

# 用户画像与聊天机器人

谢 幸

微软亚洲研究院

# 用户大数据



A screenshot of a digital calendar interface. At the top, there are several colored boxes representing different calendar categories: Tennis Club (orange), Student D (purple), and Student E (yellow). Below this, a specific event for "Tennis Club" is listed for "10am". There are options to "Display only this Calendar", "Hide this calendar from the list", "Calendar settings", "Create event on this calendar", and "Share this Calendar". The "Share this Calendar" option is highlighted with a red oval.



A screenshot of a mobile website for movie tickets. The top navigation bar shows "Fandango.com" and "Mobile Ticket". The main content area displays movie information for "Man of Steel (2013)" and "The Hobbit: The Desolation of Smaug (2013)". It shows showtimes, ticket prices, and a "Play Trailer" button. Below this, a section for "Top Billed Cast" is shown. On the right side, there is a QR code and a summary of ticket details: "The Book of Eli", Date: Jan 15, 2010, Time: 5:45 PM, Adult: 2 Adult, Confirmation #: 1070292, Auditorium: 7, Theater: Regal Union Square Stadium 14, Address: 1000 Union Street.

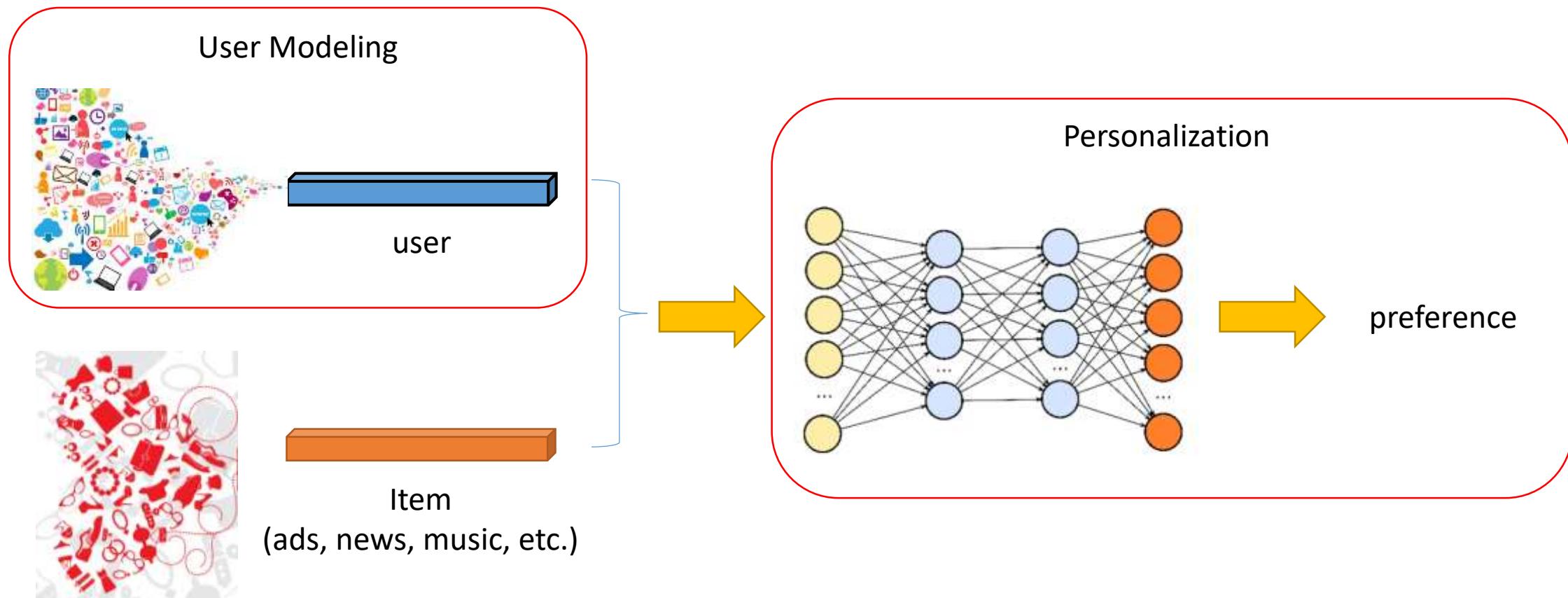


# 用户画像



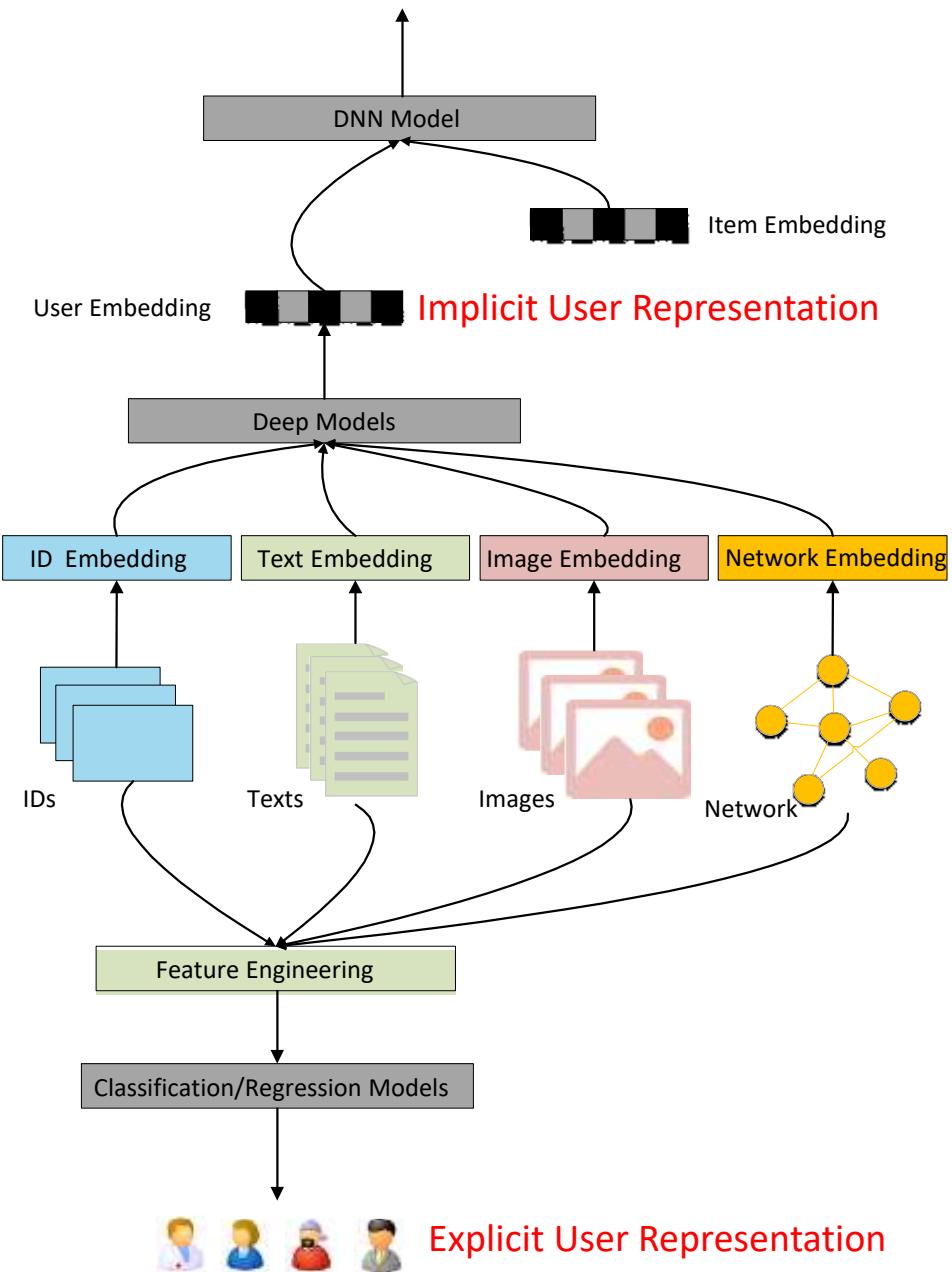
胡思乱想 白举纲 it民工 陈赫 前端开发 腹黑女  
自由主义 85后 死理性派 有条不紊 书法 移动互联网  
恐怖小说 机器学习 爱自拍 段子 诺基亚 足球 理性  
恐怖 小说 产品经理 科技 李晨 读书 诺基亚  
足球 新闻 苹果 软件 程序员 软件 兴趣  
理性 足球 android 手机 javascript 球迷  
雕刻 球迷 windowsphone 安卓  
法医 微软 宜兴  
单纯  
互联网

