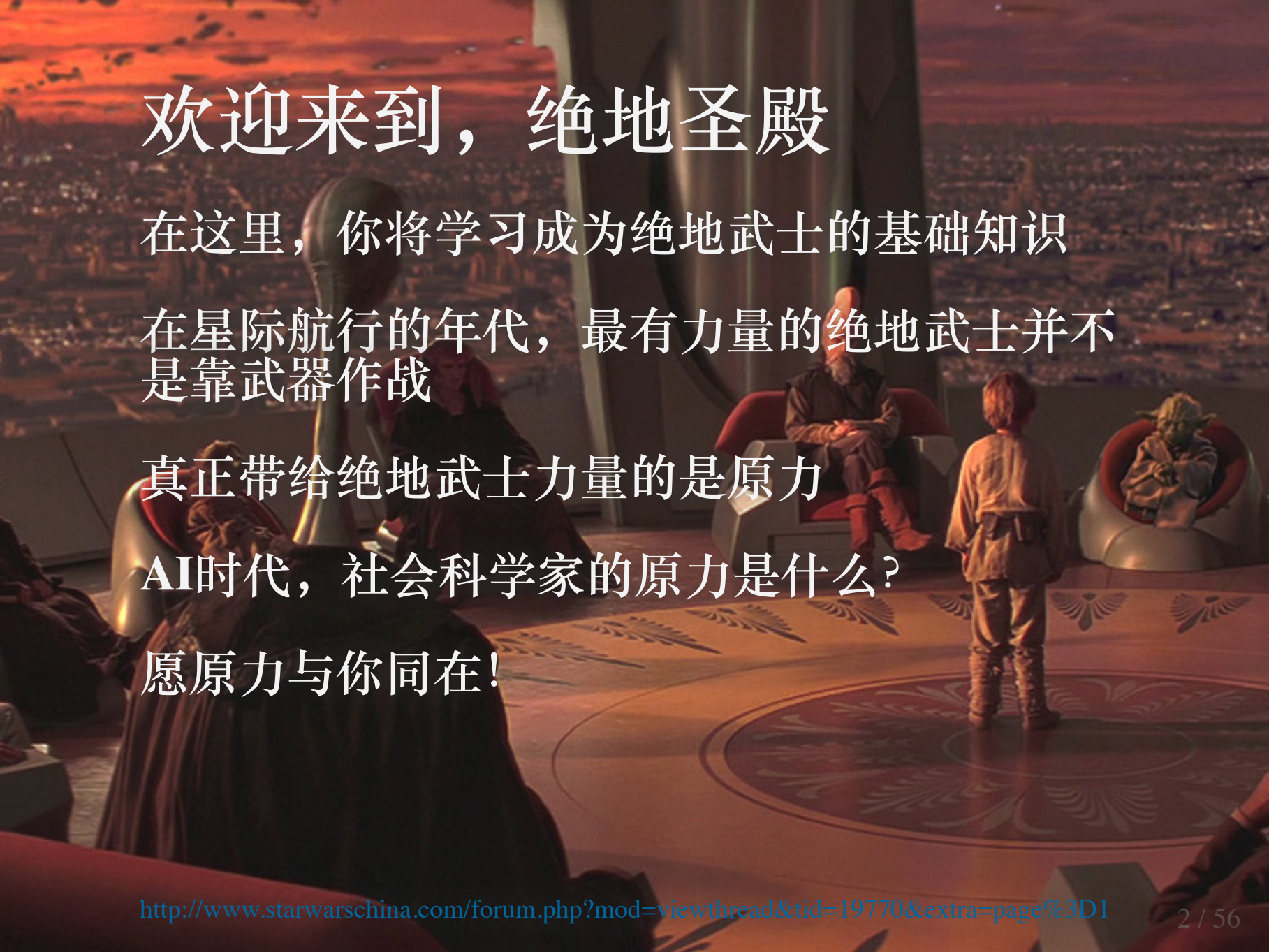

# AI时代社会科学研究的挑战与机遇

## 科学、社会科学、数据工程、工具箱

史冬波

2020年3月2日



# 欢迎来到，绝地圣殿

在这里，你将学习成为绝地武士的基础知识

在星际航行的年代，最有力量的绝地武士并不是靠武器作战

真正带给绝地武士力量的是原力

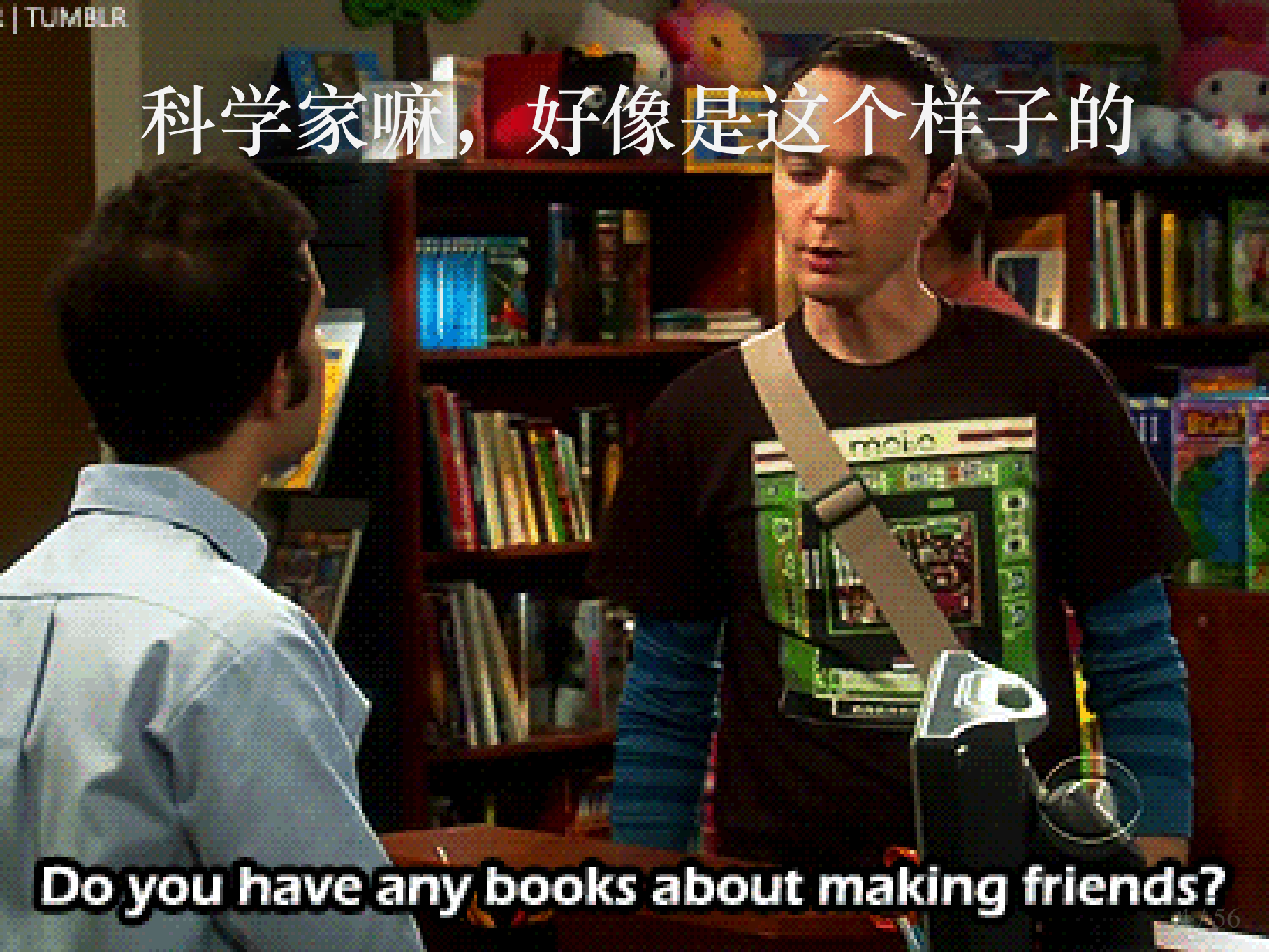
AI时代，社会科学家的原力是什么?

愿原力与你同在!

http://www.starwarschina.com/forum.php?mod=viewthread&tid=19770&extra=page%3D1

# 什么是原力?

首先要从社会科学是什么谈起，社会科学，拆开来看，就是社会+科学


| TUMBLR
科学家嘛，好像是这个样子的
Do you have any books about making friends?

莫非，社会科学家就是很社会的科学家！

# 今天的主要内容

**1.现代科学的特征与转型** 宇称不守恒实验、暗物质测量、烟草与肺癌、最低工资的争论、高速公路的限速

**2.对社会科学的再认知** 科学不是玄学、科学化程度决定了鄙视链、社会科学研究对国家民族的重要性前所未有

**3.人工智能带给社会科学的冲击以及机会** 人工智能的三要素（大数据、深度学习、云计算）、总体、工具、因果推断

**4.数据工程的原则与思想-原力！** 可复现、自动化、正交分工、最佳工具

**5.社会科学家是可以也是应当和科学家对等对话的！** 同一个世界，同一个梦想，更难的数据和对象

# 宇称不守恒实验

科技猿人-2019年12月29日

