
 - 新加坡国立大学博士后
 - fanj@ruc.edu.cn



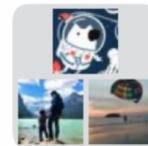
- 任课教师：塔娜
 - 新闻学院讲师
 - 清华大学计算机系博士
 - tanayun@ruc.edu.cn

研究方向
大数据技术
2007年至今

研究方向
计算传播
2013年至今

教学小组

- 助教：张大方
 - 信息学院2017级本科生
 - 电邮：
 - 2017202034@ruc.edu.cn
- 答疑安排
 - 地点：信息楼500
 - 时间：群内沟通
- 课程主页
 - 课件下载、通知发布
 - <https://github.com/fanju1984/cc/wiki>

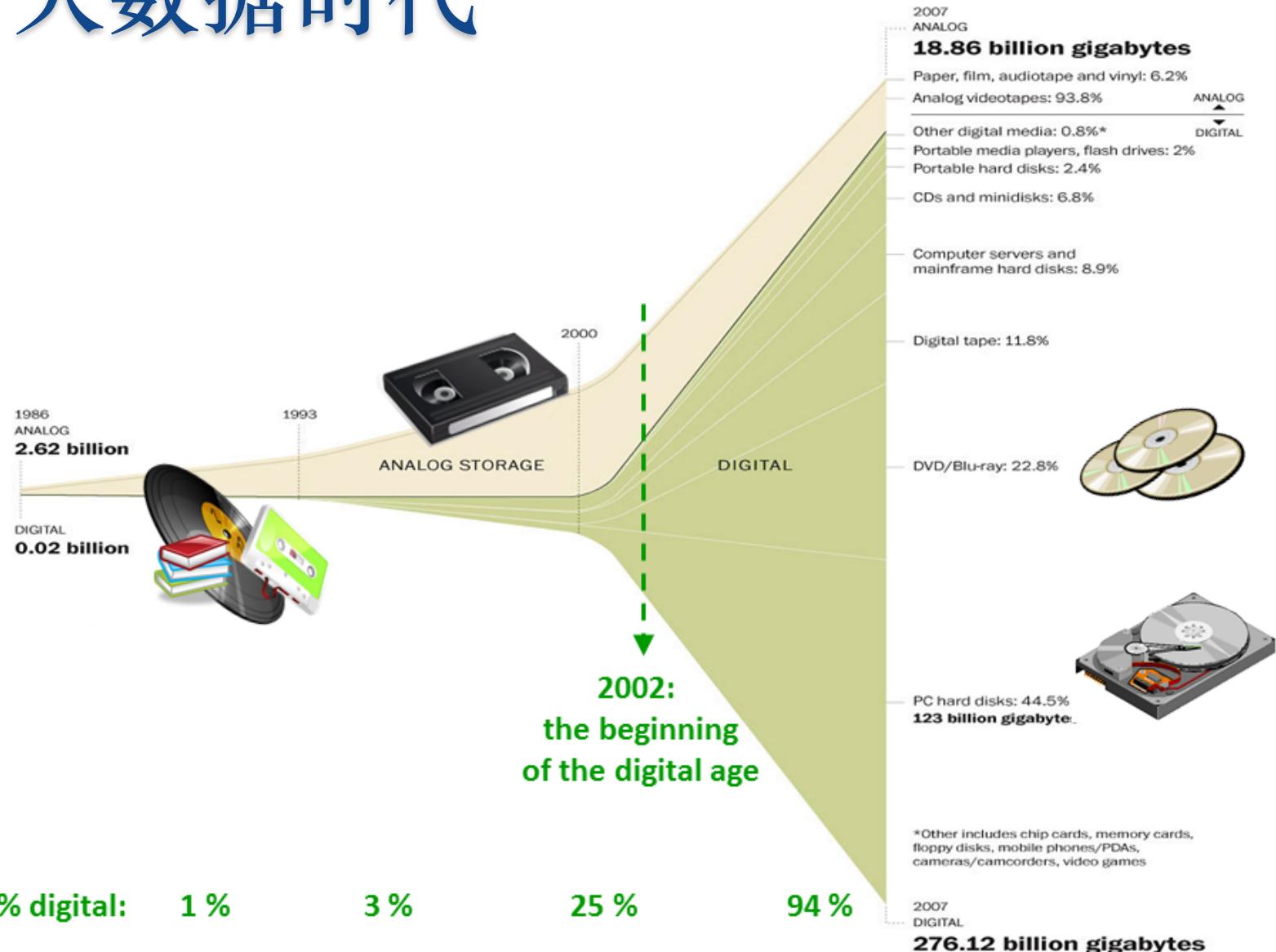


2019计算传播课程群



Valid until 9/22 and will update upon joining group

大数据时代



小计算

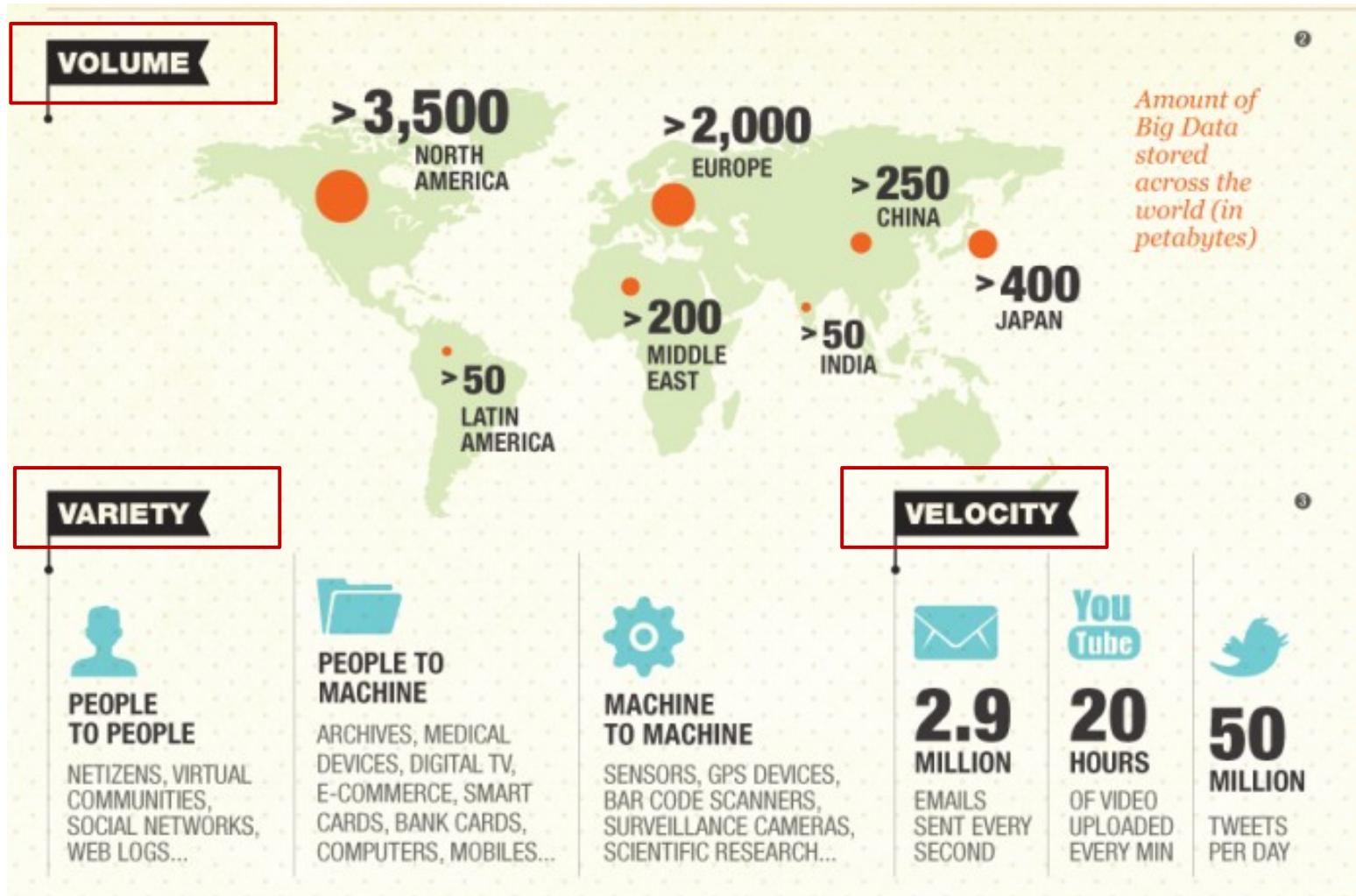
- 数字信息量每2.5年翻一倍
 - 2007年: $276.12 \text{ billion GB} = 2.76 * 10^{14} \text{ MB}$
 - 2019年: $8.8 * 10^{15} \text{ MB}$
- 假设一首MP3歌曲时长4分钟，占用空间的大小是10MB。如果2019年的数字信息量全部用来存储歌曲，够一个人不眠不休听多久？

$$\frac{8.8 \times 10^{15} \times 4}{10 \times 60 \times 24 \times 365} = 6.7 \times 10^9 \text{ 年} \quad \text{67亿年!}$$

地球年龄：46亿年



大数据的“3V”



Source: <https://visual.ly/community/infographic/technology/big-data>

计算传播理论与实务

人们的行为发生了深刻变化.....



