第1届全国高校大数据教学研讨会（BDTS2017）大会特邀报告

2017年5月12日至13日，第1届全国高校大数据教学研讨会（BDTS2017）在厦门大学科艺中心音乐厅隆重举行。本届研讨会由教育部高等学校计算机类专业教育指导委员会主办，厦门大学、厦门理工学院、贵州师范大学、人民邮电出版社联合承办，旨在搭建专业的大数据教学交流平台，汇聚全国高校大数据教学精英力量，共同探讨大数据专业和课程体系建设，为加快推进全国高校大数据教学发展贡献力量。来自全国300多所院校的400余名教师参加了本次研讨会。

厦门大学谭绍滨校长助理、人民邮电出版社教育中心营销部肖稳副主任，北京大学、中国科学院、厦门大学、华东师范大学、同济大学等重点院校的6位大数据教学知名专家，以及来自国内知名大数据企业的3名业界专家出席会议并做特邀大会报告。厦门大学林子雨助理教授主持会议。

更多内容请访问大会官网：http://dblab.xmu.edu.cn/post/bigdata2017/

中科院 朱廷劭 研究员 在做大会特邀报告

行为大数据挖掘及其在心理学的应用

朱廷劭

中国科学院心理研究所计算网络心理实验室

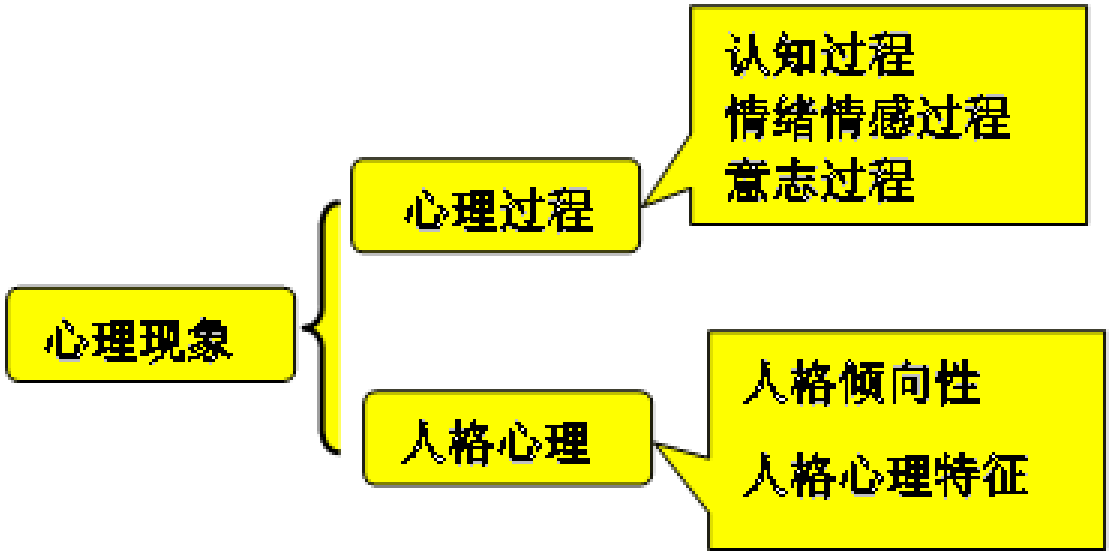
Computational CyberPsychology Lab (CCPL),