# 用户画像与个性化服务

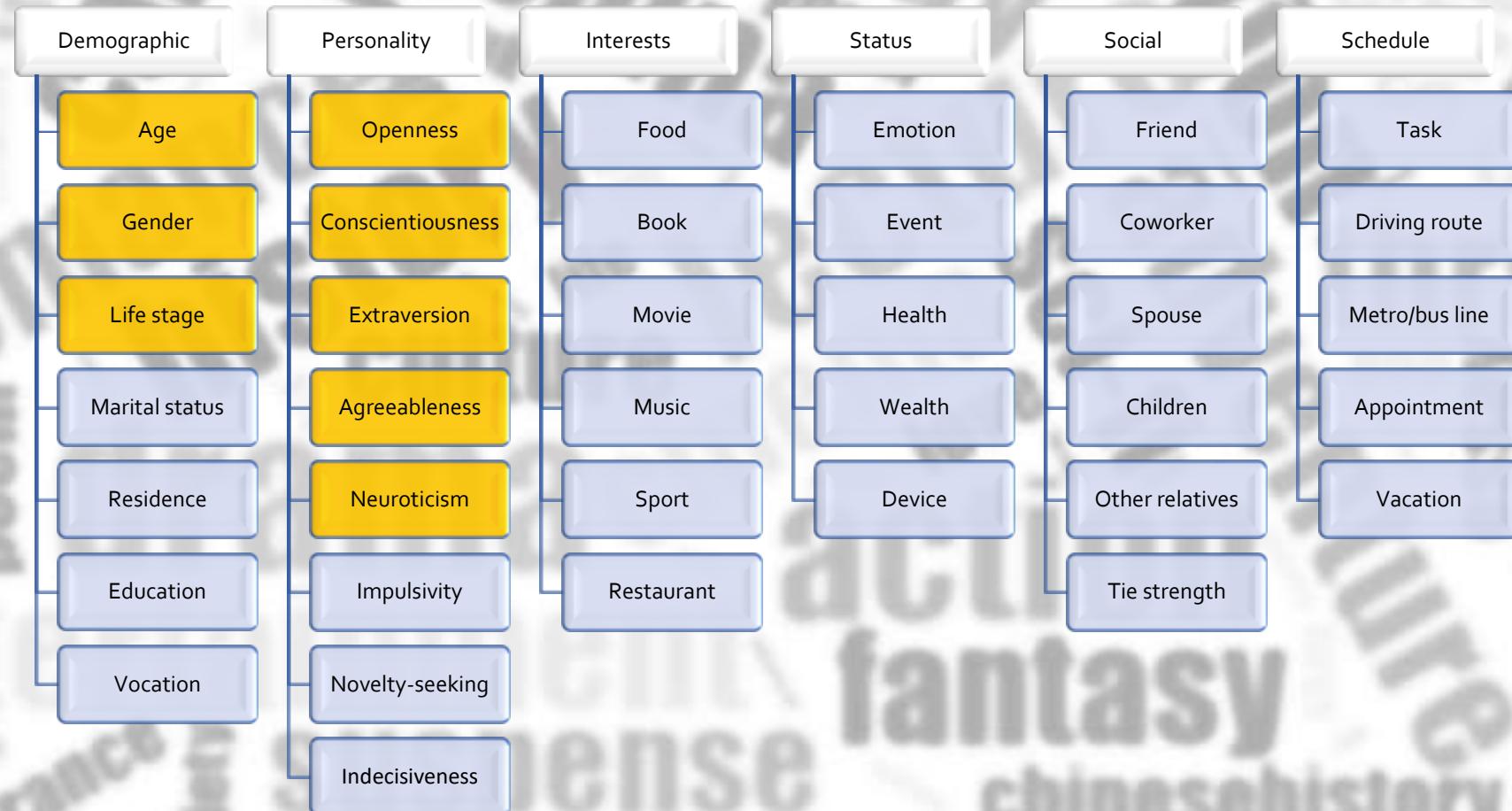


# 显式画像与隐式画像

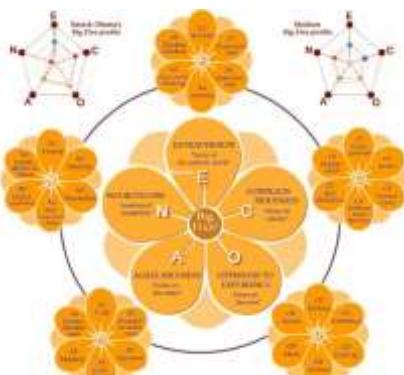
Representation	Pros	Cons
Explicit	<ul style="list-style-type: none"> <li>Easy to understand;</li> <li>Can be directly bidden by advertisers</li> </ul>	<ul style="list-style-type: none"> <li>Hard to obtain training data;</li> <li>Difficult to satisfy complex and global needs;</li> </ul>
Implicit	<ul style="list-style-type: none"> <li>Unified and heterogenous user representation;</li> <li>End-to-end learning</li> </ul>	<ul style="list-style-type: none"> <li>Difficult to explain;</li> <li>Need to fine-tune in each task</li> </ul>