教宗本笃十六世



教宗方济各

Source: <http://www.businessinsider.com/vatican-square-2005-and-2013-2013-3>

数字痕迹 (Digital Traces)

- 纽约证交所每个交易日生成1 TB的交易数据
- Facebook每天大概接收用户 500+T的社交数据，主要是照片、视频、文本消息、评论等
- 2019年春晚全球观众参与百度 APP互动次数达到208亿次

个体与群体社会行为
更全面的展示



计算社会科学

- A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.
- 大数据 + 社会科学



SOCIAL SCIENCE

Computational Social Science

David Lazer,¹ Alex Pentland,² Lada Adamic,³ Simon Aron,^{3,4} Albert-László Barabási,⁵ Devon Bruck,² Nicholas Christakis,¹ Noshir Contractor,⁷ James Fowler,⁸ Myron Gutmann,² Tony Jebara,⁹ Gary King,¹⁰ Michael Macy,¹¹ Deb Roy,⁷ Marshall Van Alstyne^{1,12}

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring. In Internet companies such as Google and Yahoo, and in govern-

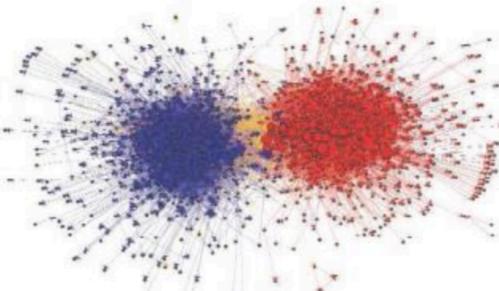
A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

ment agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be

critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understand-

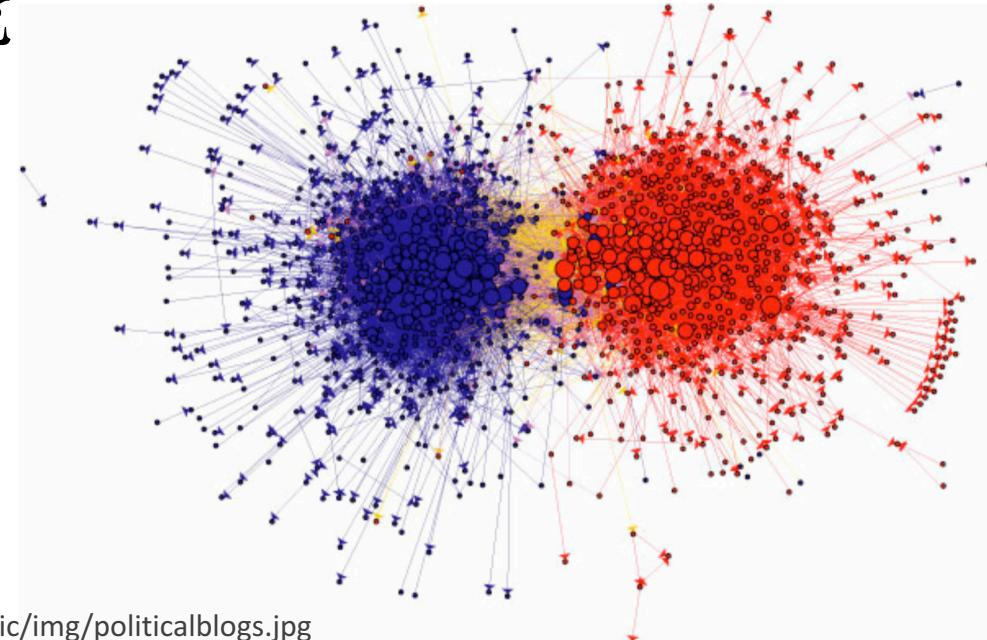
ing of individuals and collectives? What are the



计算社会科学

- 数据驱动的社会科学 Data-Driven Social Science
 - 从海量数据挖掘中理解社会现象的有价值知识
 - 综合应用社会科学和计算技术解释社会现象
 - 终结定量、定性方法的分野
 - 应用导向的研究

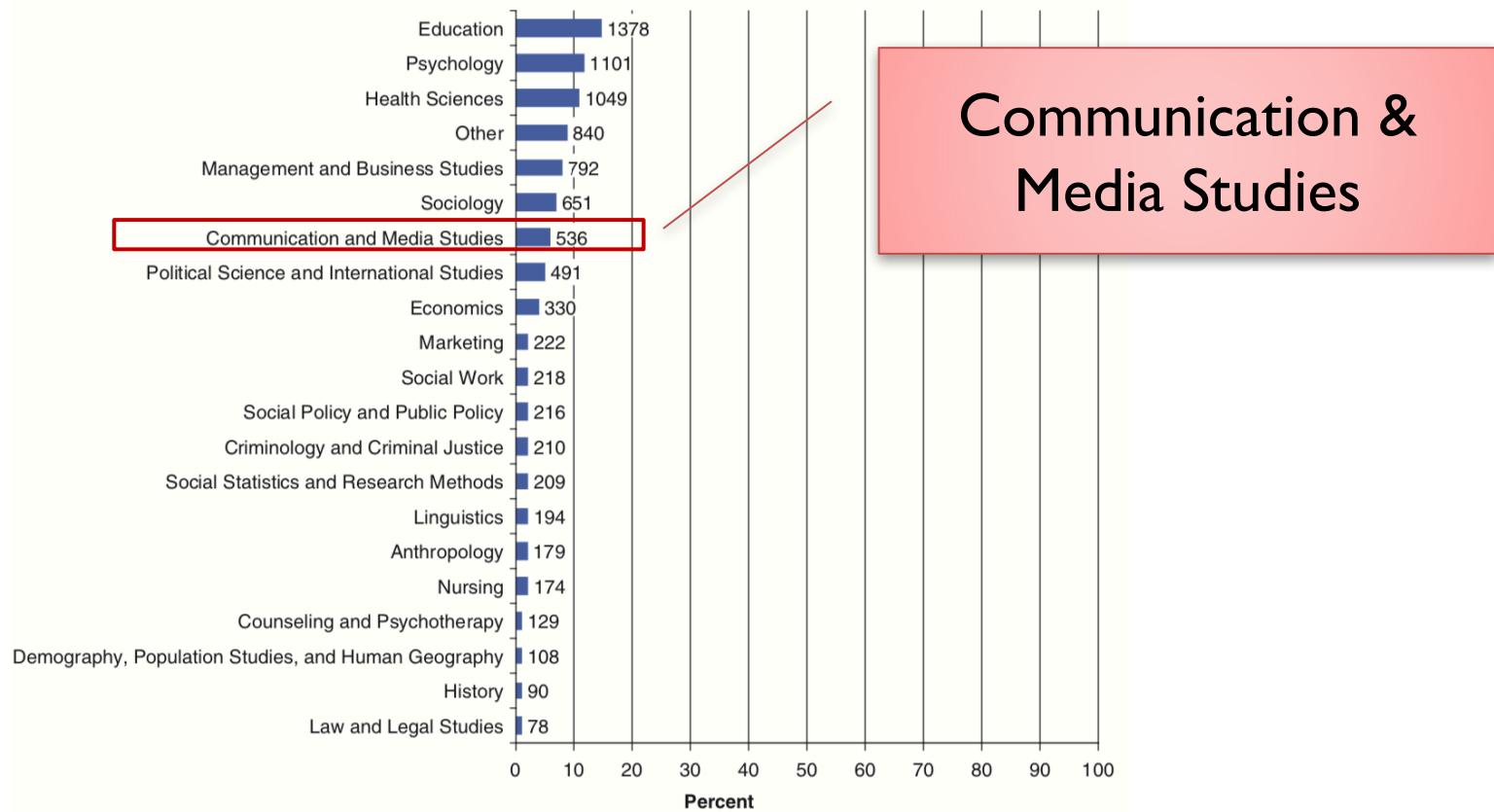
2004年美国大选前
，政治博客之间的
链接网络。**你能发
现什么特点？**



Source: <http://www-personal.umich.edu/ladamic/img/politicalblogs.jpg>

谁在做计算社会科学？

- 2016年SAGE发布白皮书Who is doing computational social science? Trends in big data research



内容提要

- I.1 计算传播学：大数据带来了什么？
- I.2 教学计划
- I.3 考核要求

计算传播学





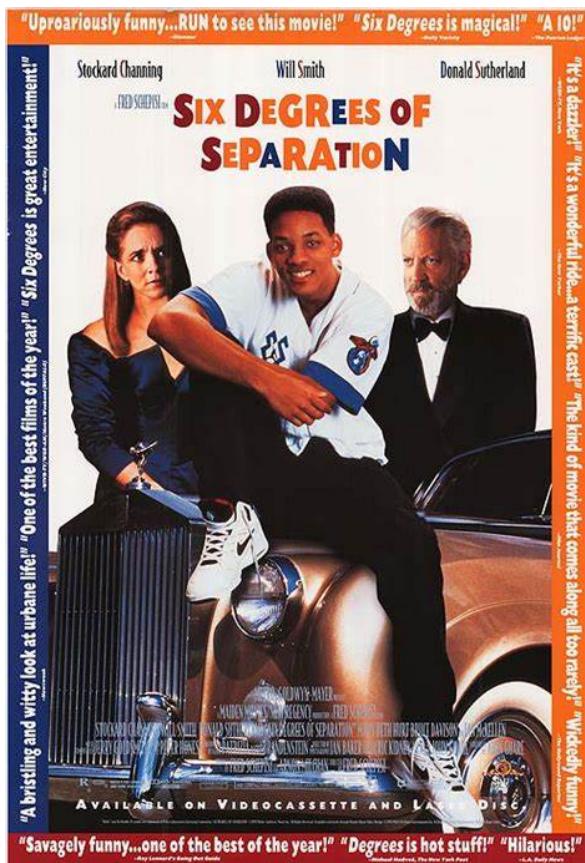
大数据带来了.....

新数据

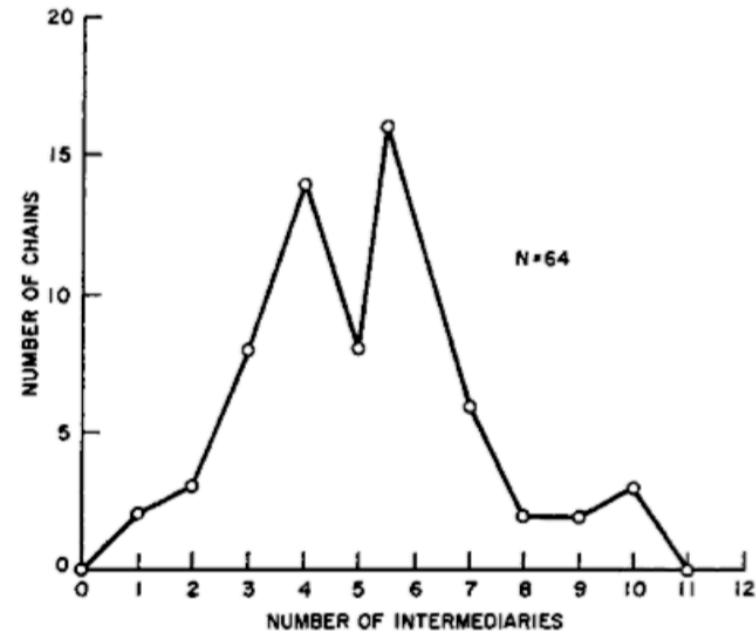
- 大规模地收集与分析人类传播行为
 - 分分秒秒的人类交互行为
 - 无处不在的地理位置信息

传统传播学收集数据的局限性

- 规模小：大多数研究基于小规模数据样本
 - 举例：六度分隔理论（Six Degrees of Separation）



“在这个世界上，任意两个人之间，只隔着六个人”