Institute of Psychology, Chinese Academy of Sciences

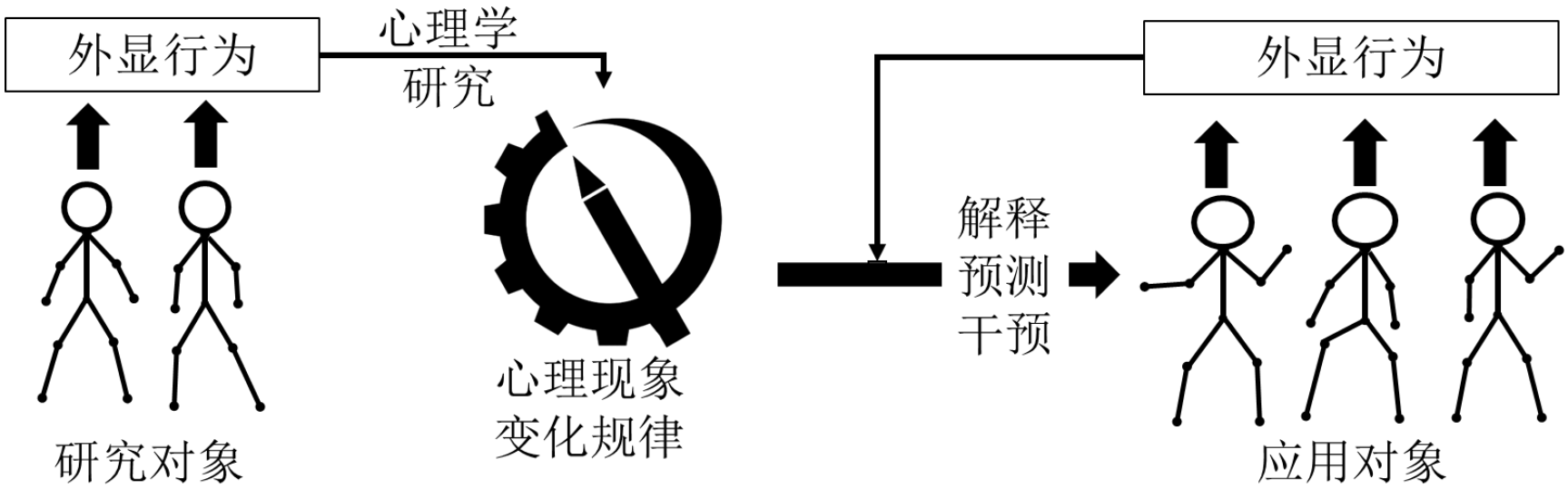
心理学的定义

- **心理学是研究心理现象发生、发展和活动规律的科学。**

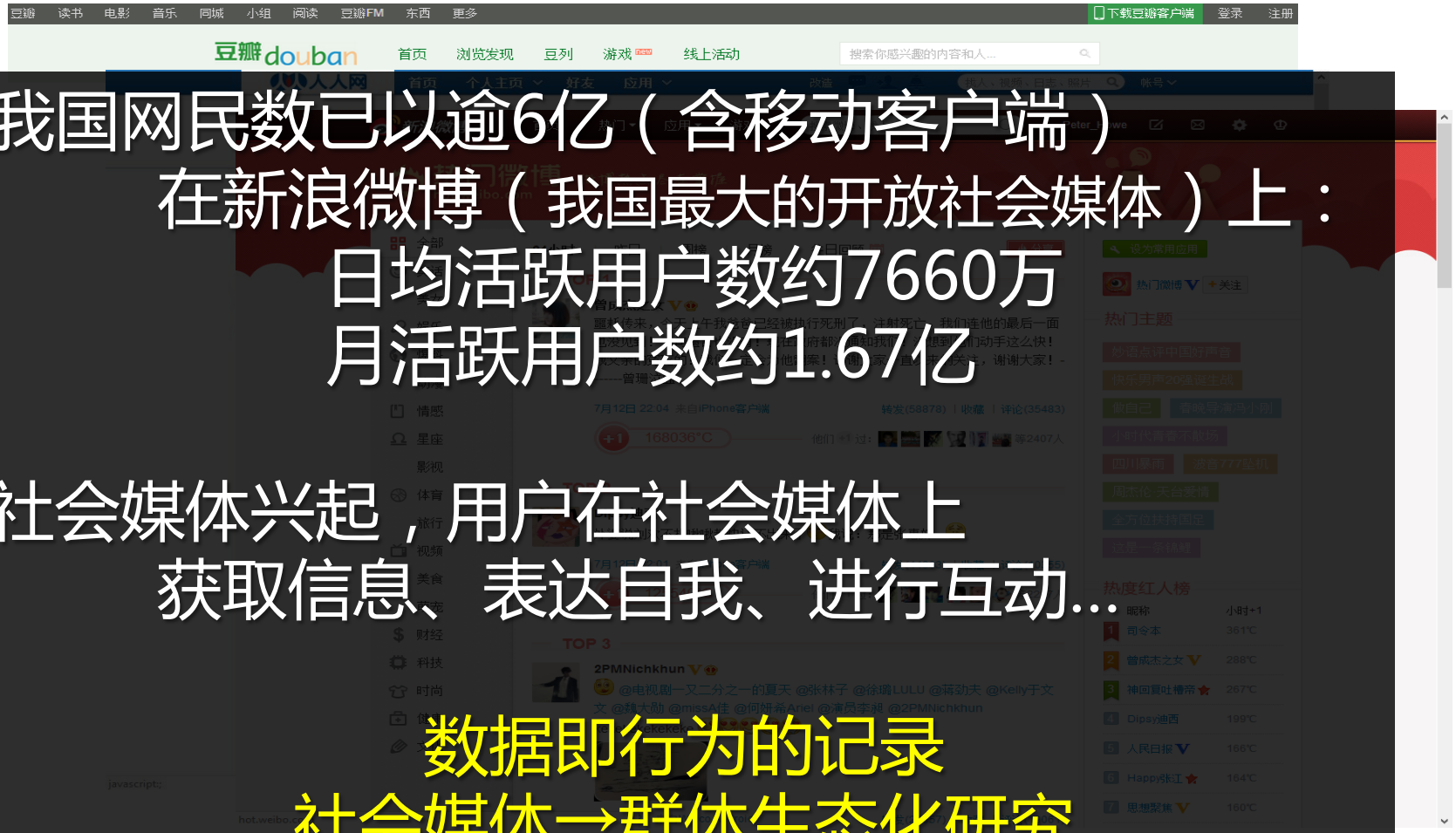
- 人的心理现象，就是指心理活动经常表现出来的各种形式、形态或状态。



心理与行为关系



网络为开展心理生态化研究带来新机遇



我国网民数已以逾6亿（含移动客户端）

在新浪微博（我国最大的开放社会媒体）上：

日均活跃用户数约7660万

月活跃用户数约1.67亿

社会媒体兴起，用户在社会媒体上

获取信息、表达自我、进行互动…

数据即行为的记录

社会媒体→群体生态化研究

大数据进行分析的优势

长期，纵向跟踪

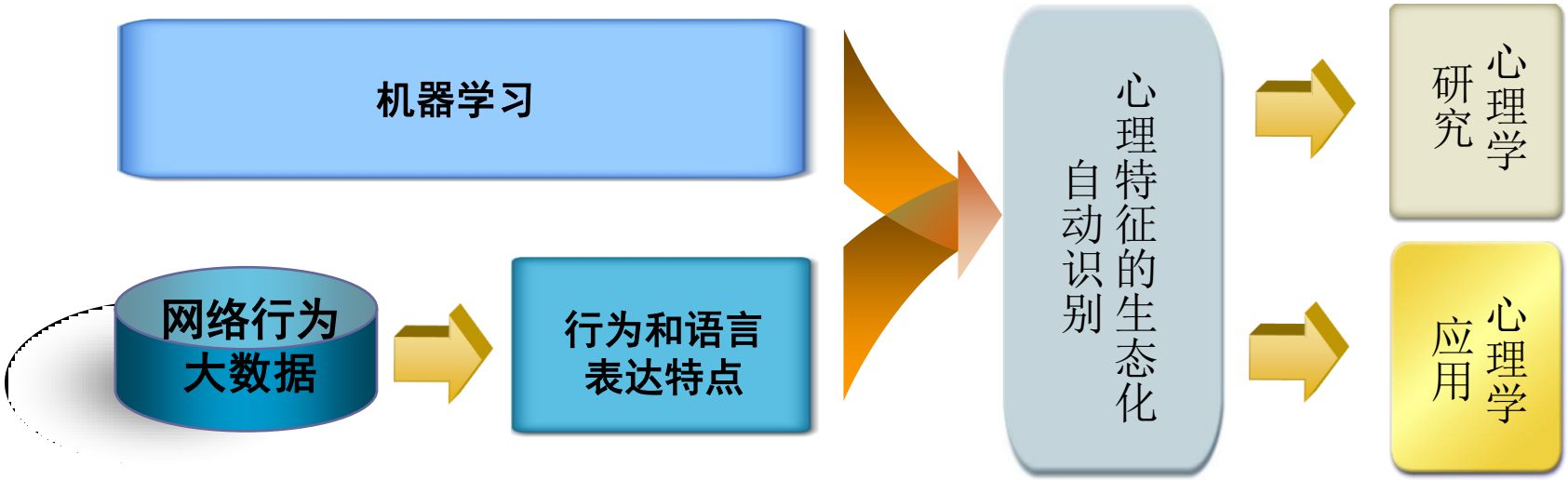
“做一天的好人并不难，难的是做一辈子好人”