# 实验的重要意义

- 宇称不守恒的发现打破了人们对上帝绝对对称的信念，实际上这是一个非常惊人的事情，迫使人们重新思考对称的问题，这一转向导致了后来许多深刻的发现。
- 杨振宁和李政道在1956年10月发表了《对于弱相互作用中宇称守恒的质疑》的论文，吴健雄随后给了实验验证，诺组委立马把1957年的诺贝尔奖颁给了35岁的杨振宁和31岁的李政道。
- 作一个不严谨的对比说明。爱因斯坦在1905年提出来光量子说和狭义相对论，1910年俄国科学家列别捷夫完成光压实验，证实了光量子假说，但是组委一直拖拖拉拉到1921年才授予爱伊斯坦诺贝尔奖。
- 对中国人而言还有一个意义，引用杨振宁先生的说法，即“我得诺贝尔奖最重要的贡献就是帮助中国人改变了自觉不如人的这个心理”，这里的中国人是指得奖的时候，两人均为中国国籍。

# 定义左右

设想一下如下的场景，你在红岸基地工作，收到了一封外新人的“来信”：

- 当然，你的第一反应是？不要回答！不要回答！不要回答！
- 我们假设不是黑暗森林法则。于是你和外星人谈笑风声，外星人问你，你们人类有多高？这个时候应该怎么回答？
- 你可以告诉外星人，我们中国人的平均身高大概是1.7米。可是外星人不知道什么是“米”，这样难不住你，你可以告诉他你找个氢原子，1.7米就是氢原子半径的340亿倍。
- 接下来，你给外星人听了一首《野狼disco》，外星人问，怎么在左边划条龙呢？这个时候犯难了。平时我们是怎么定义左右的呢？“左手的方向是左”，这一个死记硬背呀！
- 为什么定义长度很容易？定义左右这么难呢？这是因为氢原子半径是一个物理规律不变量，全宇宙是统一的。但是左右是一对镜像关系，就好像我们面对一面镜子，录下来一段录像。只看录像的时候，是没办法翻遍哪里是镜子里，哪里是镜子外的。因为镜像世界与现实世界服从同样的物理规律，这就是宇（空间）称（对称）守恒。我们看到的物理规律都是满足空间对称性的。