计算传播学的数据优势

- 收集大规模的个体/群体行为数据

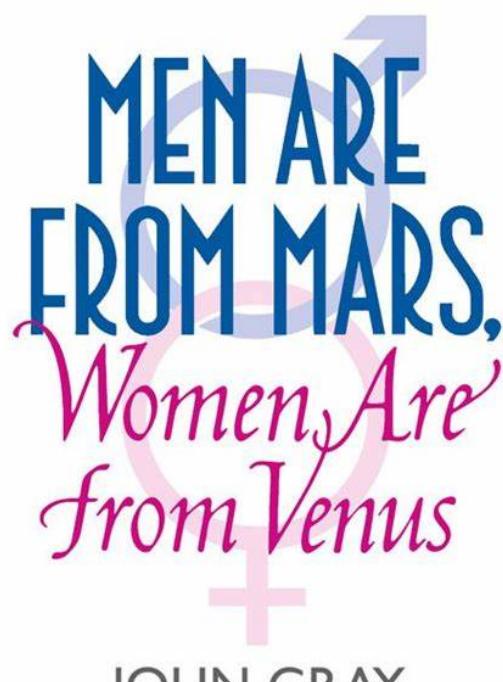


4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

传统传播学收集数据的局限性

- 有偏差：大多数研究基于自我报告式数据
 - 举例：性别研究（Gender Study）

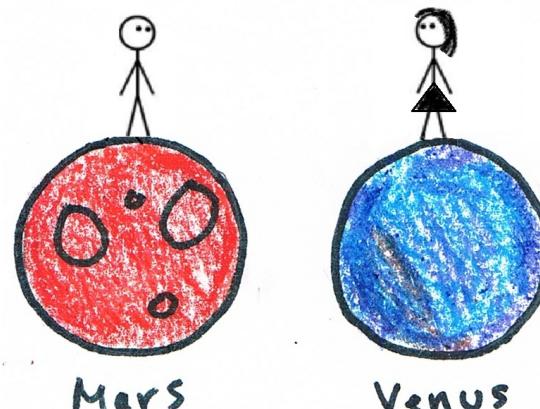
THE DEFINITIVE GUIDE TO RELATIONSHIPS



OVER 15 MILLION COPIES SOLD WORLDWIDE

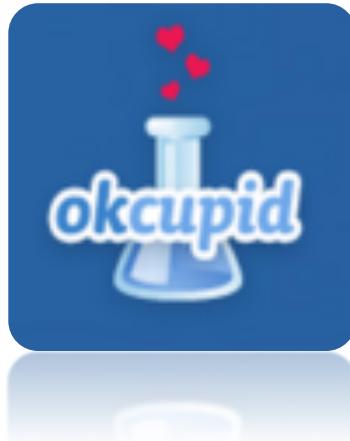
不同性别之间对于配偶的期待存在着显著差异

我的年龄 vs. 配偶年龄



计算传播学的数据优势

- 收集个体自发产生的行为数据
 - 比如：在线交友网站的数据.....

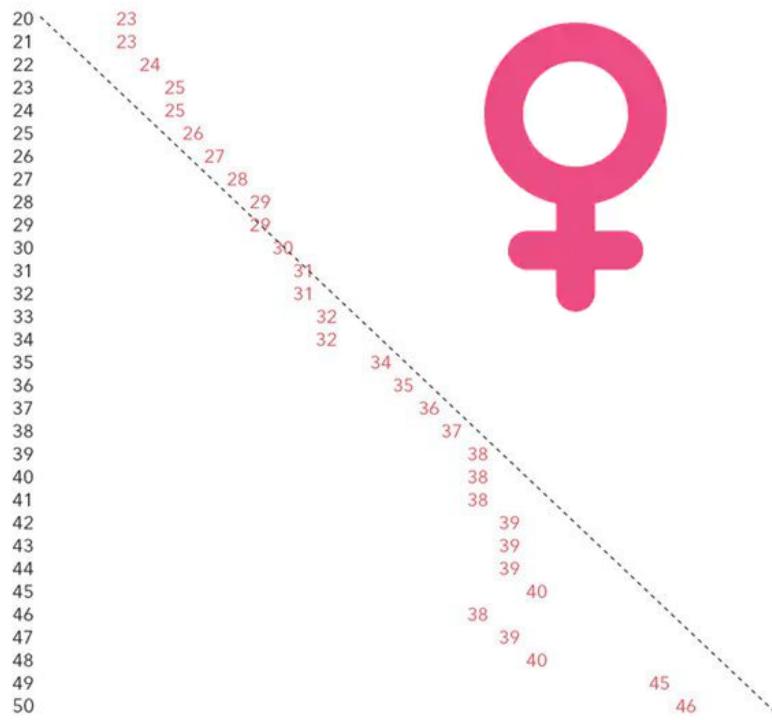


A screenshot of the OkCupid website. At the top, there's a navigation bar with links for "Browse Matches", "Messages", "Visitors", "Quickmatch", and "Events". Below the navigation is a search bar and a "Promote me" button. The main content area shows a user's profile. On the left, there are sections for "You might like" and "You recently visited", each displaying small profile pictures. In the center, there's a "My self-summary" section with a green checkmark icon, followed by a detailed text about the user's background and interests. To the right, there's a "My Details" section with a grid of checkboxes and text entries for various demographic and lifestyle details. A sidebar on the right shows a profile picture and a message indicating 514 search results in the last 24 hours.

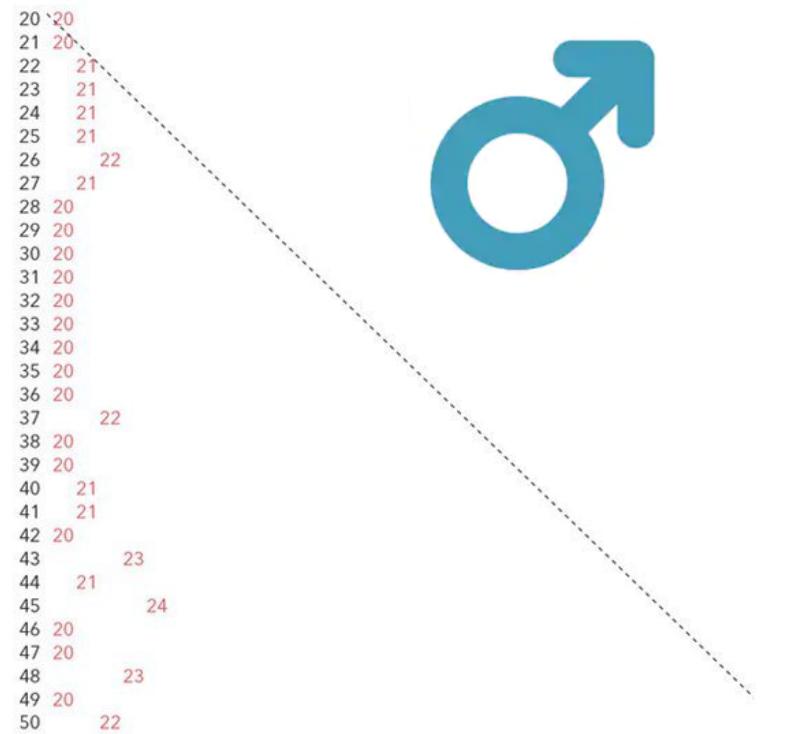
计算传播学的数据优势

- 收集个体自发产生的行为数据
 - 比如：在线交友网站的数据.....