生态化

时间可回溯

穿山甲公主归案

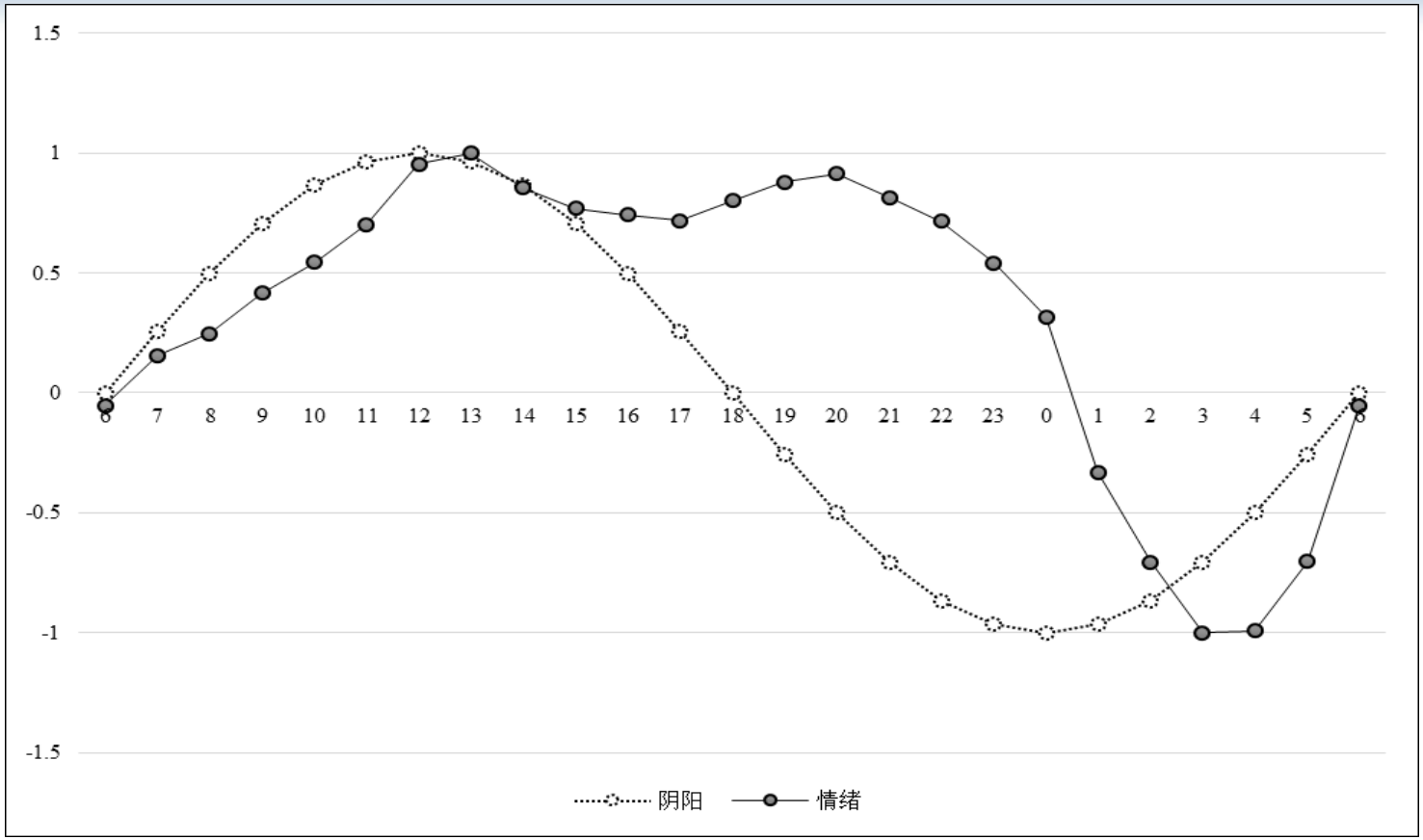
为心理学提供了最具生态化的研究平台



提纲

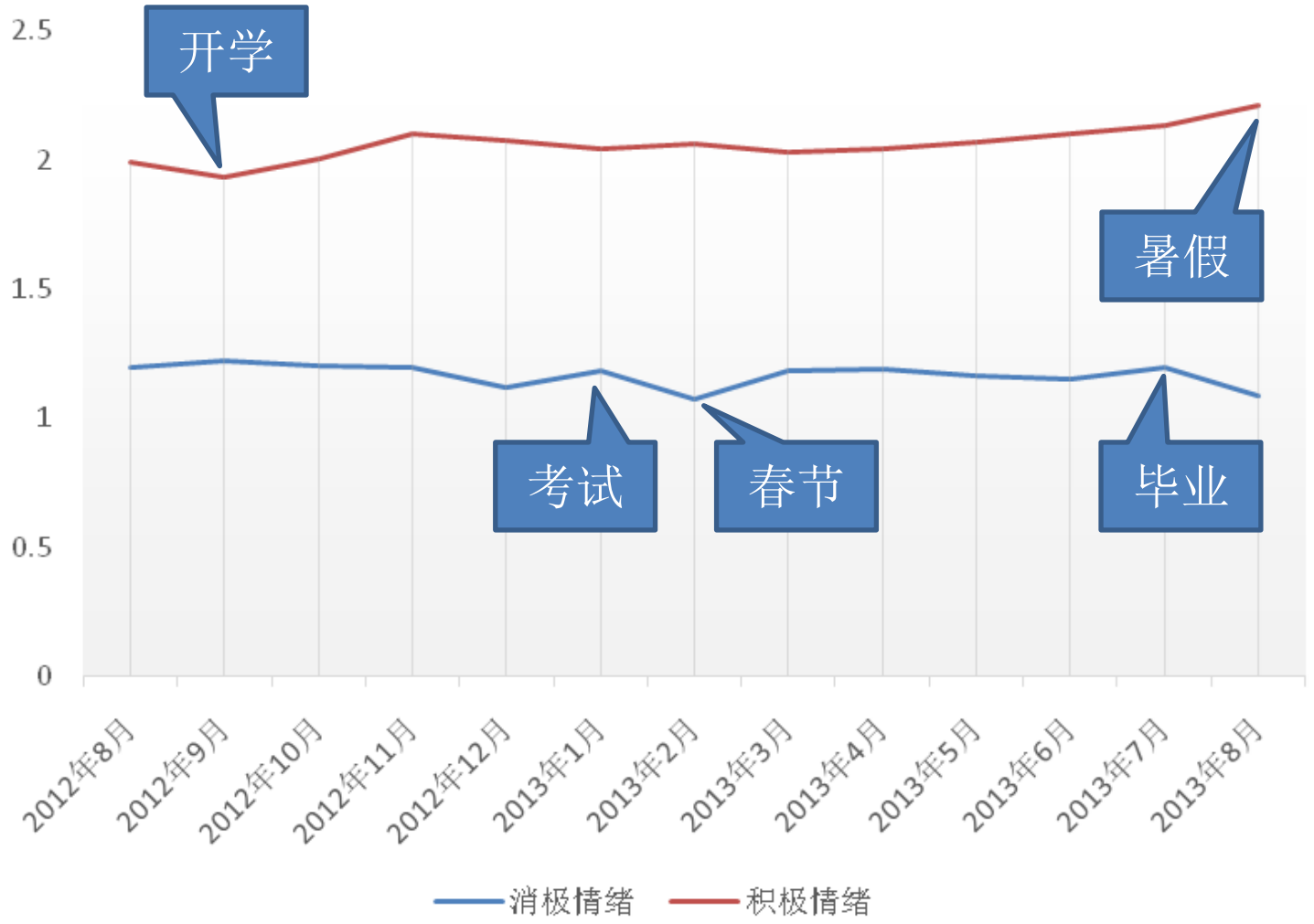
- 行为及语言表达特点
- 心理特征的生态化自动识别
- 心理学研究---家庭暴力
- 心理学应用---自杀预防

情绪在一天中的变化



从凌晨3点到下午15点这个时间段里，人们的情绪与阴阳变化的趋势是相似，但从下午17点开始至凌晨2点则不同。古人“日出而作，日落而息”，而今人们在夜幕降临之后还有很长一段的活动时间。

大学生群体的情绪表达



2013中国大学生『微博』发展报告

独生子女与非独生子女真的差别很大吗？

独生子女群体和非独生子女群体在**互粉数**上的T检验呈现显著性差异(p=.000)，但在``粉丝数"和``关注数"上未呈现显著性差异，这就说明这两个群体在社交圈子上并没有显著性差异。

在自我呈现方面，即``描述评估"、``描述长度"、``昵称长度"和``标签数"，独生子女在这四个相关网络行为上都没有显著性差异。

在网络社交活动方面，独生子女在发``**微博数**"和`` **图片数**"上显著超过了非独生子女(p=.003和p=.000)，但在其他网络社交活动上，其差异性并不显著。