PHYSICAL REVIEW VOLUME 104, NUMBER 1 OCTOBER 1, 1956

# Question of Parity Conservation in Weak Interactions*

T. D. LEE, *Columbia University, New York, New York*

AND

C. N. YANG,† *Brookhaven National Laboratory, Upton, New York*

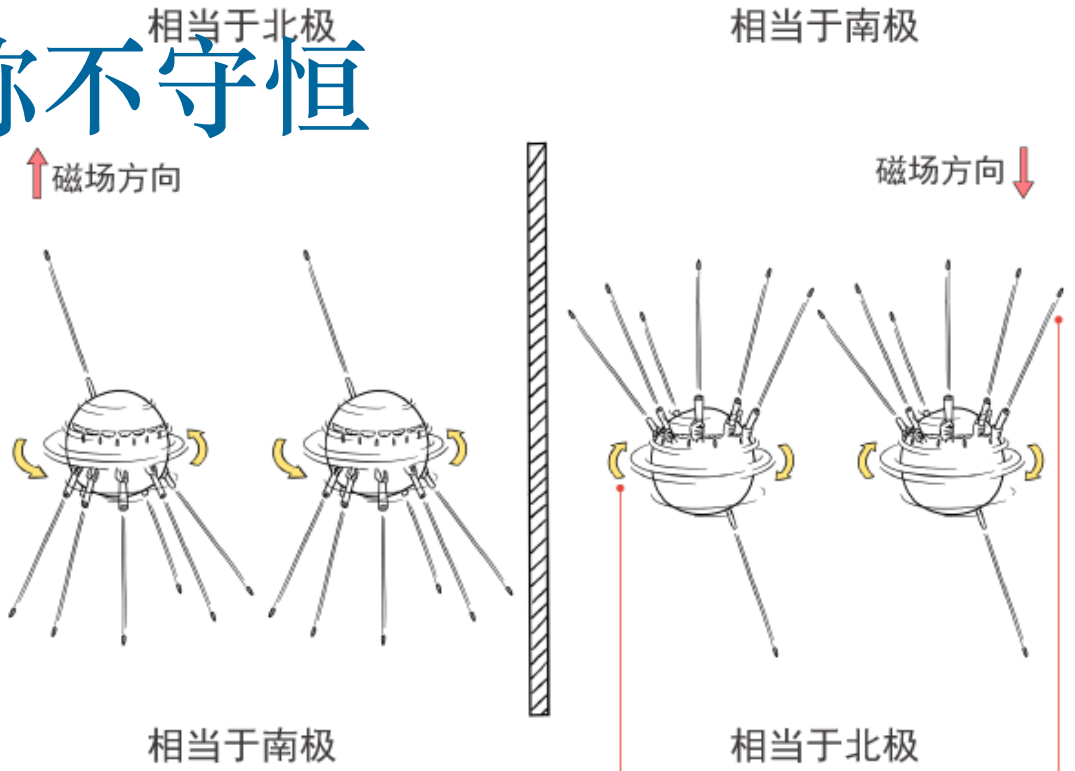
(Received June 22, 1956)

The question of parity conservation in $\beta$ decays and in hyperon and meson decays is examined. Possible experiments are suggested which might test parity conservation in these interactions.

# 宇称并不守恒

- Lee and Yang(1956)年的论文告诉我们，可以在极低温的环境下给大量的Co-60原子加磁场，然后观察他们发射出的电子角度分布。假设宇称是不守恒的，那么电子将会从磁场相反的方向发射出。这个时候我们可以把电子发射方向成为上，磁场方向成为下。而我们在定义磁场方向的时候用到了安培定则，也就是说定义磁场方向为上的手，便是右手。
- 简单来说，所有涉及弱相互作用（质子和中子发生转化的作用便是弱相互作用）的规律都是宇称不守恒的。
- 李杨的贡献便是指出弱相互作用可能是宇称不守恒的，而且建议了用来验证该假设的实验（上述实验）。几个月后，著名的华人实验物理学家吴健雄女士就用实验证实了这个理论。
- 杨振宁之所以有这样的怀疑是因为当时实验中发现了两种例子，即theta粒子和tau粒子，但是实验发现这两种粒子的所有物理性质，除了宇称以外都是一样的。那很自然的问题就是，这是两种例子还是一种粒子。如果是两种粒子，为什么所有性质都相同?

# 宇称不守恒




钴-60原子转动方向相反，相当于照镜子。
一照镜子，原子的“南北极”就对调了。

但是β射线总是往南极走，这就破坏了镜像对称。


于是，你就可以和外星人一起开心看香密了！当然你也可以教他一起左边一起划个龙，右边划一道彩虹。

# 科学革命的结构

- 回到上面的实验过程，我们可以发现实际上科学的进展，是这样的一个过程，首先是出现一个基于鲜有科学理论无法解释的现象（两种粒子之谜），然后不得不更新的理论假设（宇称不守恒），新的假设被验证（Go-60实验），诞生新的理论
- 这个过程被库恩概括为：前科学时期——常规科学——反常与危机——科学革命——新的常规科学
- 假设的验证，往往是在理论指导下完成的，这种科学的研究范式又被称为“第二范式”（吉姆·格雷，图灵奖得主）

