



# Criminal Justice Theory, Empirical Study, Machine Learning and Judge Decision-Making

报告人：陈铭阳

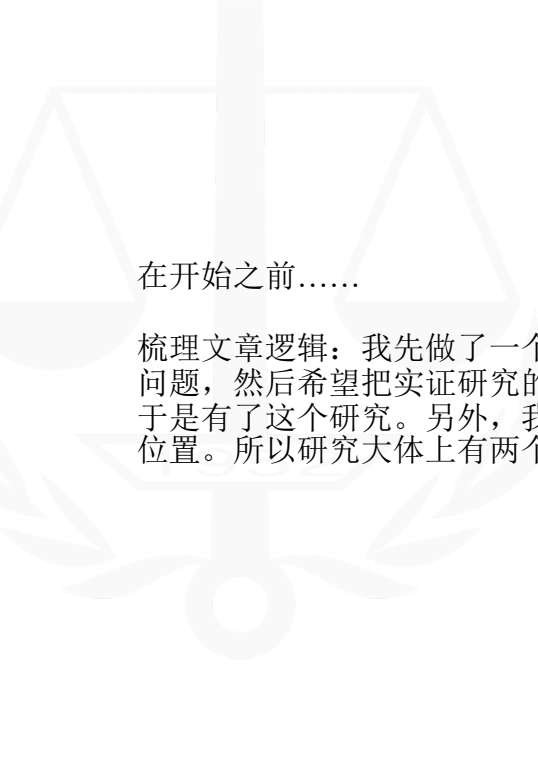
第七届迈向数据法学研讨会

2024年11月30日，浙江杭州



本研究将回答以下三个问题：

1. 系统性回答 “中国刑事司法判决是否实现了同案同判？”
2. 把同案同判的实证研究用于数据的筛选是否能提高机器学习模型的表现？
3. 经过筛选后的模型能否帮助纠正法官的量刑偏误？



在开始之前.....

梳理文章逻辑：我先做了一个实证研究系统性回答了中国判决存在同案同判问题，然后希望把实证研究的方法结合到机器学习模型当中解决这个问题，于是有了这个研究。另外，我们也希望能找到法学的从业者在建模任务中的位置。所以研究大体上有两个部分，一个是实证问题，一个是模型问题。

## 研究背景1：刑事司法的同案同判

梳理完整理论框架：案情→说理→刑期

The Cinderella Complex:

Word Embeddings Reveal Gender Stereotypes in Movies and Books

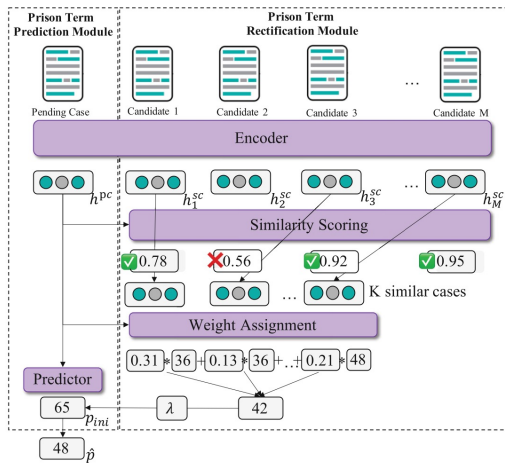
Huimin Xu<sup>1</sup>, Zhang Zhang<sup>2</sup>, Lingfei Wu<sup>3\*</sup>, Cheng-Jun Wang<sup>1\*</sup>



1. 灰姑娘情结问题：法学目前的研究当然可以如同文学研究一样，通过阅读一个一个的案件来回答“是否同案同判”的问题，但是这种方法无法给出系统性的答案。
2. 先前的研究往往忽略了案情到说理的推理是否一致，只关注于刑期和案情之间的关系。

## 研究背景2: Garbage in, Garbage out

图注: Similar Case Based Prison Term Prediction工作原理



如果Candidate样本是不同案同判的案件（说明出现了量刑偏倚，属于低质量案件），那么就会导致garbage in, garbage out的问题。

如果样本用前面的实证研究方法筛选然后进行机器学习模型的训练，是否可能进一步提高模型的表现？

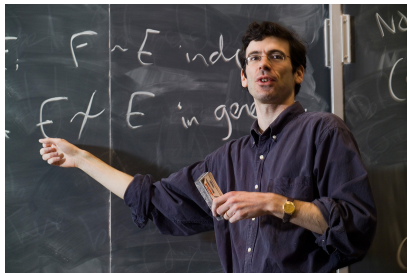
### 研究背景3：把法外因素变成“噪音”

NBER WORKING PAPER SERIES

HUMAN DECISIONS AND MACHINE PREDICTIONS

Jon Kleinberg  
Himabindu Lakkaraju  
Jure Leskovec  
Jens Ludwig  
Sendhil Mullainathan