在网络使用习惯上，独生子女**发第一条微博**比非独生子女相对更早(p=.003)，但在其他习惯上，则都没有显著性差异。

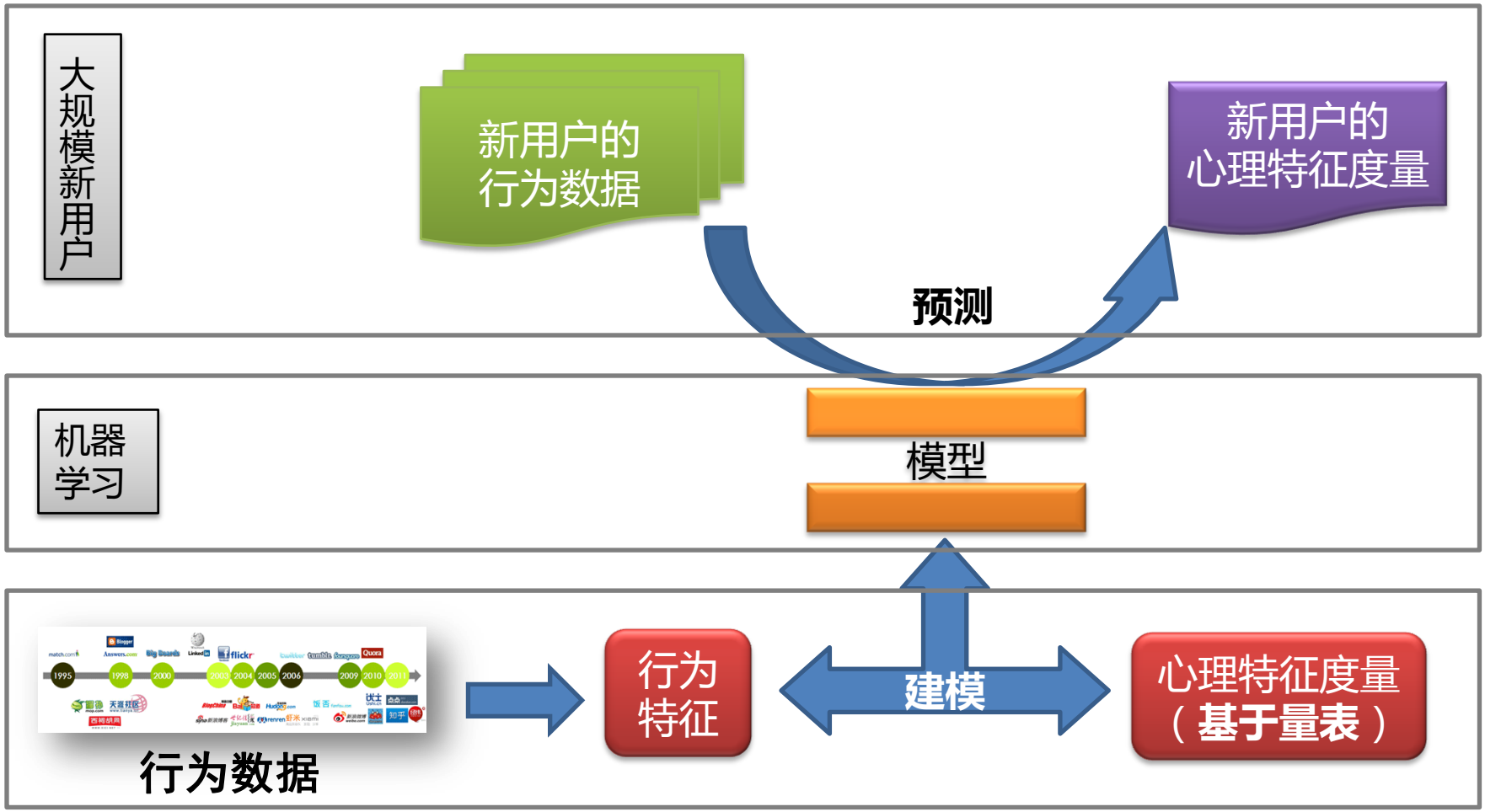
微博用户对于二胎的态度

		%	(n)微博
二胎态度			N=12,497
	Neutral	43.27	(5,407)
	Negative	30.88	(3,859)
	Positive	25.85	(3,231)
二胎倾向			N=4,978
	Negative	62.88	(3,130)
	Positive	37.12	(1,848)
二胎阻碍			N=3,429
	Cost and burden	56.55	(1,939)
	Public welfare benefits	17.35	(595)
	Condition of a child's environment	12.63	(433)
	Change of traditional opinion on childbirth	6.36	(218)
	Health condition	5.16	(177)
	Opinion of important people	1.95	(67)

提纲

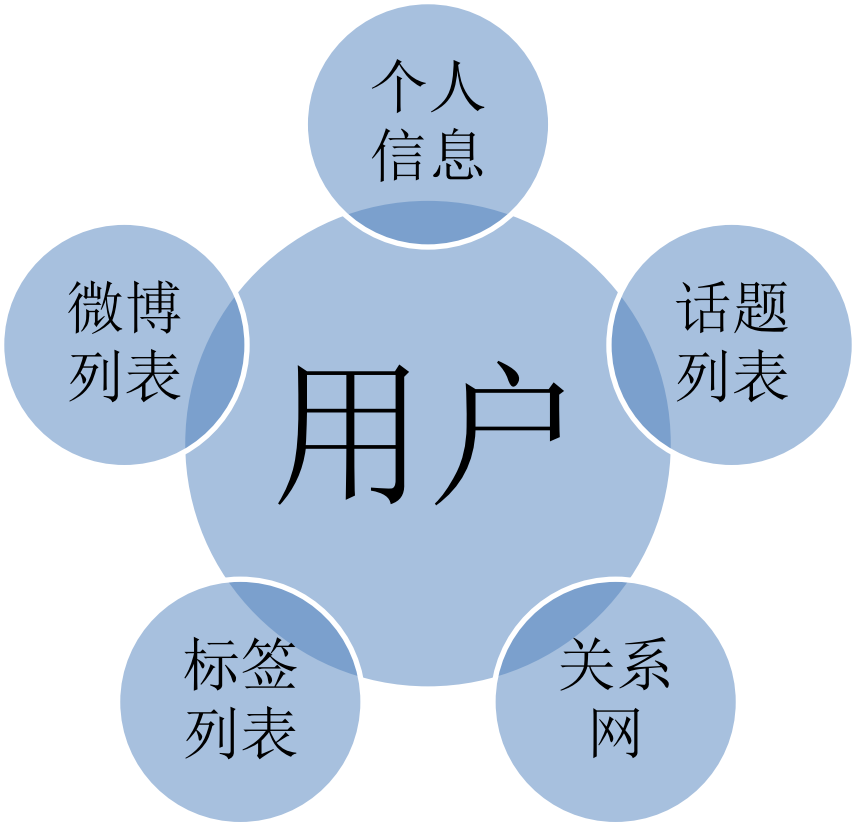
- 行为及语言表达特点
- 心理特征的生态化自动识别
- 心理学研究---家庭暴力
- 心理学应用---自杀预防

大数据机器学习运用于心理特征预测的两阶段过程

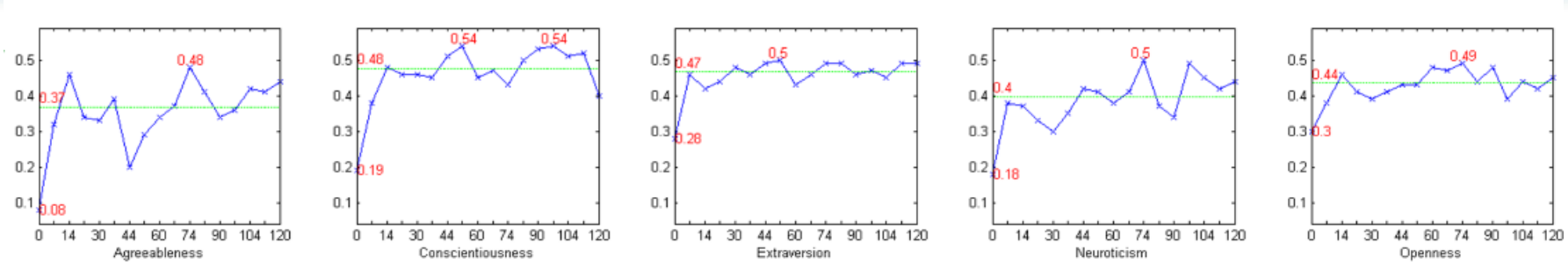


活跃用户数据下载

106万用户的微博数据（8T）
（2012.08-2012.09）



大数据对你的了解超过你的家人



Correlation Coefficient of PaceRegression Model

- 本研究中建立模型的**预测水平可以接近甚至超过用户的家庭成员对其的了解**。
  - 有心理学研究表明，现实生活中，用户的家庭成员对用户人格各维度的评价，与其自身通过问卷测量的相关系数在0.5左右 *[Connelly 2010]* 。

- 本文在中文社会媒体平台（微博）上建立的研究模型，预测能力**与国际上的学者同期**在英文社会媒体平台上建立的预测模型相比，**预测效果相当或略优**。
  - 在本研究开展的同时，国际上亦有学者在Facebook、Twitter等社会媒体平台上开展同类研究。截止目前，我们调研发现，研究者在社会媒体平台上建立的预测模型，能够以相关系数约0.4~0.6的平均水平预测用户在人格等心理特特征上的得分 *[Wu and Kosinski 2015]* 。

幸福感预测

回归模型评估

- 随机猜: -0.05~0.05
- Baseline：[0.17, 0.30] (仅仅使用人口统计学信息)
- 最佳组合: [0.27, 0.60]
- 社会心理学研究认可: [0.39, 0.68]