a woman's age vs. the age of the men who look best to her



a man's age vs. the age of the women who look best to him



传统传播学收集数据的局限性

- 一次性：大多数研究基于一次性的数据收集
 - 举例：情绪变化（Mood Vary）



人的情绪在一天的不同时刻
是会发生变化的

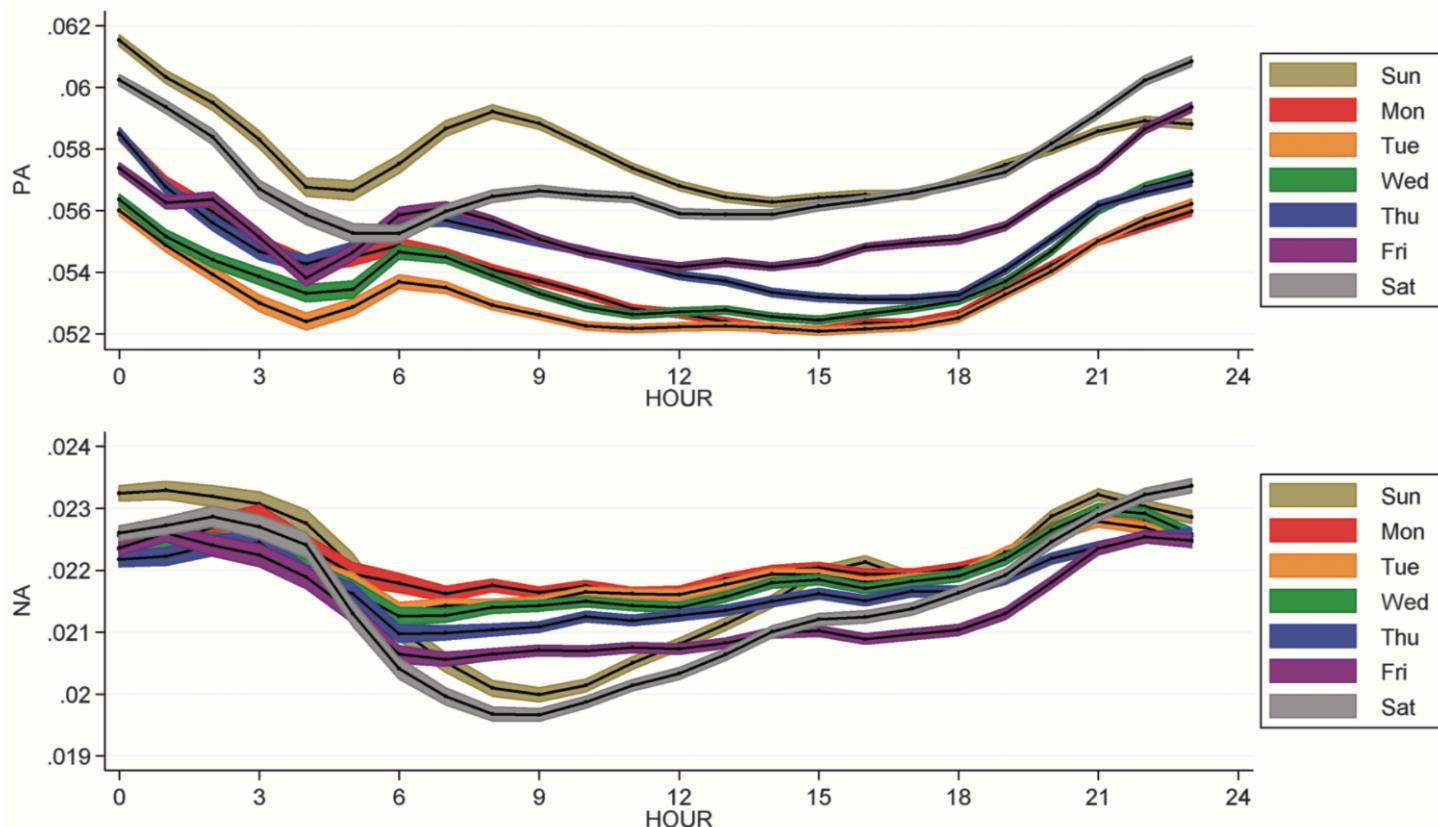
传统方法：

- 选定一组人群作为样本
- 记录他们的情绪变化

除了规模小、有偏差之外，
如何对人群进行持续观察也
成为传统方法的局限性。

传统传播学收集数据的局限性

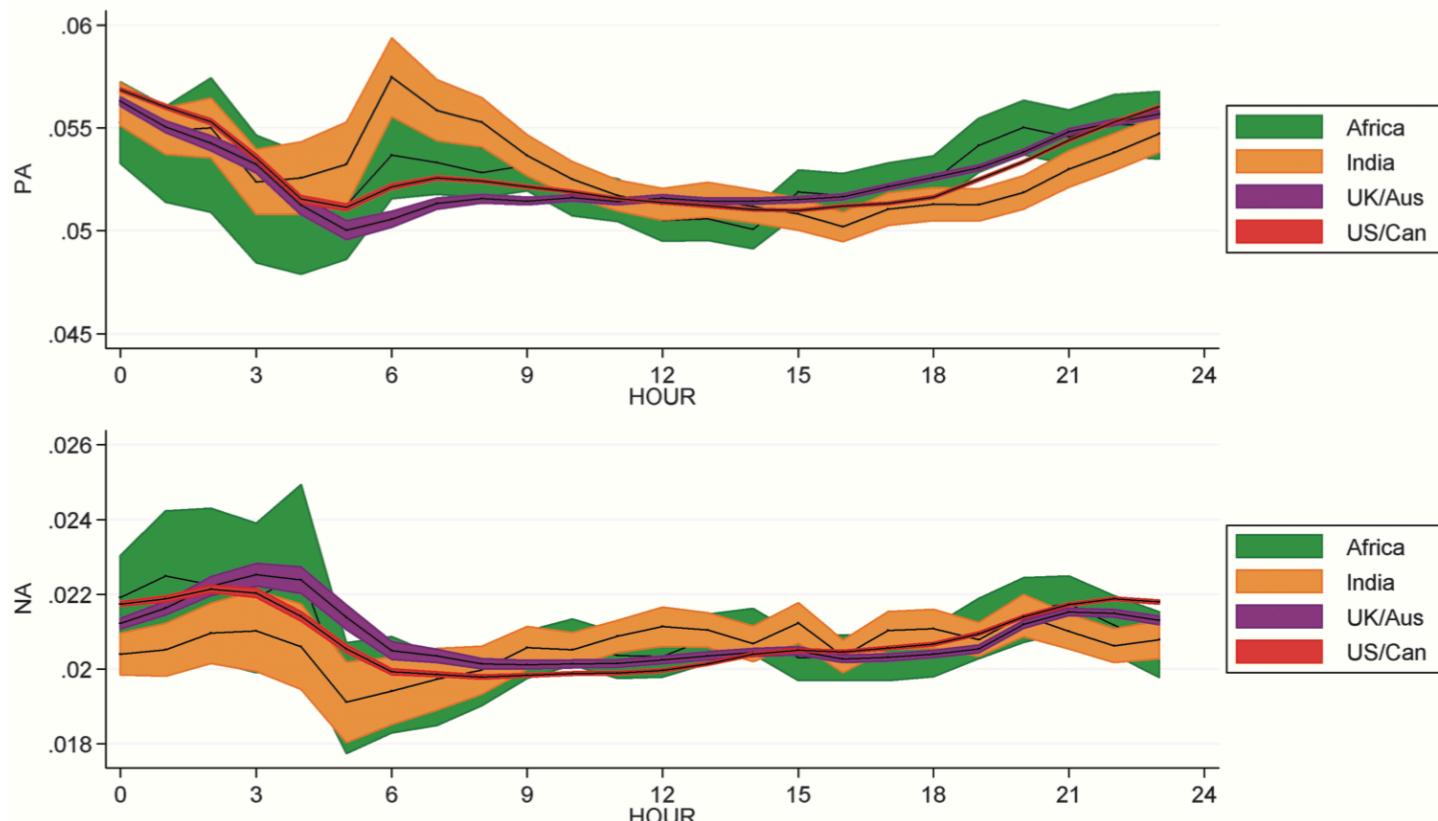
- 收集个体持续产生的行为数据
 - 针对用户每天发布的Tweets进行文本分析



Source: Golder and Macy: Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. Science, VOL 333, 2011

传统传播学收集数据的局限性

- 收集个体持续产生的行为数据
 - 针对用户每天发布的Tweets进行文本分析



Source: Golder and Macy: Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. Science, VOL 333, 2011

计算传播理论与实务



大数据带来了.....

新方法

- Big Data is not only about data
 - 挖掘人类行为背后的模式和法则
 - 解释模式背后的生成机制与基本原理

规模很重要！

- One thousand data instances
- One million data instances
- One billion data instances
- One trillion data instances
- Those are not different **numbers**, those are different **mindsets**

思考题

- 上述数据规模的变化，给我们在数据收集、处理和分析时的思维方式带来了怎样的变化？

新的研究方法

数据科学

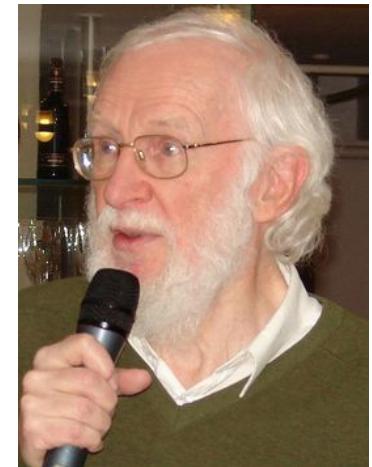
- 数据科学是一门新兴学科，它的定义仍在发展
- The exact role, background, and skill-set, of a data scientist are still in the process of being defined

数据科学与大数据

- 数据科学是早于大数据出现的一个概念
- 大数据热潮促进了人们对数据科学的重视，并促进了独立的数据科学交叉学科的发展
- 大数据是数据科学的一部分

现状和历史

- 1974年，著名计算机科学家、图灵奖获得者Peter Naur在其著作 *Concise Survey of Computer Methods* 的前言中首次明确提出了数据科学(Data Science)的概念：
- 数据科学是一门基于数据处理的科学
- “The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.”

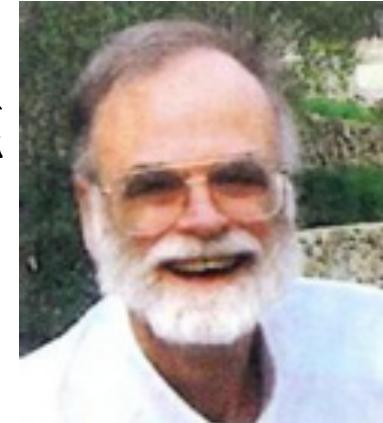


现状和历史

- The technical areas of data science are those that have an impact on how a data analyst analyses data:
 - (1) Statistical theory;
 - (2) Statistical models;
 - (3) Statistical and machine-learning methods;
 - (4) Algorithms for statistical and machine-learning methods, and optimization;
 - (5) Computational systems for data analysis;
 - (6) Live analyses of data where results are judged by the findings, not the methodology and systems that were used.

现状和历史

- 2007年，图灵奖得主Jim Gray提出数据密集型科学为科学的第四范式
 - Empirical (实验科学，实验归纳) - 钻模取火
 - Theoretical (理论科学，归纳总结) - 牛顿三定律
 - Computational (计算科学，计算机仿真) - 模拟核试验，天气预报
 - And now data-driven (**data-Intensive**)
- Data-driven science is the "**fourth paradigm**" of science that uses the **computational analysis of large data** as primary scientific method and "to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other".



小科普：什么是图灵奖



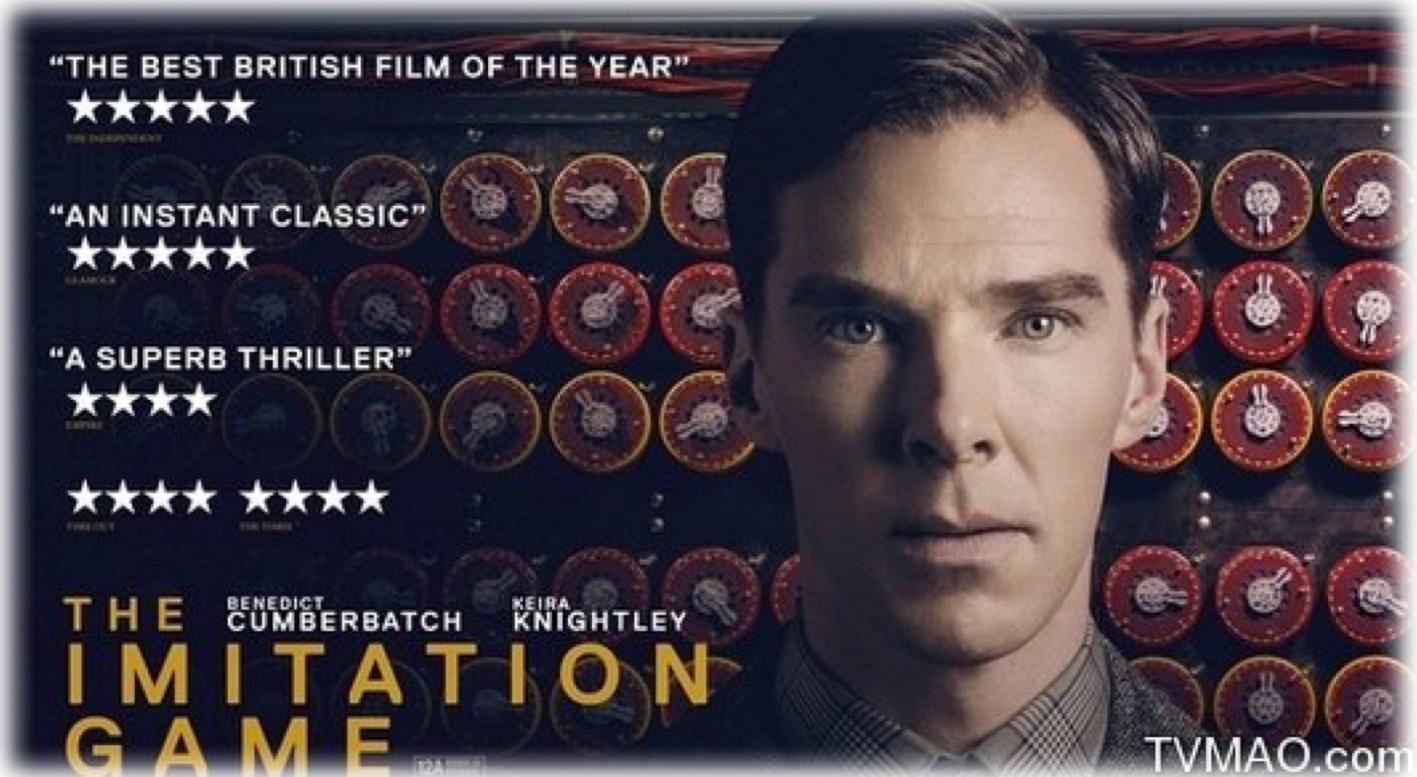
计算机领域的诺贝尔奖

阿兰·图灵是谁

- 英国司法部长克里斯2013年12月24日宣布，英国女王伊莉莎白二世赦免上世纪50年代因同性恋行为被定罪的英国著名数学家、密码学家、计算机科学之父阿兰·图灵(Alan Turing)。图灵在去世59年后，终于获得了英国皇室的“皇家赦免”。
- 阿兰·图灵是谁？