# 探寻暗物质

上海交通大学PandaX暗物质实验，《超级实验室》（第一集征程）

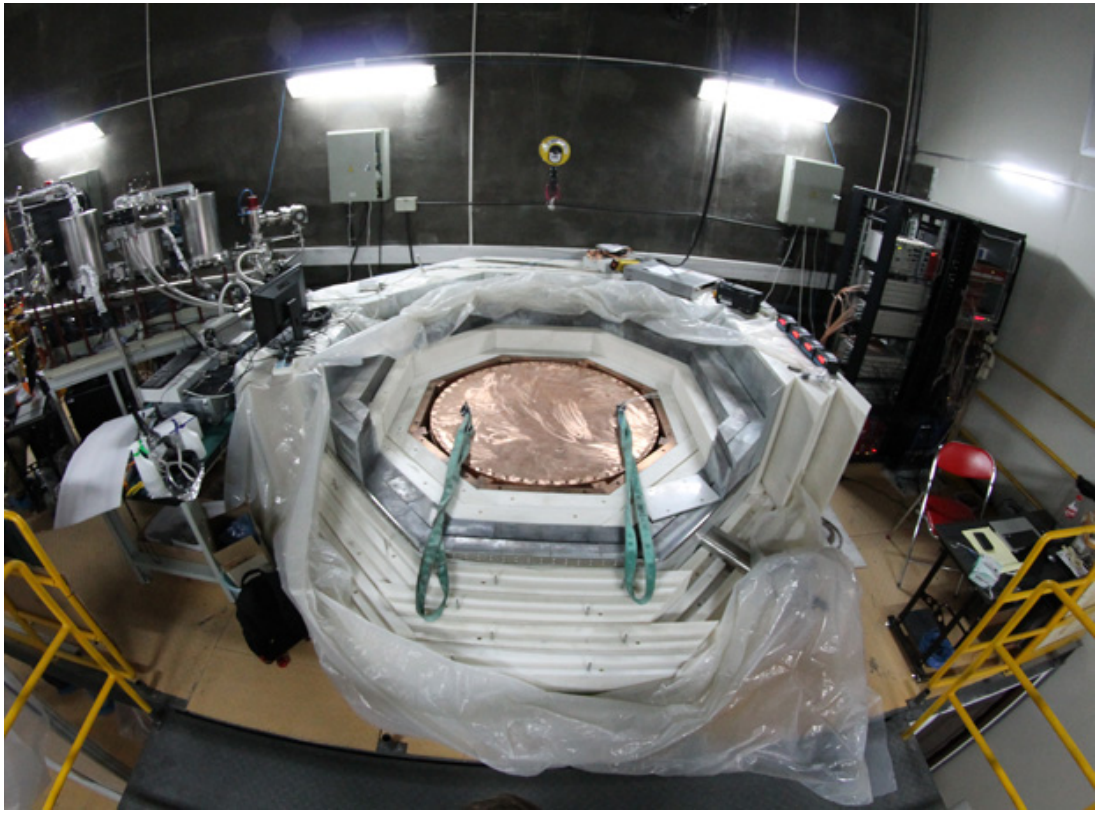
# 暗物质

根据最新天文学和宇宙学的研究，暗物质代表了宇宙中约85%的物质含量。但是，由于与普通物质没有直接电磁相互作用，所以暗物质不发光，类似于“幽灵”，无法用通常的办法看到。

科学家们普遍认为，暗物质粒子与普通物质之间极有可能存在一种微弱的相互作用。但这个作用太过微弱，暗物质粒子可以轻松穿过地球，这使探测起来异常困难。

这种探测实验最困难的地方是探测器中可能存在多种外来干扰，这些干扰也会产生光电信号。**造一座最黑的房子，等待那束最弱的光**。

# PandaX 实验



# PandaX 实验

PandaX实验用氙原子作为探测靶子，采取“守株待兔”的方式，探测弥散在地球周围的成千上万亿的暗物质粒子可能碰撞到氙原子上而发生的微弱信号。碰撞会转化为氙原子的反冲能，在探测器中发光、发电。光和电信号都可以通过灵敏的光电管作为“事件”记录下来。

实验室建在2400米深的山体下（通过岩石屏蔽宇宙背景辐射以及人类活动干扰），用上百吨的高纯材料把探测器层层包围起来（这些对暗物质探测毫无影响），在从2016年3月到6月底近100天的运行中，PandaX探测器记录了约**3千万次**的事件。这些事件可以高效地通过计算机数据处理的办法，进行甄别，一一排除。

最后剩下的可疑事件只有一个，该事件发生在北京时间2016年6月11号3点3分6秒。通过细致分析，这个事件来源于探测材料的放射性，而不是暗物质。

# PandaX 实验

负责实验数据分析的上海交通大学特聘教授刘江来表示，这次数据分析挑战性最高的地方在于利用一个全新的探测器对所有探测到的事件进行**“模式识别”**，用前所未有的精度来甄别暗物质信号和背景“噪声”。