Feature Set	Algorithm	Emotional Well-being		Positive Functioning					
		P.A.	N.A.	S.A.	P.L.	E.M	P.R.	P.G.	A.I.
B	StepWise	.24	.16	.21	.13	.22	.18	.21	.21
	LASSO	.16	.11	.15	.00	.19	.14	.00	.16
	MARS	.16	.08	.14	.04	.05	.14	.07	.13
	SVR	.19	.13	.12	.06	.17	.13	.10	.15
L [±1Week]	StepWise	.22	.16	.15	.23	.20	.16	.23	.25
	LASSO	.17	.10	.15	.16	.14	.00	.22	.16
	MARS	.19	.10	.12	.14	.15	.14	.21	.12
	SVR	.11	.09	.08	.21	.20	.12	.17	.18
D+B	StepWise	.27	.22	.23	.16	.30	.21	.20	.30
	LASSO	.20	.16	.19	.11	.24	.18	.12	.25
	MARS	.19	.06	.17	.02	.20	.12	.06	.23
	SVR	.13	.13	.11	.21	.13	.17	.24	.07
D+L [±1Week]	StepWise	.24	.19	.22	.26	.25	.20	.22	.28
	LASSO	.20	.13	.19	.23	.22	.20	.23	.25
	MARS	.16	.07	.04	.09	.06	.17	.14	.18
	SVR	.24	.11	.16	.16	.30	.22	.19	.24
B+L [±1Week]	StepWise	.23	.21	.18	.22	.18	.19	.21	.26
	LASSO	.24	.20	.11	.20	.17	.18	.24	.20
	MARS	.10	.03	.09	.07	.04	.16	.00	.10
	SVR	.18	.13	.14	.11	.22	.19	.10	.22
D+B+L [±1Week]	StepWise	.45	.26	.35	.45	.41	.45	.51	.40
	LASSO	.38	.26	.29	.34	.35	.34	.42	.35
	MARS	.40	.24	.30	.43	.45	.38	.60	.40
	SVR	.41	.27	.30	.35	.38	.39	.49	.34
Best Result		.45	.27	.35	.45	.45	.45	.60	.40
Baseline (Only D)	StepWise	.23	.17	.31	.27	.23	.30	.25	.20
	LASSO	.14	.10	.21	.14	.09	.21	.15	.13
	MARS	.16	.03	.27	.19	.07	.21	.12	.13
	SVR	.19	.14	.28	.19	.08	.22	.23	.14

1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.00

“Purpose in Life” across China

Purpose in Life: High scorers have goals and a sense of directedness in life, they feel that there is meaning to their life both currently and in the past, they hold beliefs that give life purpose, and they have aims and objectives for living.

生命意义量表: 生活质量、生命价值、生活目标、生活自由

“Life of Satisfactory” across China

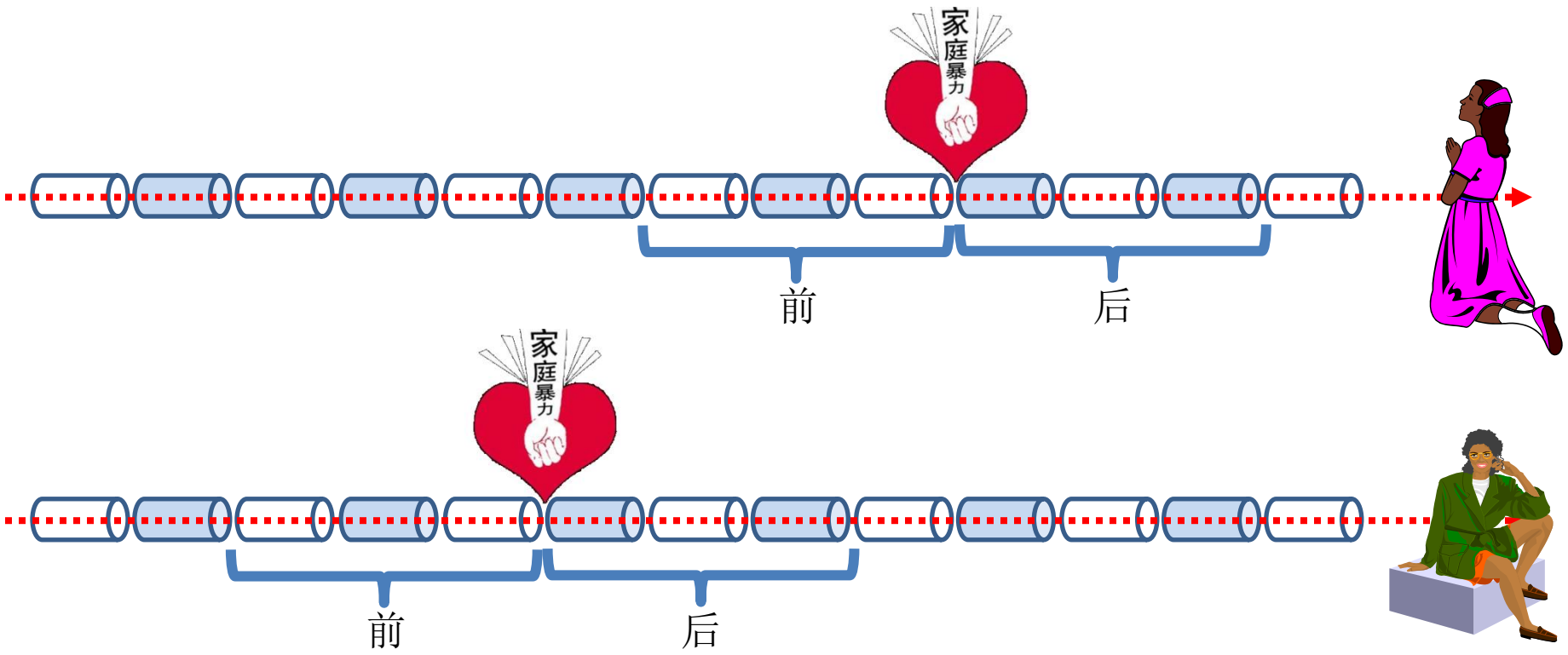
Life of Satisfactory(生活满意度)