Working Paper 23180  
<http://www.nber.org/papers/w23180>

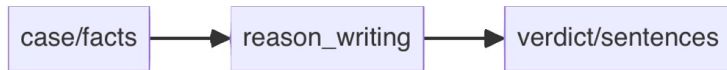


1. 机器学习模型纠正偏误的潜力：Kleinberg关于假释的预测模型研究中说明了，因为机器学习模型本身只考虑给定的特征值，而导致法官决策偏误的法外因素（性别、种族）则在大样本中会变成噪音，最终会被模型消解。
2. 如果按照研究背景2的想法引入法学理论，提高训练样本质量，减少噪音（降噪），那么完全可能制作出能够纠正法官决策偏误的模型。

研究方法1：数据处理——5000+抢劫罪裁判文书

B	C	D	E	F	G	H	I	J	K
序号	案情	说理	刑期	自首	立功	坦白	累犯	前科	认罪认罚
5	许，杨某、林某1、王某、潘某、林某2、李某及被告人唐飞航共七人携带水果刀、棒球棍等工具，驾驶一辆白色桑塔纳轿车在瑞安市逼停被害人陆某、覃某驾驶的皖E×××××油罐车，控制两名被害人后，杨某将劫取的油罐车开至滨海××附近，后将油罐车内装载的42吨柴油销赃。	被告人唐飞航以非法占有为目的，结伙以暴力、胁迫的方法劫取他人财物，数额巨大，其行为已触犯刑律，构成抢劫罪。公诉机关指控的罪名成立。被告人唐飞航在共同犯罪中起次要作用，系从犯，依法减轻处罚；归案后能如实供述自己的罪行，可依法从轻处罚；自愿认罪认罚，依法可依法	48	0	0	1	0	0	1

## 研究方法2：实证部分——文本相似度



案情文本相似度度量

说理文本相似度度量

刑期的相似度度量

RWMD距离、余弦距离等

欧几里得距离

OLS回归

建模时会再用独  
热编码进行一次  
稳健性检验



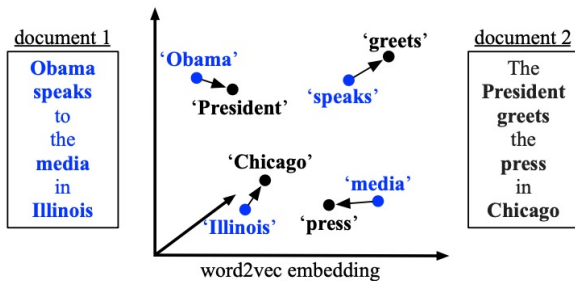
## 研究方法2：实证部分——RWMD介绍

**RWMD：**找文本和文本之间的距离有多远。  
其他相似度计算方法基本原理一致。

裁判文书的特点：高度格式化，措辞固定，相当于法官已经清洗好了（独热编码也是一样的），很适合直接进行相似度分析。

为什么文本分析？独热编码不好吗？——法官可能形成集体经验，比如我们阅读文书文本发现“持刀”这个词成为了一种集体经验。文本分析在切割文本时能够捕捉。我们也利用捕捉的信息反过来编辑规则库

图注：From Word Embeddings To Document Distances 原理



### 研究方法3：建模部分——机器学习

**R语言筛选出不同案同判的案件：**

循环所有案件作为基准，输出不能产生显著负相关的案件剔除



**Python进行：**

线性回归、Lasso、岭回归、神经网络建模  
对比剔除和不剔除的训练结果



**搭建集成学习模型回代到不同案同判的案件：**

R语言随机抽样5个案件，查看纠正的结果如何

## 研究结果1：中国实现同案同判了吗？

**Table 2: OLS of RWMD**

	No. 8	No. 1	No. 3923	No. 5071	No. 4806
C-R Cof.	0.15203	0.225497	0.116908	0.052300	0.16826
p-value	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
R-P Cof.	-5.0235	-3.3419	-1.0013	0.1867	-1.948
p-value	< 0.01	< 0.01	< 0.01	0.3982	< 0.01

**Table 3: OLS of different Algorithm**

	RWMD	Cosine	Correlation
C-R Cof.	0.225497	0.249576	0.288393
p-value	< 0.01	< 0.01	< 0.01
R-P Cof.	-3.3419	-0.98190	-0.98056
p-value	< 0.01	< 0.01	< 0.01

**Table 4: OLS of Cosine Similarity**

	No. 1017	No. 4775	No. 2177	No. 5026	No. 1533
C-R Cof.	0.081784	0.045205	0.323915	0.212049	0.226299
p-value	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
R-P Cof.	-1.0440	-0.1145	-0.6881	-0.5782	-0.6151
p-value	< 0.01	0.588	< 0.01	< 0.01	< 0.01