由于国际竞争的紧迫性，在PandaX的分析团队放弃了所有周末休息时间，每天花十四小时以上分析和讨论，在极短时间内完成了数据分析工作。

https://www.physics.sjtu.edu.cn/node/1975

# 第四范式

与之前的宇称不守恒实验不同，寻找暗物质的实验是纯粹的观测，而前者是可以控制条件的。这也恰好对应了社会科学的两种数据，观测数据和实验数据。

从海量的数据中，通过模式识别的方法找到需要的信息，三千万分之一的概率。

在高能物理领域早就出现了理论数量远远多于实验观测的情形，而基本上所有的理论都是唯象的，也就是根据曲线拟合出来的。

图领奖得主吉姆·格雷将这种数据密集型的研究概括为第四范式。在未来，科学研究将会越来越依赖数据驱动。

# 锦屏地下实验室

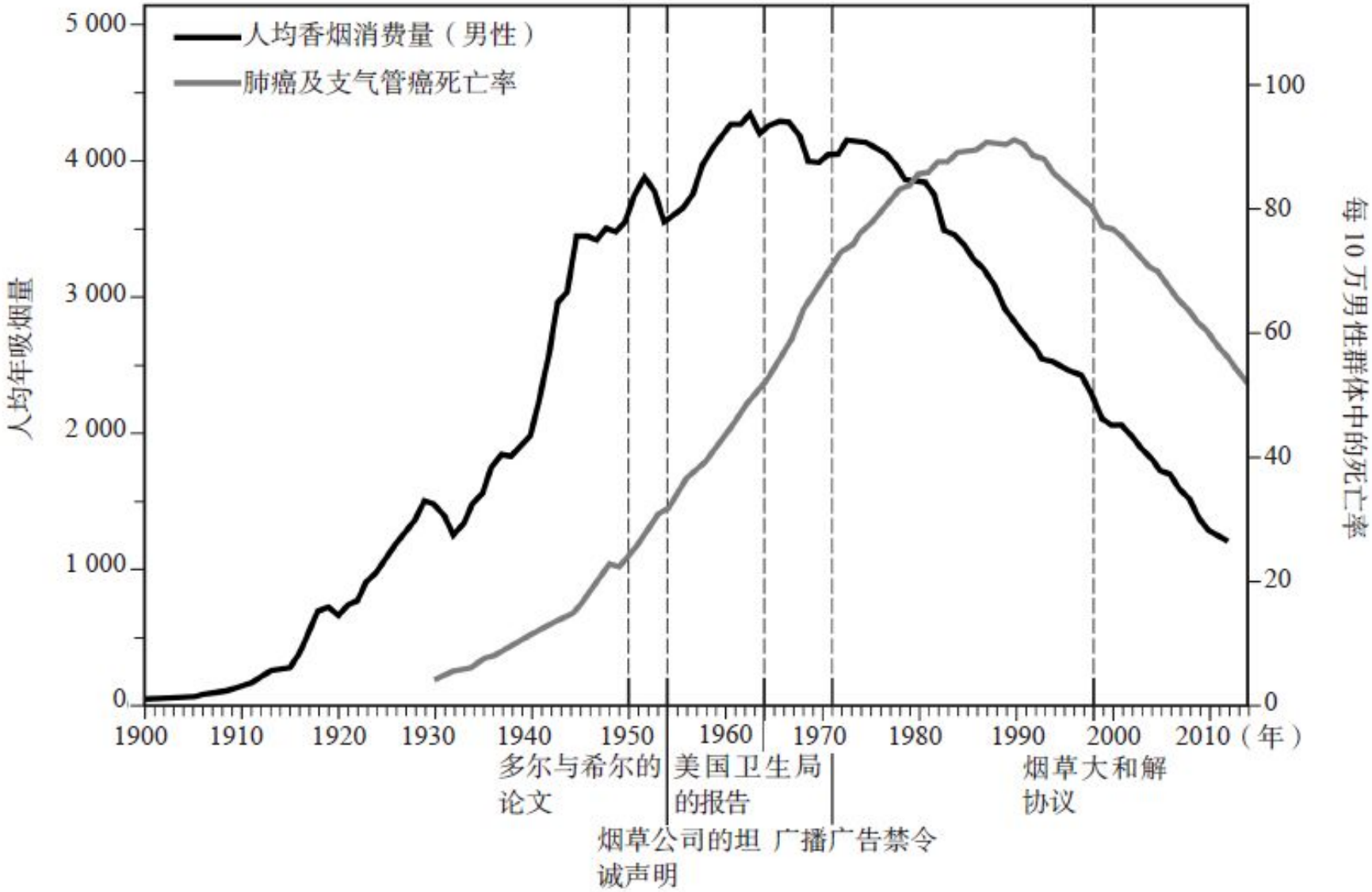
2008年8月8日，清华大学教授岳骞，在当天的奥运会开幕式直播时，看到一条插播的新闻，“锦屏水电站引水隧道顺利实现双洞贯通，隧道最大埋深2375米”，后来听说这条隧道开通之前，是需要通过直升机运送物资的。

岳骞迅速向学校领导反馈这一消息，很快清华大学与雅砻江流域水电开发公司联合建立地下实验室，成为当前世界上最深的地下实验室。

这个故事告诉我们：要看新闻联播，以及插播新闻的时候不要换台

# 烟草与肺癌

如果说物理学的研究在感观和认知上距离大家很远，那医学和公共卫生就和社会科学的研究非常接近了。


人均香烟消费量（男性）
肺癌及支气管癌死亡率
人均年吸烟量
每 10 万男性群体中的死亡率
5 000
4 000
3 000
2 000
1 000
0
100
80
60
40
20
0
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010（年）
多尔与希尔的论文
美国卫生局的报告
烟草公司的坦诚声明
广播广告禁令
烟草大和解协议

# 烟草与肺癌

实际上，在1902年，香烟仅仅占到美国烟草市场的2%，而这一数字到1952年已经扩大到81%。而在烟草出现之前，肺癌本来是一种非常罕见的疾病，而1960年，肺癌已经成为男性群体中最常见的癌症之一。

从上图中，我们可以很显然的将肺癌的爆发归咎为吸烟，但是如果回到1950年代，这可能是一件非常困难的事情。

驳斥吸烟致癌假说的一个最重要的科学主张是可能存在某些不可测量的因素，同时导致了人对尼古丁的渴求和人患肺癌。

相关关系并不等于因果关系。许多人吸了一辈子的烟，却从未患肺癌反，有些人从不吸烟却依然患上了肺癌。他们中的一些人可能是因为家族遗传而得了肺癌，另一些则是因为接触到了致癌物，还有一些人两个方面的原因都有。

1900年到1950年，确实是有多个同时发生的致癌因素：道路铺设，汽车尾气排放，以及普遍的空气污染

# 艰难的科学研究

实际上，当时科学家已经掌握了一项研究因果关系非常有力的武器：随机对照试验。最早的研究中，多尔和希尔（1948）使用对照试验的方法证明了链霉素可以治疗结核病。

随机对照试验，简单来说就是随机将研究对象随机分为两组，一组接受某种干预，而另一组只接受“安慰剂”，根据两组对象的干预结果来判断干预的因果效应。

但是这样的思路，这个方法在研究吸烟对肺癌的影响时显然是不适合的。

退一步，1950年，希尔为肺癌病人选择了一组健康的志愿者作为对照组，调查其行为与病史。研究结果令人震惊。在649名接受采访的肺癌患者中，除两人外其余均为吸烟者。这一结果在统计学上与随机水平相去甚远，是一件极不可能发生之事。多尔和希尔情不自禁地计算出了该结果的精确度：1 500 000:1。此外，平均而言，肺癌患者的吸烟量也比对照组成员的吸烟量更大，但采访也显示，肺癌患者吸入烟雾的比例相对较小

# 艰难的科学研究

之后，该研究在不同的被国家重复了19次，但是这一研究结果依然存在很大偏倚。因为这是一个回溯性的研究，所以患者回忆会存在偏倚，可能患者更容易回忆起自己的行为；另一方面，已住院的癌症患者是否可以代表人口中的吸烟群体。何况**肺癌患者吸入烟雾的比例相对较小**。统计学家费舍尔非常激烈的批评希尔的研究，“一项偏倚的研究重复19次也不能证明任何问题”

希尔从1951年开始向6万名英国医生通过问卷调查追踪其吸烟习惯和健康状况，结果5年后，重度吸烟者死于肺癌的概率是不吸烟者的24倍。而美国的证据则更严重。

# 体质假说

前瞻性研究仍未能将吸烟者与其他各方面都相同的不吸烟者进行比较。事实上，这种比较是否可行值得怀疑。毕竟，吸烟是吸烟者的一种自我选择。在许多方面，他们都可能与不吸烟者有着基因或“体质”上的不同，比如有更多的冒险行为，更易饮酒过量等。其中一些行为同样可能会对健康造成不良影响，而这些不良影响可能被错误地归咎于吸烟。对于怀疑论者来说，这是一个特别便利的论据，因为体质假说几乎不可检验。直到2000 年人类基因组测序工程开启之后，寻找与肺癌相关的基因才成为可能。