提纲

- 行为及语言表达特点
- 心理特征的生态化自动识别
- 心理学研究---家庭暴力
- 心理学应用---自杀预防

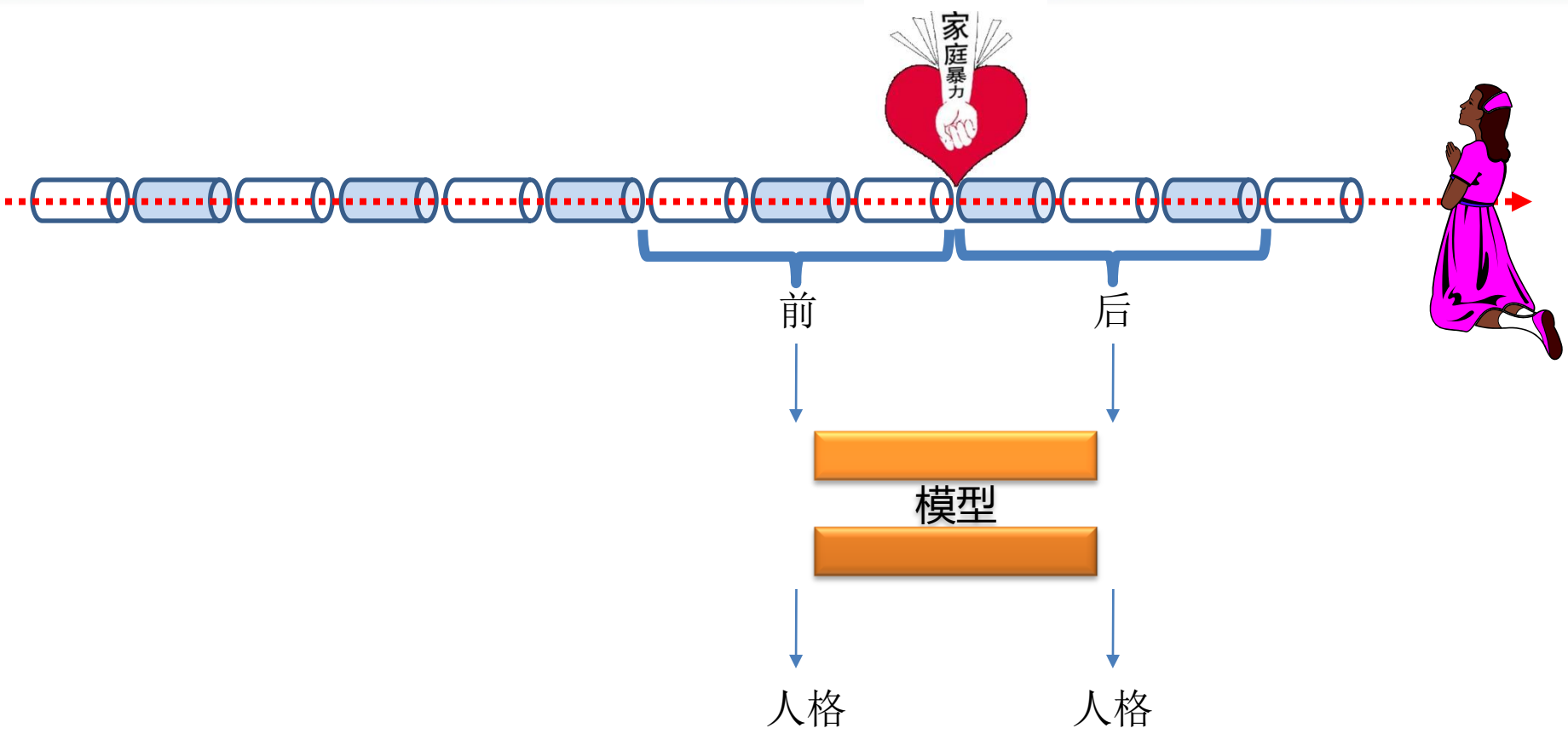
利用微博数据研究家庭暴力对心理的影响

个体遭受家庭暴力的时间点不同，传统方法难以获取

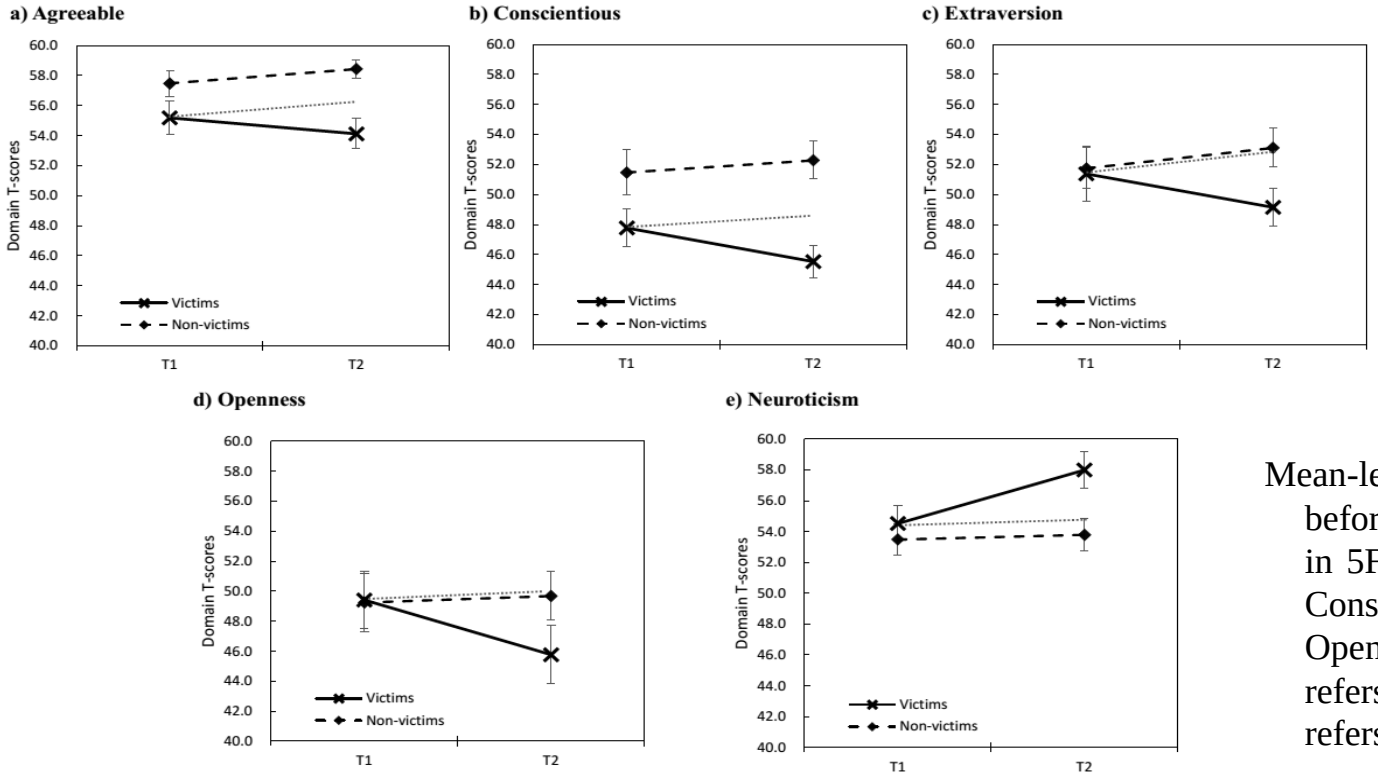


穿越时空来填量表

利用微博数据计算受害者前后的心理特征



家庭暴力前后的人格变化



Mean-level change of personality before and after domestic violence in 5FM domains: a) Agreeable, b) Conscientious, c) Extraversion, d) Openness and e) Neuroticism. T1 refers the month before DV, T2 refers the month after DV.

家暴组与非家暴组区别

Indexes	Dimensions		The victim group（n=232）		The non-victim group（n=232）	
			M ±SD	*t*	M ±SD	*t*
Depression	Depression	Before	15.07 ±2.22	*-4.080***	14.77 ±4.55	*0.129*
		After	15.59 ±2.06	-	14.68 ±2.13	-
Suicide Probability	Suicide ideation	Before	13.47 ±0.95	*-2.138**	13.41 ±1.34	*-0.824*
		After	13.58 ±1.03	-	13.47 ±1.27	-
	Hostility	Before	11.50 ±1.04	*-2.675***	11.53 ±1.74	*0.509*
		After	11.68 ±1.16	-	11.57 ±1.61	-
	Negative self-evaluation	Before	19.68 ±2.43	*-1.118*	19.95 ±3.38	*0.28*
		After	19.83 ±2.41	-	19.89 ±2.94	-
	Hopeless	Before	24.93 ±2.33	*-2.283**	25.04 ±3.66	*-0.408*
		After	25.23 ±2.52	-	25.11 ±3.38	-
Life Satisfaction	Satisfaction with life	Before	12.13±0.73	*3.087***	12.23±1.12	*0.459*
		After	12.00±0.67	-	12.19±1.12	-

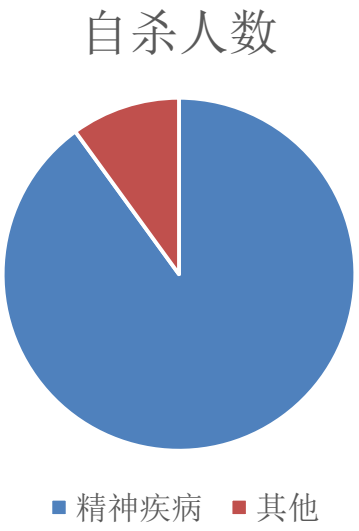
不同家暴类型的影响

INDEXES		DEPRESSION	SUICIDE PROBABILITY				SATISFACTION WITH LIFE
			Suicide ideation	Hostility	Negative self-evaluation	Hopeless	
Intimate partner violence（n=40）	**Before**	14.88 ±1.89	13.20 ±0.84	11.34 ±1.08	19.79 ±2.57	24.50 ±1.55	12.07 ±0.70
	After	15.53 ±2.20	13.26 ±0.84	11.41 ±1.09	19.92 ±2.37	24.80 ±1.79	12.01 ±0.61
	t	*-2.648***	*-0.518*	*-0.35*	*-0.333*	*-1.109*	*0.582*
Child abuse（n=151）	**Before**	15.06 ±2.28	13.60 ±0.99	11.60 ±1.06	19.72 ±2.68	25.17 ±2.55	12.13 ±0.78
	After	15.56 ±1.87	13.66 ±1.05	11.75 ±1.17	19.95 ±2.41	25.44 ±2.69	11.95 ±0.70
	t	*-2.981***	*-1.066*	*-1.812*	*-1.314*	*-1.674*	*3.293***
Exposure to DV（n=41）	**Before**	15.30 ±2.34	13.26 ±0.84	11.25 ±1.06	19.38 ±2.16	24.51 ±1.99	12.18 ±0.59
	After	15.79 ±2.59	13.56 ±0.90	11.69 ±1.16	19.30 ±2.43	24.86 ±2.43	12.16 ±0.59
	t	*-1.602*	*-2.478**	*-2.551**	*0.323*	*-1.13*	*0.229*