# 研究结果1：中国实现同案同判了吗？working paper的稳健性检验

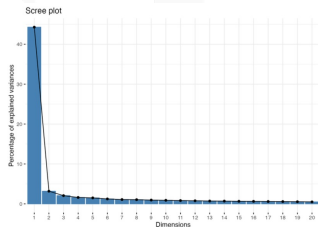


Figure A7.1: reason PCA

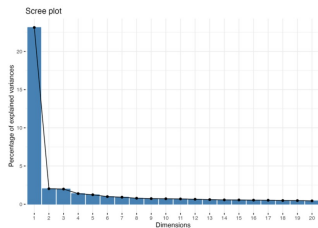


Figure A7.2: case PCA

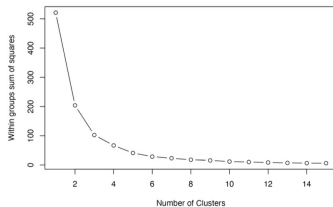


Figure A6.1: elbow method

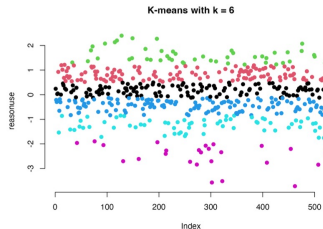


Figure A6.2: clusters

## K-means and Logistic regression

Table A4.4.1: k-means clusters

Cluster No.	Value (standardized)
1	1.6081276
2	0.8280640
3	0.1999997
4	-0.4044666
5	-1.2030767
6	-2.4947546

Table A4.4.2: logistic regression

	Model 1	Model 2
Cluster 5	7.208*	-0.161
Cluster 4	11.063**	-0.175
Cluster 3	12.027***	-0.241
Cluster 2	14.157***	-0.284
Cluster 1	20.424***	-0.296
Pr (>Chi-sq)	0.0000***	0.7196
Num.Obs.	522	518

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

## 研究结果1：中国实现同案同判了吗？

图注：独热编码处理的数据输出的刑期不符合同案同判的案件序号



24  
35  
50  
53  
60  
72  
73  
81  
82  
100  
137  
138  
139  
147  
150  
179  
180  
188  
190  
211  
221  
223  
226  
235  
248  
251  
264  
302  
314  
315

图注：什么案件更容易被误判？

```
1 numHigh <- 0
2 for (i in unreasonable) {
3   if (caseset$刑期[i] < 60) {
4     numHigh <- numHigh + 1
5   }
6 } # 循环计算低刑期案件数量
7 print(numHigh)
8 # 输出结果: 117
9 print(numHigh/length(unreasonable))
10 # 输出结果: 0.1808346
11 numHigh2 <- 0
12 for (i in 1:length(caseset$刑期)) {
13   if (caseset$刑期[i] < 60) {
14     numHigh2 <- numHigh2 + 1
15   }
16 } # 循环计算低刑期案件数量
17 print(numHigh2)
18 # 输出结果: 3704
19 print(numHigh2/length(caseset$刑期))
20 # 输出结果: 0.7289904
```

## 研究结果2：机器学习模型表现提高了吗？

**Table 5: Comparison with Baseline Model**

Model name	Baseline R2	Baseline MSE	Aft-Filter R2	Aft-Filter MSE
Linear Reg	0.548/0.542	674.6032	0.680/0.784	245.6639
Ridge Reg	0.553/0.532	721.6177	0.701/0.687	257.6409
Lasso Reg	0.531/0.542	745.7592	0.682/0.682	264.0610
MLP Reg	0.764/0.504	747.2403	0.870/0.701	252.2937

Tasks	Law Articles		Charges		Prison Terms
Evaluation Metrics	$F_{\text{micro}}$	$F_{\text{macro}}$	$F_{\text{micro}}$	$F_{\text{macro}}$	Score
nevermore	<b>0.958</b>	<b>0.781</b>	<b>0.962</b>	<b>0.836</b>	77.57
jiachx	0.952	0.748	0.958	0.815	69.64
xlzhang	0.952	0.760	0.958	0.811	69.64
HFL	0.953	0.769	0.958	0.811	<b>77.70</b>
大师兄	0.945	0.757	0.951	0.816	73.16
安徽高院类案指引研发团队	0.946	0.756	0.950	0.803	72.24
AI judge	0.952	0.766	0.956	0.811	—
只看看不说话	0.948	0.738	0.954	0.801	77.54
DG	0.945	0.717	0.949	0.755	76.18
SXU AILAW	0.940	0.728	0.950	0.791	76.49
中电28所联合部落	0.934	0.740	0.937	0.772	75.77

Table 1: Performance of participants on CAIL2018 .