康菲尔德把目标直接对准了费舍尔的体质假说。如果吸烟者患肺癌的风险为常人的9 倍，那么在吸烟者中，这种混杂因子存在的概率也需要至少比常人高出9 倍，如此才能解释这种患病风险的差异。让我们思考一下这意味着什么：如果有11% 的不吸烟者携带“吸烟基因”，那么就至少有99%的吸烟者一定携带吸烟基因。而如果有12%的不吸烟者碰巧携带这种基因，那么从数学的角度看，“吸烟基因”就不可能完全解释吸烟和癌症之间的相关。

后来解密的文件显示烟草公司内部早已知晓其危害，却在蓄意欺骗公众。

《为什么:关于因果关系的新科学》P147-152

# 医学与公共卫生研究的启示

- 人体作为一个比原子、分子复杂的多的系统，已经无法通过数学模型来进行建模，而数学模型内在蕴含了因果关系
- 理论简化成为简单直接的因果关系，这种简化可能是由于研究对象过于复杂造成的
- 医学和公共卫生的研究中，因果推断需要依赖统计方法，而且其要求是及其苛刻的，是一种纯粹唯象的方法。
- 随机对照试验是一个强有力的武器，但在很多情形下无法使用
- 可复现性是科学进步的重要基础

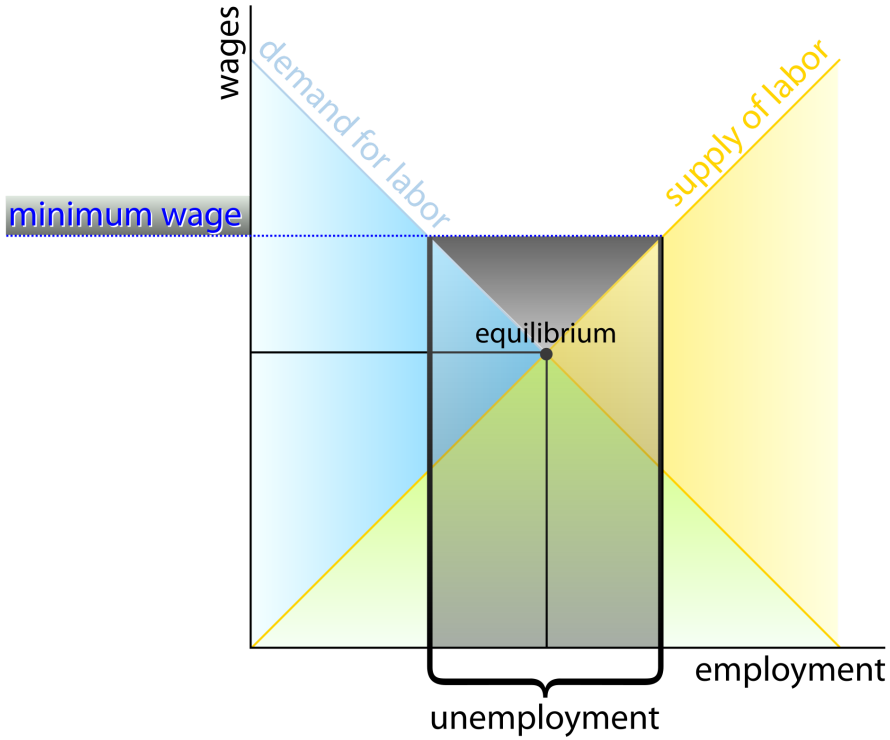
# 最低工资

# 民主党总统候选人

7/8支持15美元最低工资方案，1/8未确定态度

# 最低工资

新古典模型


wages
demand for labor
supply of labor
minimum wage
equilibrium
employment
unemployment

# 最低工资

新古典的模型认为，当最低工资高于市场均衡工资的时候，会减少就业岗位，从而提升失业率。这项旨在改善低收入人群福利的政策，反而会造成他们失业！

2015年之前，美国西雅图市的最低工资常年维持在9.47美元，2015年，市长将这一数字提升到11美元，2016年再提升到13美元，2017年更是提升到15美元。

2017年的一篇工作NBER论文中，UW的研究团队使用华盛顿州政府的个人工资数据证实，2016年提升最低工资导致西雅图的低收入工作群体失业率上升5%， 而平均工作时间减少10%。[1]

然后就在这篇论文上线的前一周，伯克利的经济学研究团队发布了一个完全一样话题的论文，但是不同的是，他们没有发现最低工资改革对失业率的影响。他们的研究没有使用行政数据，而使用的统计数据。[2]

[1]Jardim E, Long M C, Plotnick R, et al. Minimum wage increases, wages, and low-wage employment: Evidence from Seattle[R]. National Bureau of Economic Research, 2017.
[2]Reich M, Allegretto S, Godoey A. Seattle's Minimum Wage Experience 2015-16[J]. Available at SSRN 3043388, 2017.