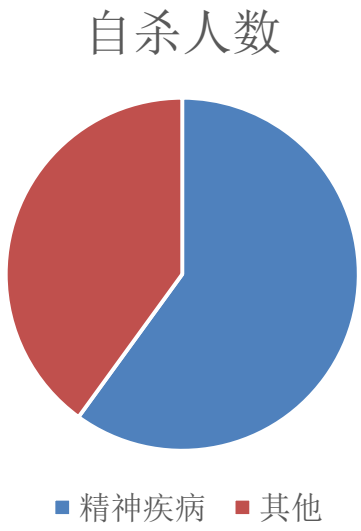
提纲

- 行为及语言表达特点
- 心理特征的生态化自动识别
- 心理学研究---家庭暴力
- 心理学应用---自杀预防

自杀是一个严重的社会问题



国外



国内

微博带来的新机遇

- 自杀预防
  - 多一种渠道捕捉自杀的讯号
  - 多一种渠道传递关怀
  - 多一种渠道提升公众对自杀问题的认识，鼓励有自杀念头的人士向他人求助。
  - 多一种渠道提供危机干预服务。

- 自杀研究
  - 收集自有自杀意念人士的一手数据，深入了解他们的心态和行为；
  - 监测大众对自杀信息的反应，深入研究自杀的传染性问题；
  - 监测、试验自杀预防类信息在人群的扩散，研究各种基于网络的自杀预防、干预措施的有效性。

自杀死亡和无自杀意念新浪用户的微博特征差异

31

自杀死亡组

贝克自杀意念量表 SSI

自杀可能性量表 SPS

抑郁自评量表 SDS

30

对照组

微博特征差异

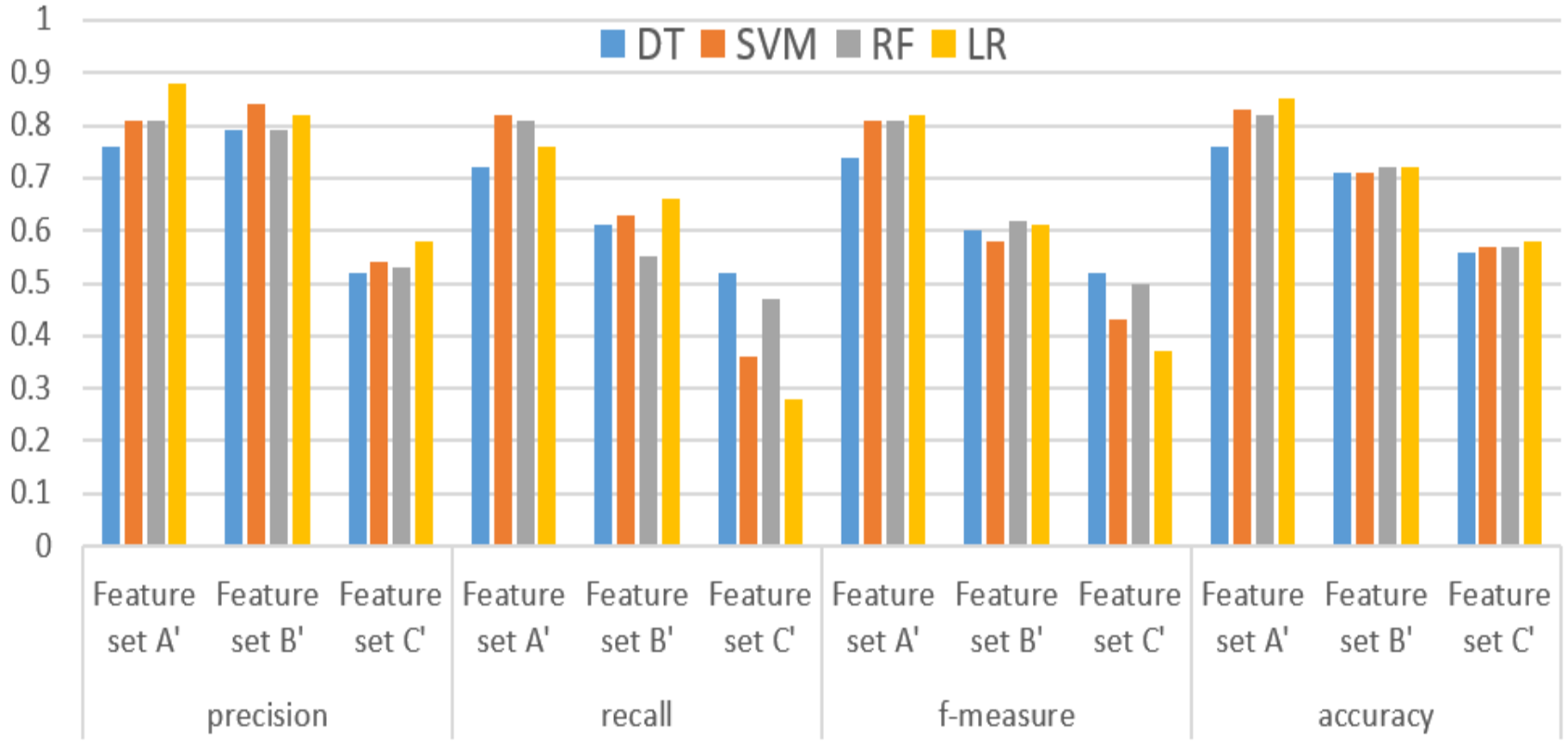
行为特征	自杀死亡组（n=31）		对照组（n=30）		U值	W值	Z值	P值
	M	Q_R	M	Q_R				
微博链接率	0.04	0.08	0.06	0.08	314.0	810.0	-2.18	0.029*
微博互动率	0.60	0.54	0.69	0.36	313.0	809.0	-2.19	0.028*
自我关注度	0.47	0.50	0.30	0.20	286.0	751.0	-2.58	0.010*

语言特征	自杀死亡组（n=31）	对照组（n=30）	t值	P值
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

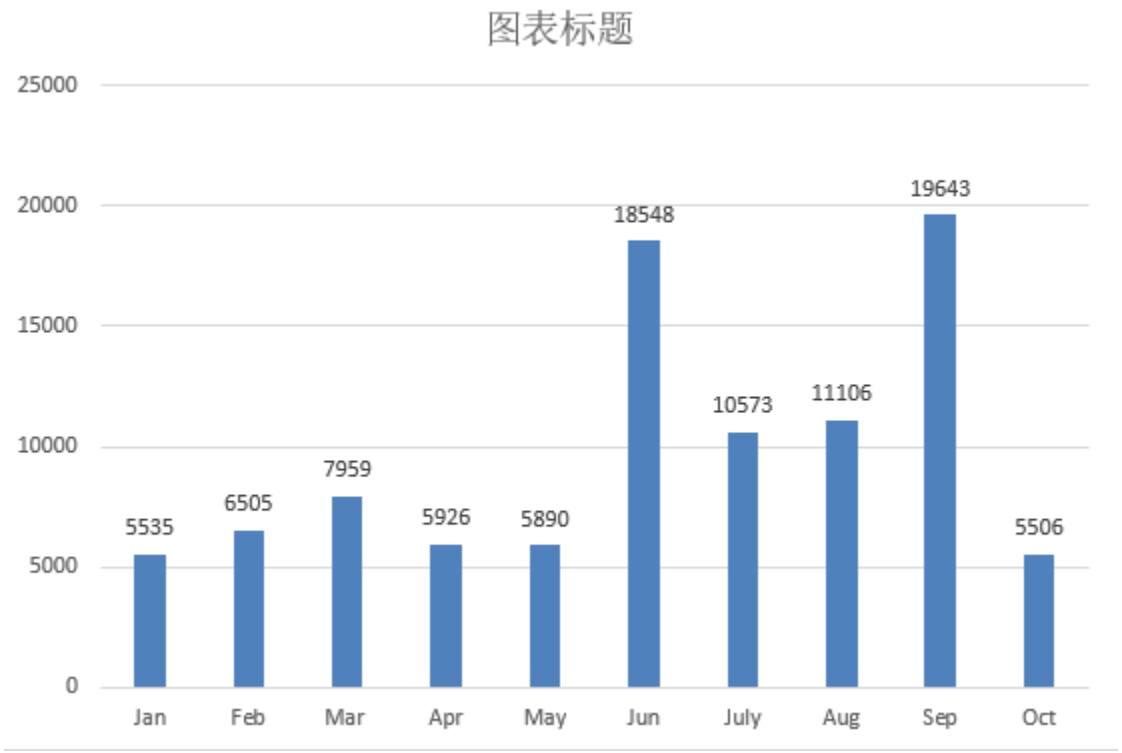
自杀死亡用户的微博互动更少，更加关注自我，更频繁地使用表达排除意义的词语，从情感层面上有更多负性表达，使用更多与死亡、宗教相关而更少与工作相关的表达。这些特点对于开展针对网络用户的自杀学研究有重要的启示