测试集得分：86.336

训练集得分：89.707

测试集得分：84.825

训练集得分：90.790

## 研究结果3：机器学习模型能纠正偏误吗？

图注：测试网站



## 研究结果3：机器学习模型能纠正偏误吗？

搭建集成学习模型，然后回代测试：

**Table 6: TLCA-test result**

	No. 1224	No. 4166	No. 3770	No. 1173	No. 2788
Judges' decision	180	37	36	171	180
Staking model	96	59	92	84	104
Pass TLCA-test? (Y/N)	Y	Y	N	Y	Y

测试模型的网页：<http://118.25.58.138>

图注：机器学习纠正结果符合法学生的分析吗？

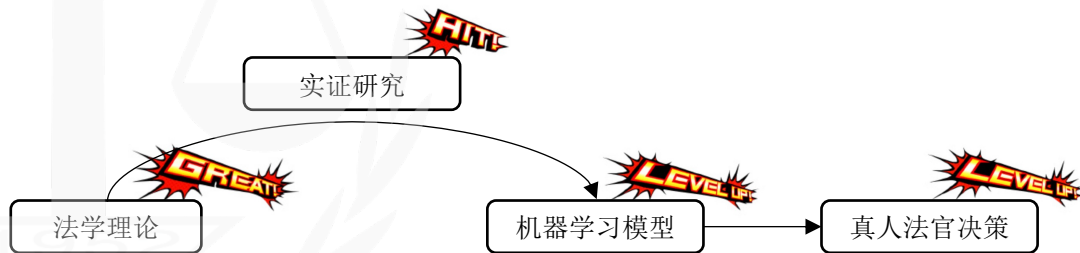
在 1224 号案件中，被告人仅仅是持刀入户抢劫，并且是初犯，也没有造成人员的伤亡，数额也并不巨大，只是存在法定的结果加重犯情节，罪不至无期徒刑，但是法官却为其期判处了无期徒刑，量刑明显偏高。在纠正后，量刑调整为 96 个月。4166 号案件中，被告人结伙持刀殴打被害人，法官为其量刑 37 个月，明显偏低，在其他没有持刀也没有进行殴打的案件中，被告是 36 个月的有期徒刑（比如不属于不符合合同案同判的 4162 号案件），更何况本案中对我们提取的是主犯，法官在说理中指出需要加重刑罚，却判处和一般抢劫相仿的刑期，不合理，纠正后为 59 个月的刑期。3770 号案件案情为入户盗窃后转化型抢劫，并且持刀，同时法官存在要件认定的一个小错误，即该案件根据司法解释应为既遂，而不是未遂。因此应当判处更高的刑期，而不是短短的 36 个月。这个案件没有显著负相关很可能是说理本身存在一些瑕疵，但是他能够出现负相关。1173 号案件为持刀抢劫，数额巨大，同时及时缴纳了罚金，除此之外没有其他法定的量刑情节，然而量刑居然达到了 171 个月，明显偏高，纠正后为 84 个月。2788 号案件的量刑为 180 个月，我们检验发现是提取逻辑让这个案件存在一定的误差，实际提取为 48 个月。此案存在入户抢劫，被认定为未遂，但是实际上本案已经既遂。我们修正为既遂，也是 105 个月的刑期。因此无论是 180 个月还是 48 个月，都不符合该案例刑期。



## 研究结论与贡献

1. 在中国，以抢劫罪为例，司法判决的案情和说理基本实现了同案同判的目标。但是说理和刑期却无法稳定出现显著负相关。
2. 在不同案同判的案件中，比例较高的案件为高刑期案件，这一定程度上反映出，政策应当关注于对严重犯罪的自由裁量之限制上。
3. 通过法学理论驱动的实证研究方法筛选模型的训练样本，可以提高数据集质量，从而提高模型的表现。
4. 通过法学理论、实证研究和机器学习模型的结合，本研究证明了机器学习模型具有通过化法外因素为噪音从而纠正法官量刑偏误的潜力，探索了四者的关系。

## 研究结论与贡献





# 感谢聆听！

团队成员（按贡献排序）：

陈铭阳，中国政法大学本科（研究设计，规则库构建，文本分析，集成学习器建模，论文撰写）

宋高捷，中国政法大学数据法治研究院硕士研究生（基学习器建模，文本预处理与切割）

徐展学，浙江大学计算机系硕士研究生（基学习器建模，特征提取，网页搭建）

吴志鹏，重庆大学卓越工程师学院硕士研究生（特征提取）

王骛凯，中国政法大学本科（规则库构建，网页模型测试与测试结果报告撰写）

任子安，中国政法大学数据法治研究院硕士研究生（规则库构建，网页模型测试与测试结果报告撰写）