# 最低工资的公案

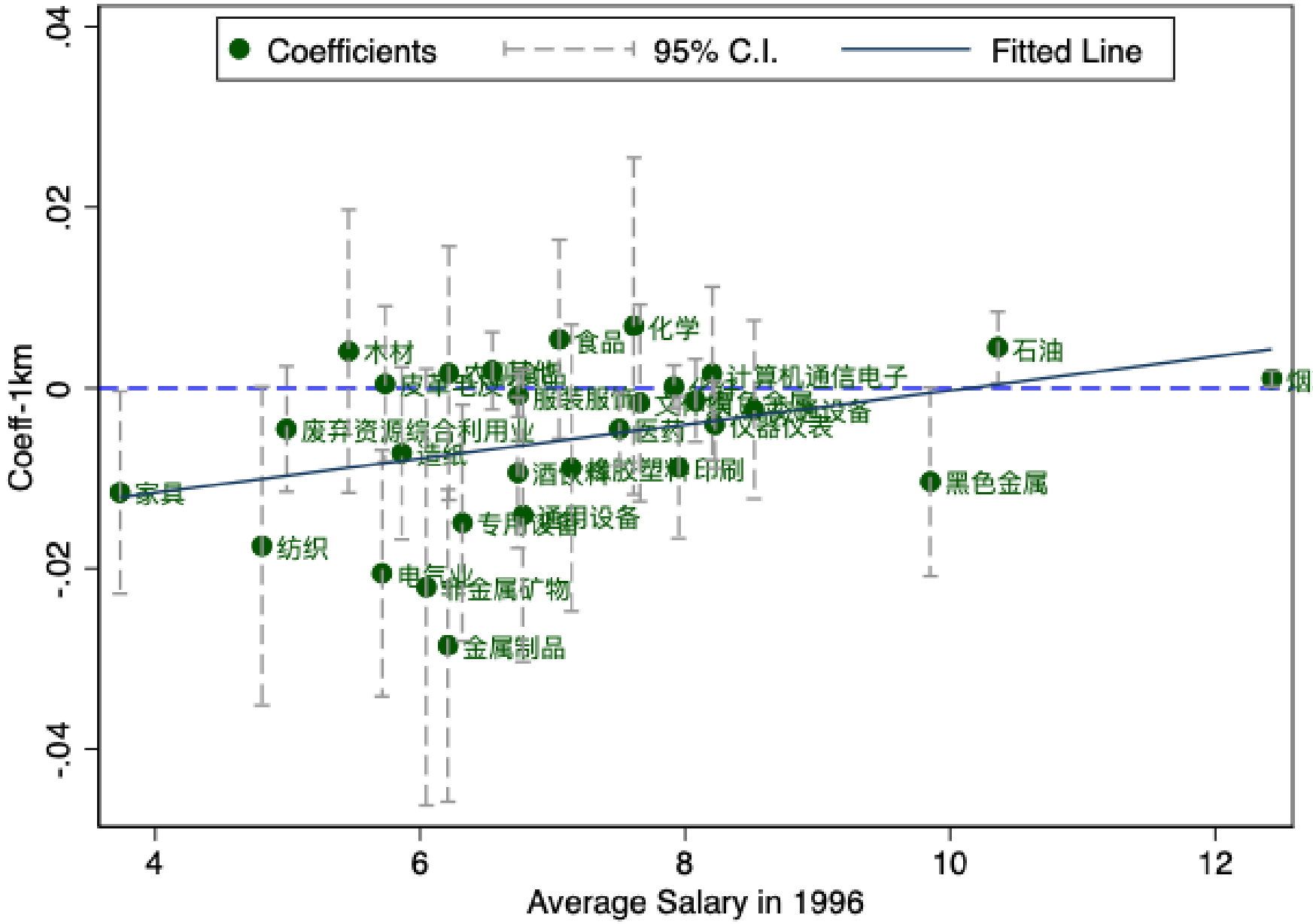
更有料的是，西雅图周刊报道，UW团队曾经在发表论文之前，讲论文的结果分享给西雅图市政府。西雅图市政府在看到研究之第一时间联系了伯克利的团队。

实际上，最低工资法案改革是市长竞选的主要承诺之一

从这个公案里面，我们发现，社会科学的实证研究是非常“弹性”的。不同的数据、不同的样本、不同的数据处理方式可能导致不同的结论。

**而不同的结论会导致不同的政策结论！**

可复现性阻碍社会科学的理论积累与科学化，当然医学也不是很乐观，美国斯坦福大学预防医学研究中心主任 Ioannidis，在 2005年发了一篇论文《Why most published research findings are false》。Ioannidis 对 1990 年至 2003 年间发表的 49篇最顶级医学论文进行了研究，最终发现只有20篇是靠谱的。Ioannidis 还不过瘾，又在不算最顶级的论文里找了432个医学研究成果，其中重复验证有效的，只有 1 个。


Coefficients
95% C.I.
Fitted Line
Coeff-1km
.04
.02
0
-.02
-.04
4
6
8
10
12
Average Salary in 1996
木材
化学
食品
石油
计算机通信电子
皮革毛皮
服装服饰
废弃资源综合利用业
医药
仪器仪表
造纸
酒饮料
橡胶塑料
印刷
黑色金属
家具
专用设备
通用设备
纺织
电气业
非金属矿物
金属制品

# 高速公路的最高时速

生命的统计价值（value of statistical）：

$$w_i = \alpha \rho_i + other\ variables$$

劳动经济学中，经常使用上面的公式来估计生命的统计价值，其中α表示一个工作的风险从0变为1的时候，需要多付给工人的工资。

这个数字在公共政策中具有重要的意义，例如在高速公路建设过程中，最高时速越高，通行时间越短，从而具有经济价值，但是必然伴随着更高的风险，更多的交通事故。

2014年开始，加州交通局决定使用3百万美元作为政策决策指导。

# 科学研究的特征

1. 科学理论可以是一套基于数学的模型，也可以是一条因果关系的命题。实际上命题往往是理论模型的推论，作为证伪的假设而存在

2. 理论推出假设，并指导实验

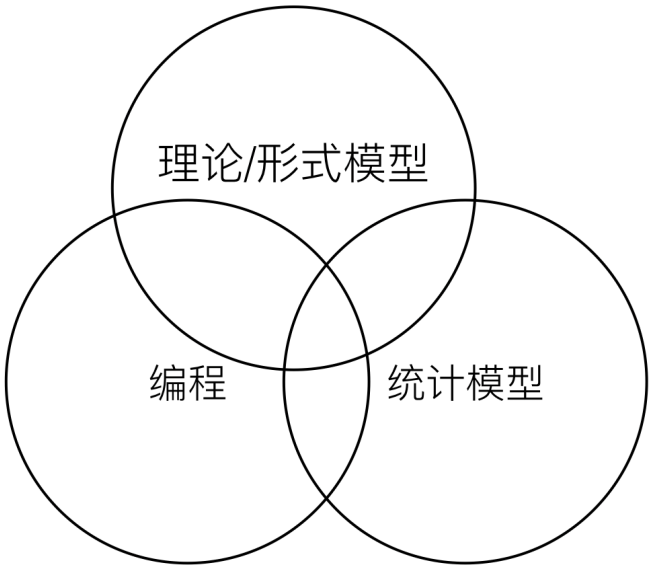
3. 以数据驱动的第四范式已经出现

4. 科学的进展是建立不断复现，争论的基础上的。只有经得起复现的研究才是有价值的科学研究。

5. 社会科学结论的弹性很强，不同的数据往往会导致不同的结论，因此自身需要更高的复现性要求。

6. 科学的方法指导下的经济研究

# 对社会科学的再认知


理论/形式模型
编程
统计模型

# （社会）科学研究三要素

1. 理论模型: 决定了选题与研究的方向，可以指导。
2. 数据（编程、实验）: 验证理论假设，或发现新的现象。决定研究的下限。
3. 统计模型: 从数据中检验假设的工具，与理论一起决定了研究的上限，在一定程度上与数据是替代关系。