阿兰·图灵是谁

- 电影：模仿游戏（2014年）



人工智能的图灵测试

- 图灵设想了一个游戏，房间里有两个人，一男一女，房间外面有一个人，这个人可以提问题，里面的两个人通过写字来回答，然后他要猜测，里面哪个人是女人。男人要设法欺骗猜测者，而女人则要设法使猜测者相信自己，所以他们都说：“我才是女人，你不要相信他。”……图灵本来想表达的意思，其实是这种模仿原则在思维和智能问题上的应用。把这一男一女，换成一个人和一台计算机，如果猜测者根据写出来的回答，无法辨别哪个是人，哪个是计算机，那么本着“公平对待机器”的思想，就必须承认计算机具有“智能”。
 - ——《阿兰·图灵传》

现状和历史

- 2009年，Google首席经济学家 Hal Varian 强调了数据科学的重要性
 - The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades.



现状和历史

- 2010年，数据科学韦恩图
- Drew Conway, CEO and founder of [Alluvium](#)

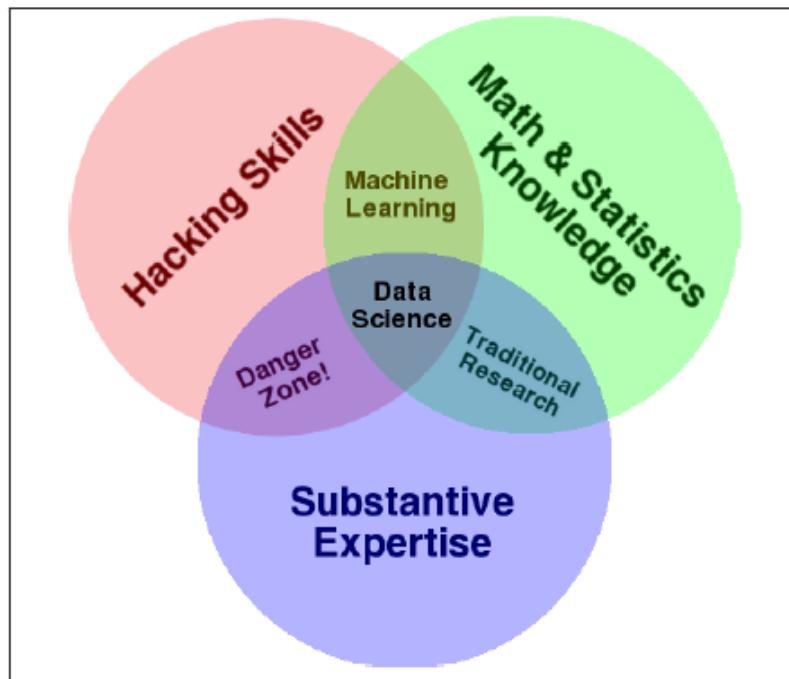
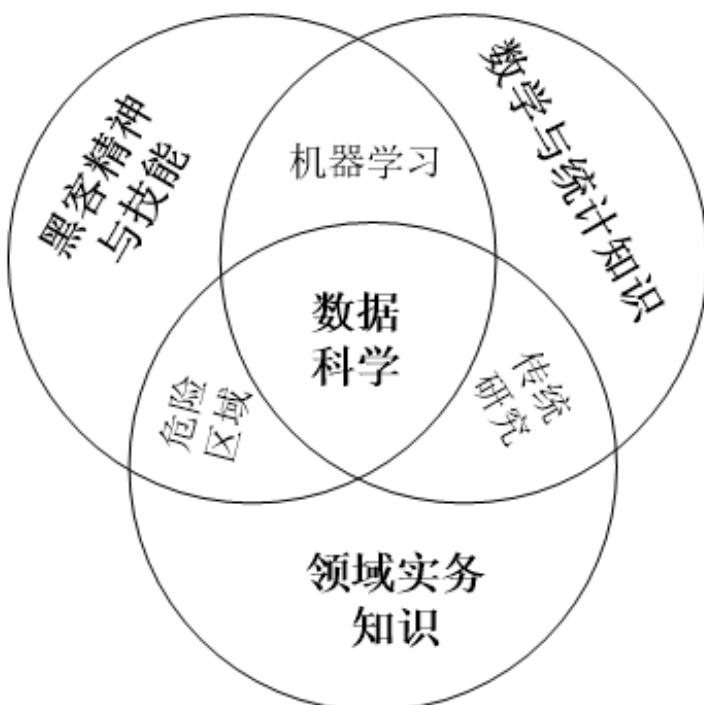


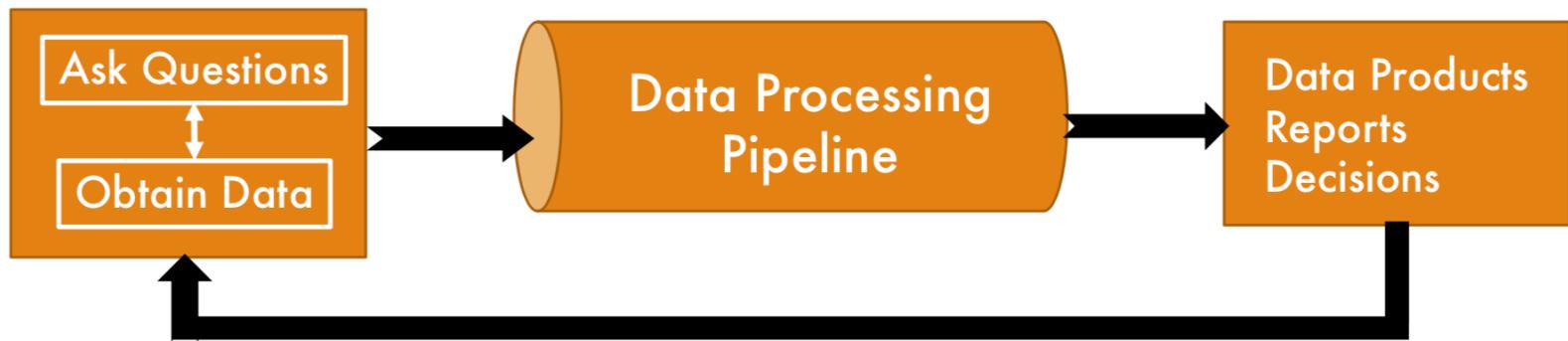
Figure 1-1. Drew Conway's Venn diagram of data science

Source: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>



数据科学的工作流程

- 循环迭代式的工作流程



- 两类工作方式
 - 先提出问题，再收集与分析相关的数据
 - 先收集数据，再分析可以回答哪些问题

数据科学与烹饪



采集



清洗



集成



分析

- 两类烹饪方式：
 - 先想好做啥菜，再采办与处理原材料
 - 先买好原材料，再看看能做成什么菜

数据科学能够回答哪些问题？

- 问题 I:

Is This A or B?

- 分类问题 (Classification)
 - 这张照片中是猫还是狗？
 - 这条评论体现了积极还是消极情绪？
 - 这个用户是否会点击给他推送的广告？

数据科学能够回答哪些问题？

- 问题2：

How much or How Many?

- 回归问题（Regression）
 - 我发完这条微博会涨多少粉？
 - 下礼拜一的气温会是多少？
 - 中国第四季度的GDP增长会是多少？

数据科学能够回答哪些问题？

- 问题3：

Is This Weird?

- 异常检测（Outlier Detection）
 - 那个群体的情绪变化与大众明显不同？
 - 那个用户最近买的商品与以往显著不同？
 - 哪条借贷记录属于欺诈行为？

数据科学能够回答哪些问题？

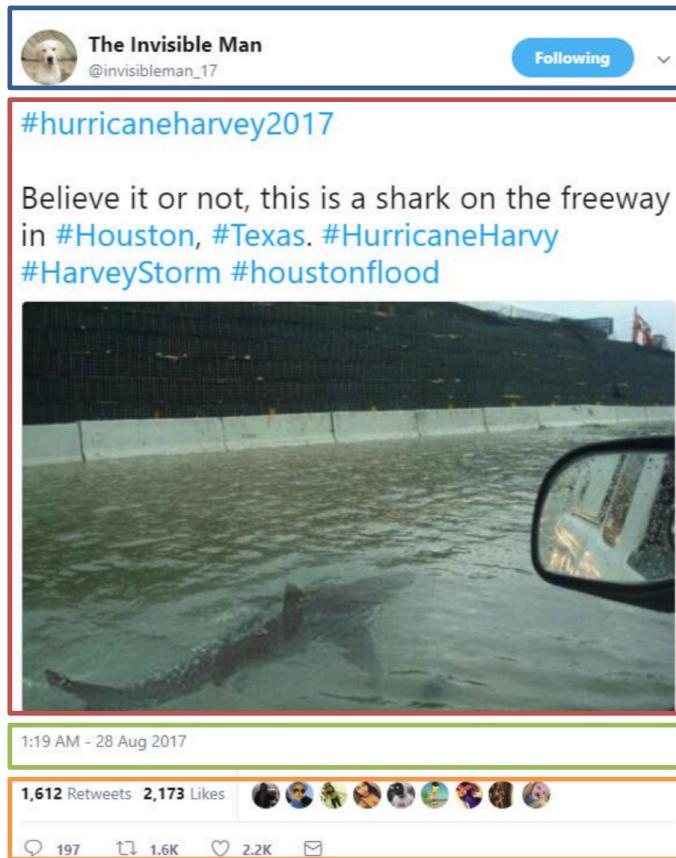
- 问题4:

How Is This Organized?