语言特[illegible]								
第二人称[illegible]								[illegible]**
人类词[illegible]								****
消极情绪词	0.59	0.44	0.38	0.16	230.0	695.0	-3.39	0.001**
愤怒词	0.12	0.22	0.06	0.04	281.0	746.0	-2.65	0.008**
悲伤词	0.15	0.16	0.07	0.06	162.0	627.0	-4.37	<0.001**
工作词	0.33	0.26	0.45	0.42	292.0	788.0	-2.50	0.013*
死亡词	0.16	0.14	0.11	0.08	269.0	734.0	-2.83	0.005**
省略号	0.02	0.08	0.08	0.20	302.5	798.5	-2.35	0.019*

单条微博自杀意念识别


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
DT
SVM
RF
LR
Feature set A'
Feature set B'
Feature set C'
precision
recall
f-measure
accuracy

微博来源



2016年 走饭第一条微博下评论更新情况

焦点小组访谈更好的编制微博私信

目的：更好的编制微博私信，提高私信的点击和反馈率

受访者招募：在心理所的4个被试QQ群内发布招募信息，通过ASIQ-4问卷进行筛选，问卷得分大于1的作为合格受访者，开展了两次访谈。

访谈共涉及主题

1、微博用户私信的使用习惯和对微博私信的态度；

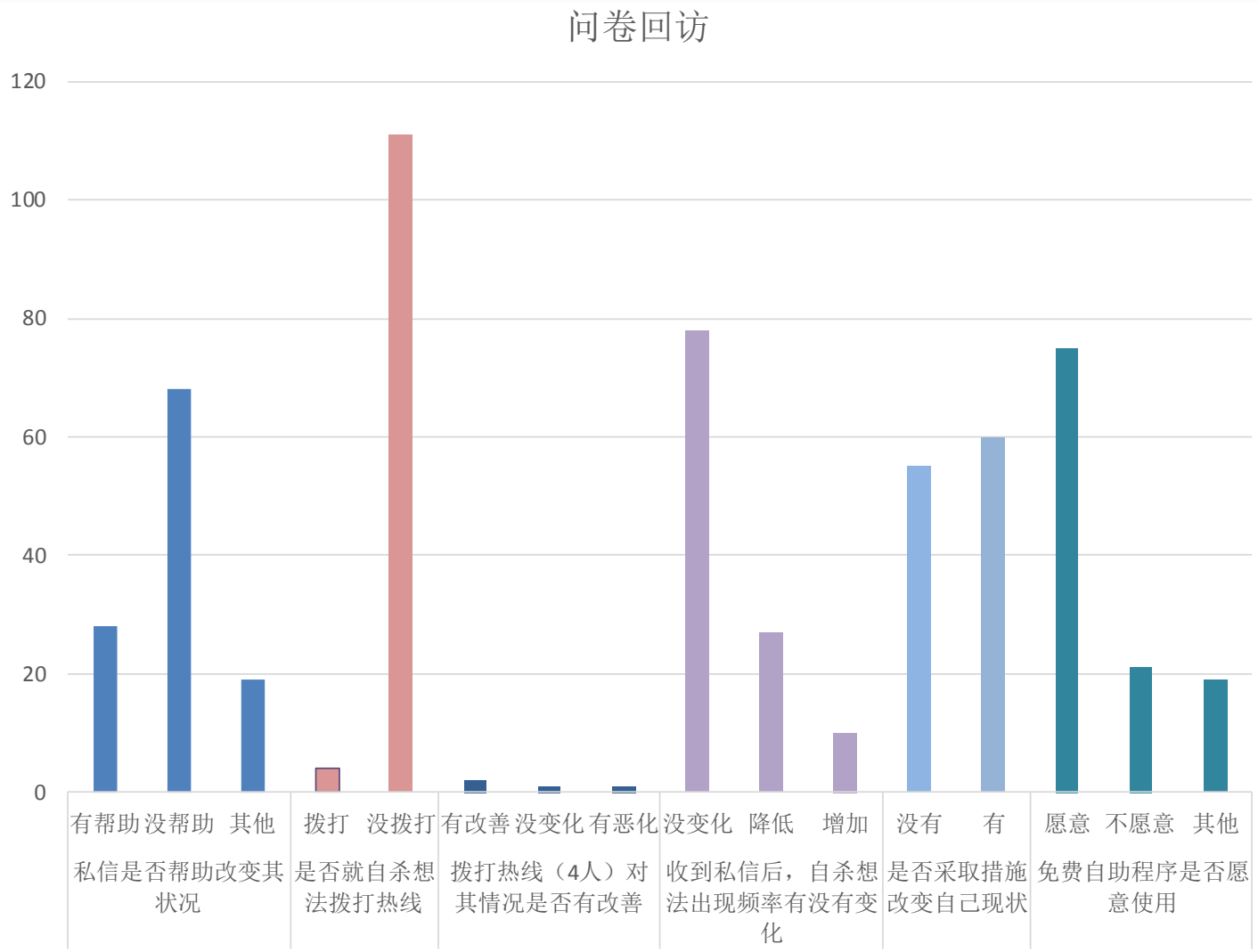
2、当前干预推送私信的优缺点。鉴于微博私信内容字数的局限性，我们将私信的形式设置为：文字+链接的形式；

3、如何更好的设计干预的私信，使得有自杀风险的用户更容易接受。

初步私信沟通

2016年11月，17日下午7点到九点, 18日下午4点到四点半，通过调用Sina API向 4222 微博用户发送私信。

私信回复情况


问卷回访
120
100
80
60
40
20
0
有帮助 没帮助 其他
私信是否帮助改变其状况
拨打 没拨打
是否就自杀想法拨打热线
有改善 没变化 有恶化
拨打热线（4人）对其情况是否有改善
没变化 降低 增加
收到私信后，自杀想法出现频率有没有变化
没有 有
是否采取措施改变自己现状
愿意 不愿意 其他
免费自助程序是否愿意使用

大数据之于心理学

- 大数据为心理学研究带来新的机遇：

  大范围、跨时空、生态化；

- 开展跨学科合作，充分利用现代信息技术；

- 注重隐私保护，恪守伦理！

在对客观数据的全面准确分析之上，开展心理学研究，将在效率和效果上实现新的飞跃。

大数据教学情况

中国科学院大学

- 2010年秋季
- 2011年秋季
- 2012年秋季
- 2013年秋季
- 2014年秋季

  计算网络心理学
- 2015年夏季

  大数据与心理学
- 2016年秋季

  应用心理学高级课程—大数据心理学

中国科学院心理研究所继续教育学院

2015年，大数据心理学专业

大数据课程设置

- 计算机基础：编程、数据分析技术、机器学习等；
- 心理学基础：心理测量以及相关心理学课程；
- 大数据应用于心理学研究的主要思路，结合实例；
- 基于目前的研究，介绍主要内容，加以文献阅读；
- 不同专业同学形成课程设计小组，互相交流。

专业设置总结（个人观点，仅供参考！）

- 大数据作为一种技术，与具体的场景相结合；
- 研究先行，将最新的研究成果与教学密切结合，教研相长；
- 结合实例，突出课程内容的实用性，不是单纯介绍技术；
- 大数据作为一种研究方法，需要我们转变研究思路。

谢谢！

Thank You！

中国科学院心理研究所计算网络心理实验室

Computational CyberPsychology Lab. (CCPL),

Institute of Psychology, Chinese Academy of Sciences

http://ccpl.psych.ac.cn