---

# 传统社会科学：理论+统计模型

- 框架与模型（假设）的关系
- 形式模型：基于数学逻辑演绎的模型
- 统计模型：基于数据归纳的模型

# 社会科学的内部纷争

- 理论数学化程度的差异
- 工具可复现性的差异
- 关心研究对象的差别
- 理论与关系的演变 物理学都已经走向数据驱动和唯象理论了，生命科学也在很大程度上依然处于唯象的阶段，社会科学却在教条的走上了“理论-假设-实证”的道路，存在极大的误导性
- 定性研究和定量研究的争论是个伪命题。金山找对叶问说，今日北方拳输给南方拳了。叶问回答：你错了，不是南北拳的问题，是你的问题！

# 国内的传统社会科学：白话版理论

- 理论基本靠思辨
- 概念基本靠字典
- 实证基本靠描述
- 知网的论文尽可能少看，最多看看文献综述即可。当然以后的质量预期会更高

# 编程：计算机科学与工程

- 追求算法简洁
- 追求运算效率
- 解决实际工程问题

# 编程+理论：管理科学与工程

- 以工程的视角解决管理学问题
- 以模拟的方式拟补实证数据的缺失

# 统计模型：传统科学

- 以概率论为基础对现实问题进行建模
- 用已有数据拟合模型参数
- 用模型进行预测

# 编程+统计模型：AI

- 大规模数据的存储与运算技术
- 快速的运算能力
- 黑箱式的模型系统（深度学习）

# 编程+理论+统计模型：下一代社会科学

- 具备中级段位的编程能力：在收集、预处理数据方面的竞争力。尤其注重数据的透明性与可复现性
- 具备过硬的统计模型基础：保证实证研究的可靠性
- 具备扎实的理论基础（形式模型能力）：思辨与研究设计的能力
- 不与程序员竞争
- 要具备与统计学家对话的能力，例如，看得懂Econometrics
- 强调理论导向，但是已经是实证时代
- 解释数据发现
- 引导模型建立
- 因果推断
- 设置议程

# AI 对社会科学研究的冲击与挑战

# AI智能发展的三要素

人工智能目前实现的就是个分类器功能，最简单的分类器就是线性回归。人工智能的大发展的得益于其算料、算力以及算法的突破。

1. 算料，即数据。传感器技术的发展，互联网的发展使得数据的采集越来越及时，积累越来越多，数据储存和运输的成本越来越低。大数据是一个商业概念，而不是一个学术概念，因此大数据很难界定，也存在太多误用。但是，无论如何大数据必须要大，一台笔记本电脑可以打开的一定不是大数据。

2. 算力，计算机集群。真正的突破在于分布式运算的突破，一百万个臭皮匠，肯定打死一个诸葛亮的水平。欧洲核子中心 (CERN) 一年用电约 1.3 TWh;北京东城区西城区2016 年用电总量10.1TWh，人口200万;史老师的实验室不需要暖气。

3. 算法，深度学习、神经网络。机器学习算法，以Alpha GO为代表。神经网络是一个黑箱，但是特别管用。

# AI 对社会科学研究的冲击

1. 最直接的冲击是人的冲击
2. 工具更新带来的恐慌
3. 对统计模型的冲击

# AI 带给社会科学的机遇

从研究上带来的机遇

1. 数据量上,从抽样到全体,例如https://www.opportunityatlas.org/
2. 测度那些不可测的变量，例如，创业质量（Dongbo Shi et al 2020）
3. 估计个体因果效应的差异（史冬波和罗亦文，2019）
4. 人工智能作为一项共性技术可以提升科研的效率

作为研究对象带来的机遇

1. 对劳动力市场的影响
2. 对经济增长的影响
3. 对社会福利（不平等）的影响
4. 对产业组织的影响

# AI其实也没有解决太多问题

1. 大部分打着大数据旗号的社会科学研究都是骗人的，数据量都不够大
2. 数据量大不解决抽样偏误问题
3. 数据量大不解决选择性偏误问题
4. AI不解决因果问题

# 数据工程

# 数据工程

数据工程顾名思义就是将数据科学的工作流程化，用软件工程的思路优化数据科学流程，具体实现中，遵循以下四个原则。

数据工程四原则

1. 可复现，以人类语言和计算机语言的形式，详细记录每一步计算。这是科学的基本精神，与可证伪性一起，是区分科学与伪科学的标志。

2. 自动化，Single Point of Truth, Don't Repeat Yourself. 不可在分析做任何重复，任何有意义的信息都应该被共享，

3. 正交分工，将数据分析任务切分为相互不影响的组成部分，分工是现代社会的基础。

4. 最佳工具，尽量使用高级语言和语法糖，为每个子任务选择合适的工具。只有在性能分析之后，才在必要时使用低级语 言进行性能加速，最佳工具会随时间变化。

# R语言

1. R语言始终是不错编程语言之一
2. R语言是学习数据科学的最佳选择，没有之一
3. R语言可以独立完成整套数据工程流程
4. R就像蝙蝠侠：侦探工作、智慧、狡黠、使用工具、动脑多于蛮力
5. Python就像超人：肌肉力量、超级力量、优雅、全面、蛮力多于用脑

社会科学家是可以也是应当和科学家对等对话的，愿原力与你同在！