- 聚类问题 (Clustering)
 - 哪些评论在话题上非常类似？
 - 哪些顾客在购物行为上非常类似？
 - 哪些用户喜欢同样类型的电影？

案例分析

• 虚假信息监测属于什么问题？



虚假信息散布者

虚假信息内容

- 文本
 - 文字
 - 标签
 - 链接
 - 表情
- 图片
- 视频 (GIF)

虚假信息时空属性

虚假信息传播者

- 日期、时间
 - 地点
- 点赞、回复
 - 转发

案例分析

- 广告互动率预测属于什么问题?
 - 微信朋友圈广告是典型的feeds流广告
 - feeds广告就是与内容混排在一起的广告：
 - 最不像广告的广告，长得最像内容的广告。
 - Feeds广告操作性简单，打扰性低，已经成为移动互联网时代主流的广告形式。
 - 建立在用户行为记录和大数据分析基础上，个性化推荐



整体-大盘-曝光量	:	1,686,709,174
整体-大盘-可视曝光	:	1,015,382,271
整体-大盘-收入 (元)	:	66,735,746

案例分析

- Target的神预测

- 2012年，明尼苏达州一家Target门店被客户投诉，一位中年男子指控Target将婴儿产品优惠券，寄给他的女儿，而他的女儿只是一个高中生，实在不可理喻。
- 但是没有过多久，他却给Target来电道歉，因为经他逼问，他女儿后承认自己真的怀孕了。这位高中生没有告诉过父亲她怀孕了，也没有在Target调查问卷上留下过类似的记录。

- Target的数据分析师开发了怀孕预测模型

- Target通过分析这位女孩的购买记录——无味湿纸巾和补镁药品就预测到了这为女顾客可能怀孕了
- 怀孕了未来就有可能需要购置婴儿服装和孕妇服装



大数据带来了.....

新問題

- Computational Communication = Computer Science + Communication Data?
 - 关联关系 VS 因果关系
 - 大数据带来的伦理问题

回顾前面的案例

- Target的数据分析师开发了怀孕预测模型
 - Target通过分析这位女孩的购买记录——无味湿纸巾和补镁药品就预测到了这位女顾客可能怀孕了
 - 因为购买了“无味湿纸巾和补镁药品”，所以就怀孕吗？
- 数据科学模型与社会科学模型的不同之处
 - Computer scientists may be interested in finding the needle in the haystack—such as [...] the right Web page to display from a search—but social scientists are more commonly interested in characterizing the haystack.

数据科学的目标

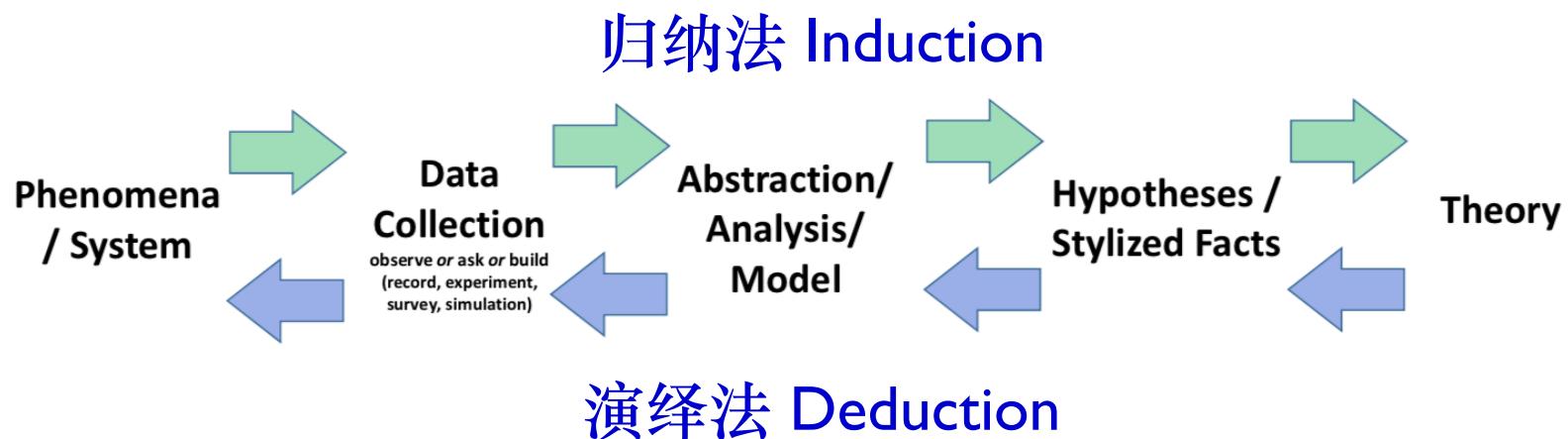
- Prediction
 - We do not care why a model makes good predictions; we just care that it does.
- 啤酒与尿布——数据挖掘的著名案例
 - “啤酒” 和 “尿布” 两个看上去没有关系的商品摆放在一起进行销售、并获得了很好的销售收益
- 为什么?
 - 解释：年轻的父亲前去超市购买尿布，顺便为自己购买啤酒
 - 这个解释一定靠谱吗？我们不关心！

回顾：计算科学认为的“智能”

- 图灵设想了一个游戏，房间里有两个人，一男一女，房间外面有一个人，这个人可以提问题，里面的两个人通过写字来回答，然后他要猜测，里面哪个人是女人。男人要设法欺骗猜测者，而女人则要设法使猜测者相信自己，所以他们都说：“我才是女人，你不要相信他。”……图灵本来想表达的意思，其实是这种模仿原则在思维和智能问题上的应用。把这一男一女，换成一个人和一台计算机，如果猜测者根据写出来的回答，无法辨别哪个是人，哪个是计算机，那么本着“公平对待机器”的思想，就必须承认计算机具有“智能”。
 - ——《阿兰·图灵传》

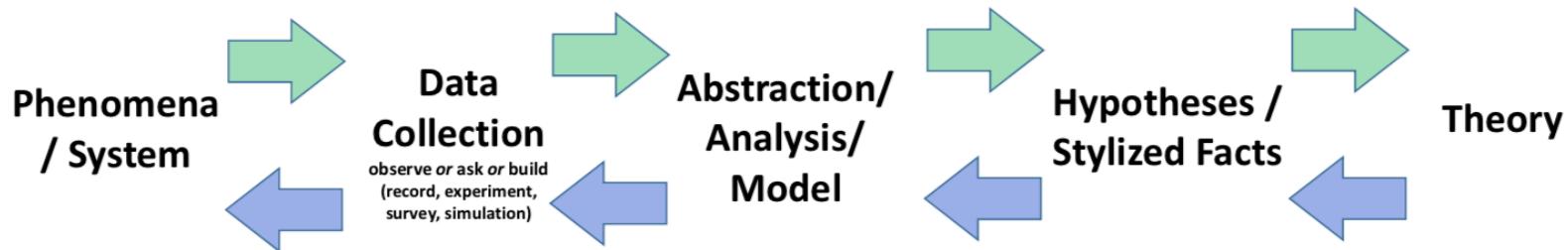
社会科学的目标

- Explanation
 - The goal is to use observed data to provide evidence in support or opposition of causal explanations
- 社会科学研究方法



本门课的侧重点

- “传播为体”：立足传播学的研究方法



- “计算为用”：使用数据科学的计算手段实现大规模数据收集与自动分析



钢铁侠
还是
终结者



案例分析

- 理论Theory: 二级传播 (Two-Step Flow)
 - 来自大众媒体的影响首先到达舆论领袖那里，舆论领袖再把他们读到和听到的内容传达给受他们影响的人
- 假设hypothesis
 - 信息在社交网络传播时，不同人的“影响力”是不同的（例如：存在大V和小透明）
- 数据收集：如收集知乎数据
- 数据分析：定义什么是“影响力”
- 数据可视化：有效地进行沟通和展示

案例分析

- 一夜之间从知乎小透明到万赞。。。需要多少个大V扶持呢？



数据伦理问题

- Quarks and cells neither mind when we discover their secrets nor protest if we alter their environments during the discovery process.
- 计算传播学能否把算法当做“黑箱”？



信息茧房

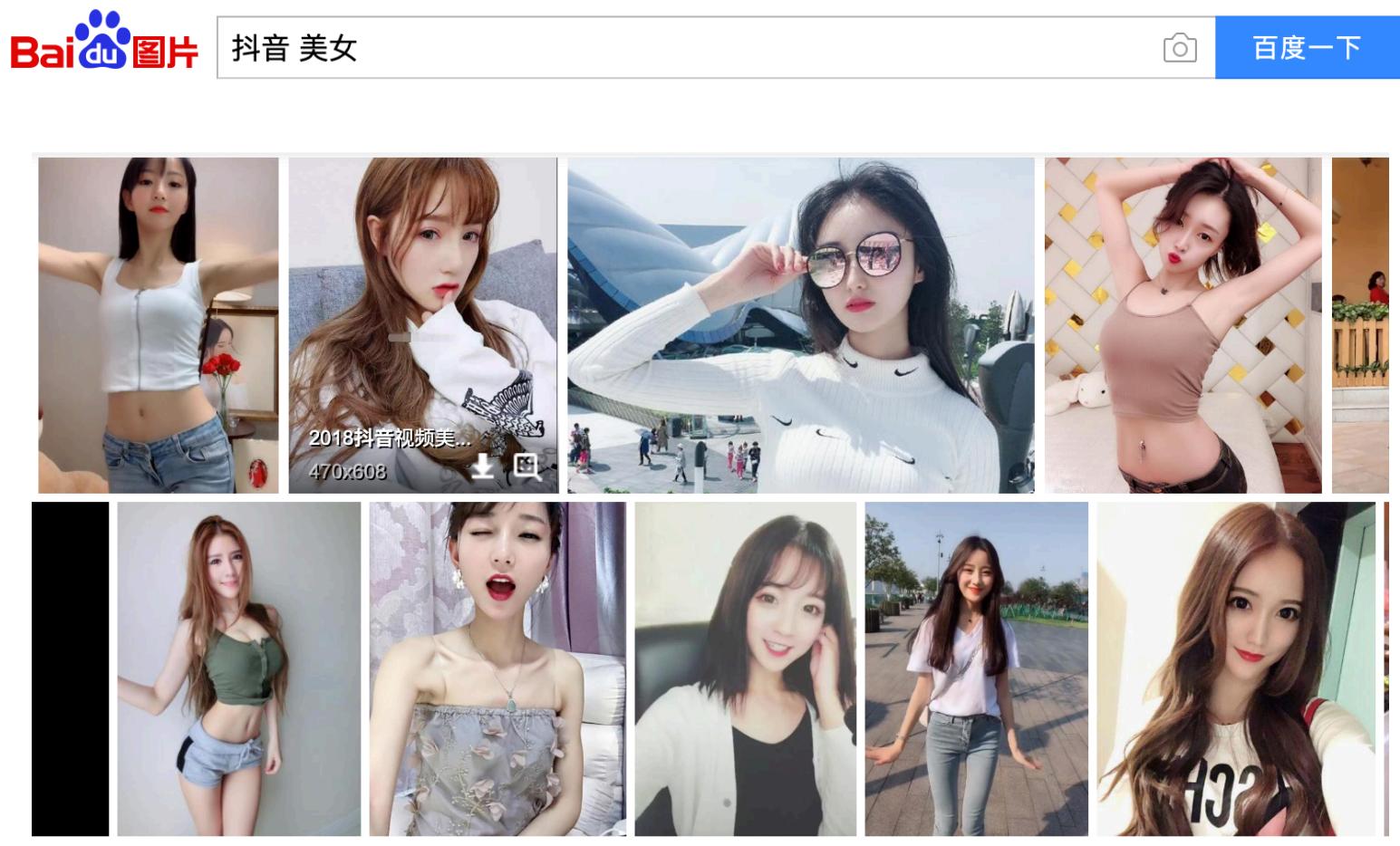
• 曾经的女神



<http://csfr1840.blog.sohu.com>



信息茧房



数据隐私



首页
快讯
▲ 资讯
视频
专题
活动

搜索

寻求报道
我要投稿
寻求融资

知料 | ZAO了一天，隐私的雷快爆了

九连环 · 2019-08-31

支付宝表示安全。

| 知料是36氪推出的新栏目，挖掘新闻背后那些你需要知道的料，欢迎继续关注。

文 | 方婷 史圣园

内容提要

- I.1 计算传播学：大数据带来了什么？
- I.2 教学计划
- I.3 考核要求

课程大纲

- 第一讲 计算传播学简介
- 第二讲 数据采集与数据预处理
- 第三讲 文本分析
 - 文本检索
 - 情感分析
 - 主题建模
- 第四讲 网络分析
 - 中心性分析
 - PageRank分析
 - 信息传播分析

课程大纲

- 第五讲 数据伦理
 - 信息茧房
 - 隐私保护
 - 算法歧视
- 实践环节
 - 实验1：依托微博/微信数据，完成某一热门事件或公共议题的文本分析（情感或主题分析）
 - 实验2：依托社交网络数据，完成社交网络上的KOL（意见领袖）分析
 - 大作业：自选

有关实践环节



你认为的程序设计

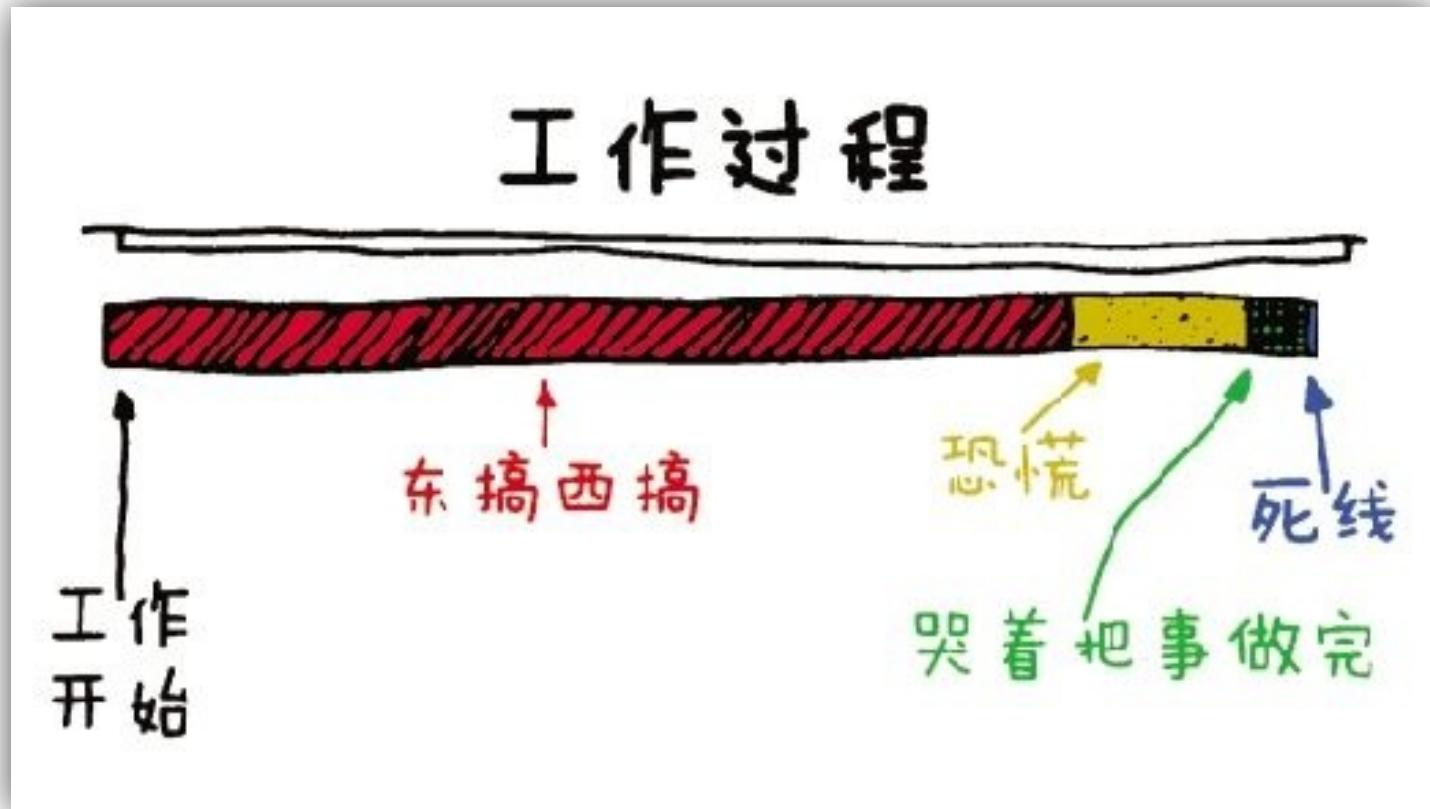


实际上的程序设计

- 战略上藐视敌人，战术上重视敌人
 - 交互式工具 + 编程辅导

温馨提示：应该避免.....

- 拖延晚期 (Procrastination Cancer)



内容提要

- I.1 计算传播学：大数据带来了什么？
- I.2 教学计划
- I.3 考核要求

本堂课怎么考？

- 期末成绩：30%
 - 期末考试
- 平时成绩：70%
 - 两次实验：30%
 - 大作业：30%
 - 课堂表现：10%

总结

- I.1 计算传播学：大数据带来了什么？
- I.2 教学计划
- I.3 考核要求

社交媒体虚假信息示例

 The Invisible Man
@invisleman_17

Following ▾

#hurricaneharvey2017

Believe it or not, this is a shark on the freeway
in #Houston, #Texas. #HurricaneHarvy
#HarveyStorm #houstonflood



1:19 AM - 28 Aug 2017

1,612 Retweets 2,173 Likes

197 1.6K 2.2K

社交媒体虚假信息包含哪些数据



虚假信息散布者

虚假信息内容

- 文本
 - 文字
 - 标签
 - 链接
 - 表情
- 图片
- 视频 (GIF)

虚假信息时空属性

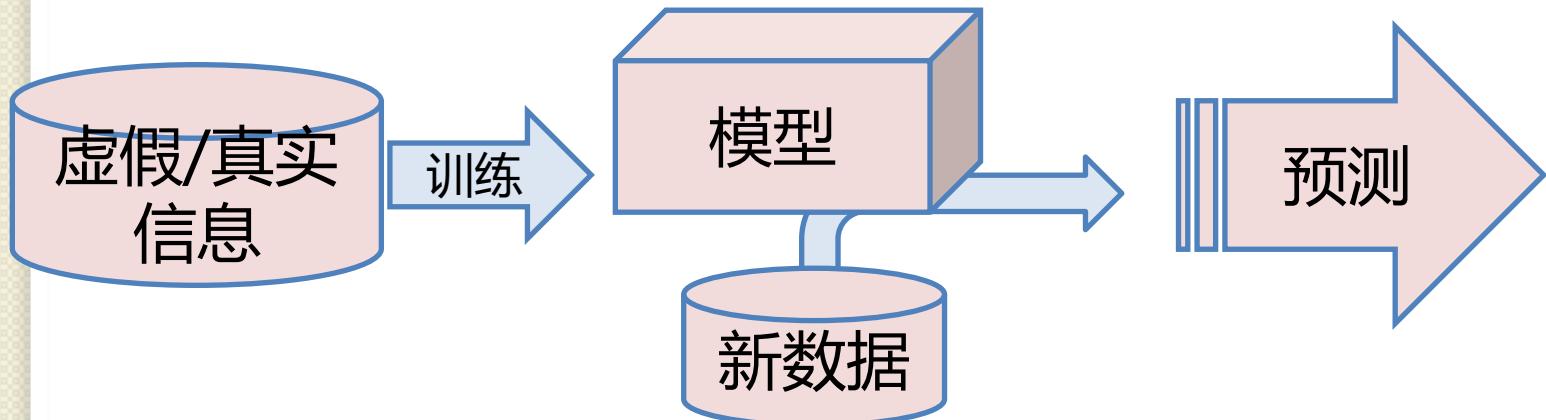
- 日期、时间
- 地点

虚假信息传播者

- 点赞、回复
- 转发

基于内容的虚假信息监测

- 建模为机器学习的**分类**问题
- 以社交媒体的帖子（post）为例
 - 输入：一组帖子的集合
 - 输出：为每个帖子打上标签（1：虚假； -1：真实）

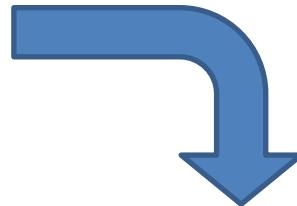


信息内容的特征工程

- 将每篇帖子变成高维空间上的一个点

#hurricaneharvey2017

Believe it or not, this is a shark on the freeway
in #Houston, #Texas. #HurricaneHarvy
#HarveyStorm #houstonflood



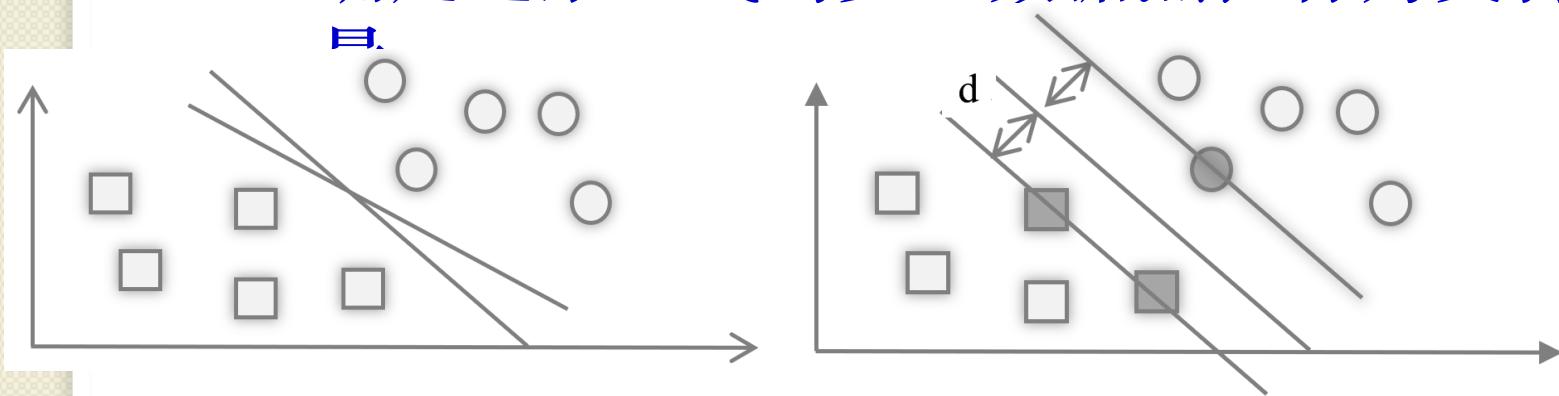
Text Feature	Example
Length of post	#words, #characters
Punctuation marks	Question mark ? Exclamation!
Emojis/Emoticons	Angry face ;-L 😠
Sentiment	Sentiment/swear/curse words
Pronoun (1 st , 2 nd , 3 rd)	I, me, myself, my, mine
URL, PageRank of domain	@AndrewGirdwood Have you heard Google was hiring people to work from home? pretty cool i thought
Mention (@)	http://dwarfurl.com/1f291
Hashtag (#)	Free President #Gbagbo #8demarzo #stopkony #SOLARSTORM #iwd #fmmedia12 #BarackObama #cnn #breakingNews #Syria #sarkozy

基于匹配的虚假信息监测

- 基本想法
 - 某个新帖子与训练数据中的虚假帖子越像，则越可能是虚假帖子
- 如何度量什么叫“像”？
 - 精确匹配：一模一样
 - 文本相关度：TF-IDF、BM25
 - 语义相关度：Word2Vec、Doc2Vec
- 缺点：召回率低

基于分类模型的虚假信息监测

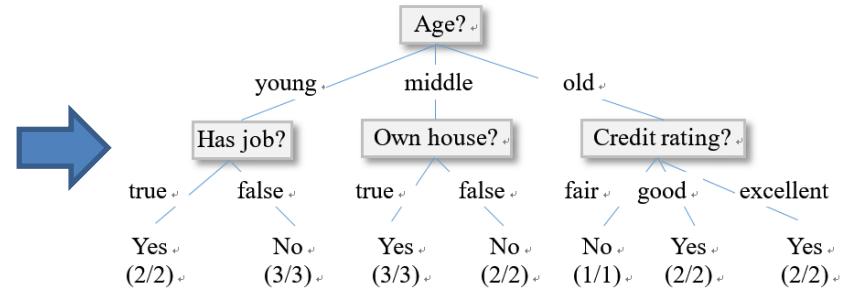
- 使用机器学习中的分类模型
- 分类模型：以SVM为例
 - 二维平面上把两类点分开，可用的直线有多条
 - 最优的那条，到两类数据点的距离都最大
 - 确定这条直线的少量数据点，称为支持向量



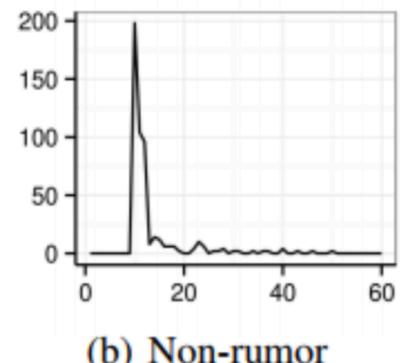
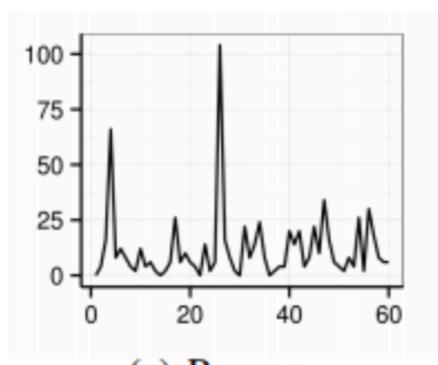
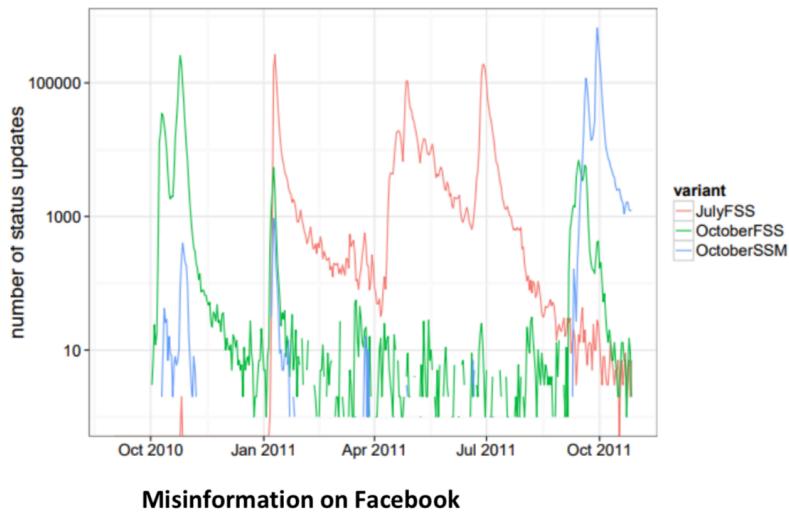
基于分类模型的虚假信息监测

- 使用机器学习中的分类模型
- 分类模型：以决策树模型为例
 - 决策树中的非叶子节点，表示对象属性的判断条件，其分支表示符合节点条件的所有对象，树的叶子节点表示对象所属的预测结果。

ID	Age	Has Job	Own House	Credit Rating	Class
1	Young	false	false	fair	No
2	Young	false	false	Good	No
3	Young	true	false	Good	Yes
4	Young	true	true	fair	Yes
5	Young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	True	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No



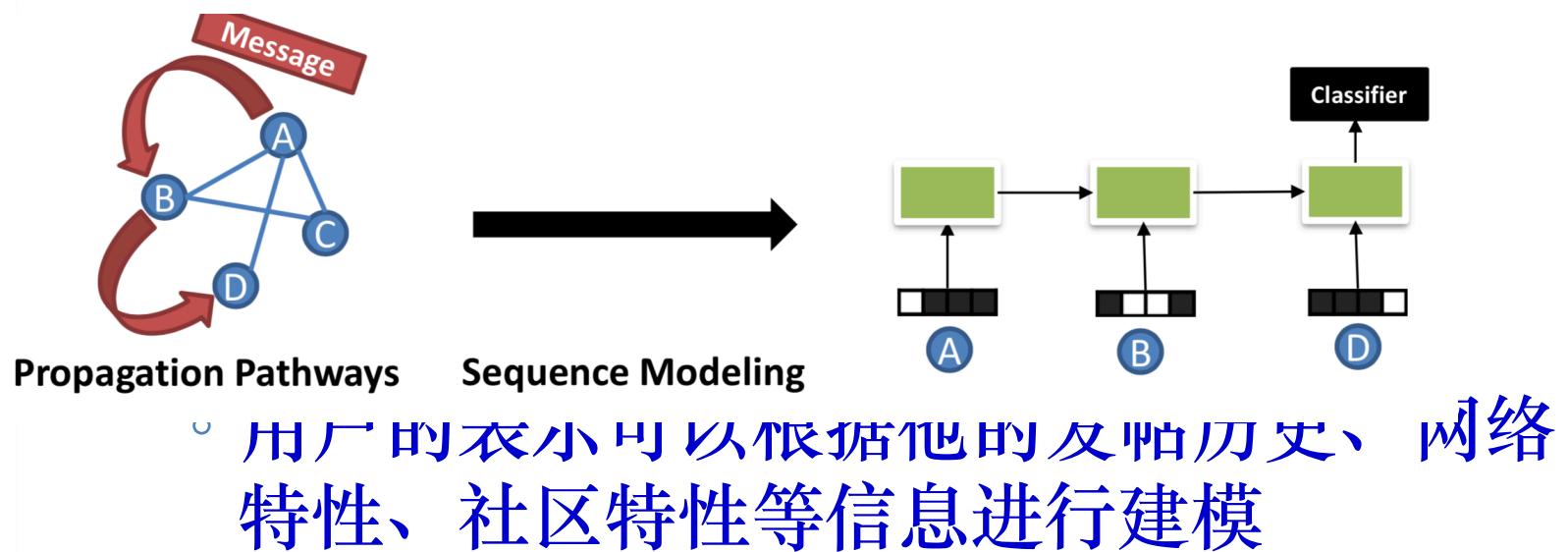
基于时空属性的检测方法



- 虚假信息在时空属性上的特性
 - 容易在很短时间或局部空间上突发
 - 同样（类似）的信息在时间上出现多次突发

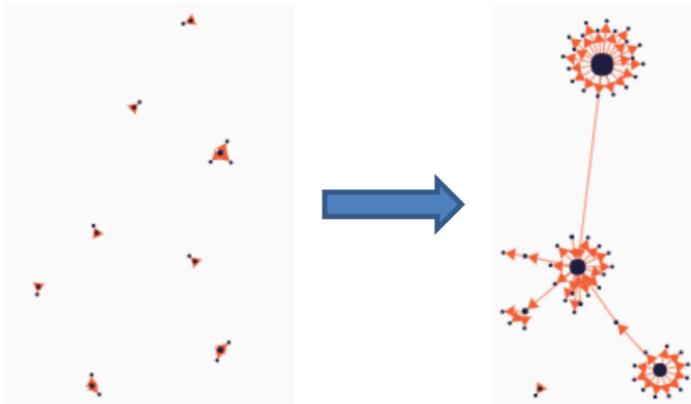
基于传播的检测方法

- 基本想法
 - 虚假信息可以由谁传播、以及在社交网络中如何进行传播来进行区分



虚假信息监测还是个开放的问题

- 早期监测缺乏可用的数据



- 缺乏高质的数据标注
 - 传统的半监督学习采用相似性传播标签
 - 然而虚假信息经常会“改头换面”