# 显式用户画像



# 相关研究工作



Big Five Personality

WSDM 2017



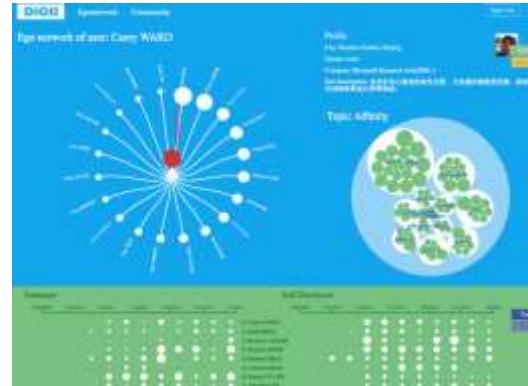
Location Interests

IJCAI 2017



Consumer Impulsivity

UbiComp 2015



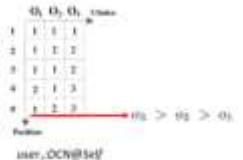
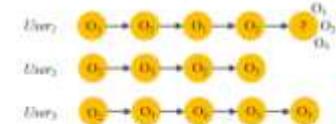
Dynamics of Online Intimacy

WSDM 2016

## Novelty Seeking Model

### Item Novelty Matrix

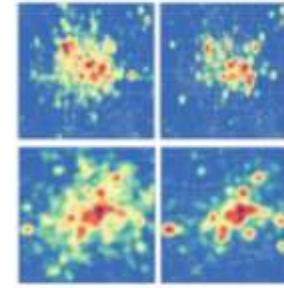
- $N \times M$  Matrix
- At the position, facing  $M$  choices, novelty is determined as a partial order
- Two factors determine the order
  - Popularity of item itself



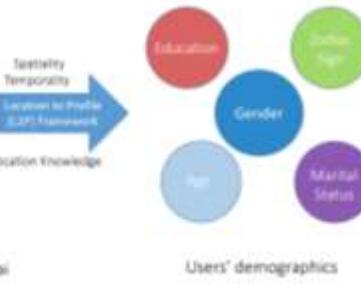
Novelty Seeking Trait

WWW 2015/WWW 2014

## Profile inference from location check-ins



Location check-ins of Beijing and Shanghai

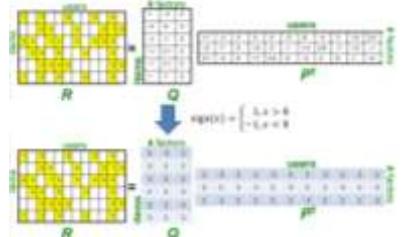


Users' demographics

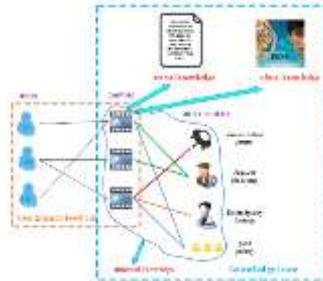
Location to Profile

WSDM 2015

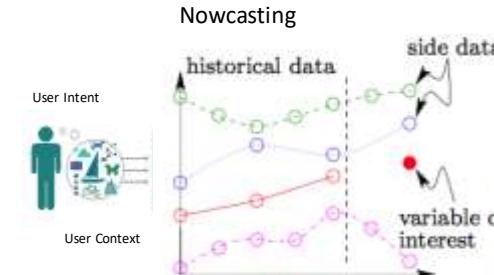
# 相关研究工作



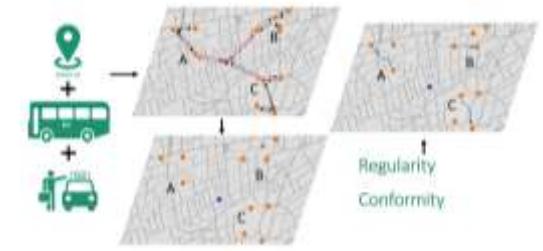
Discrete Content-aware MF  
KDD 2017



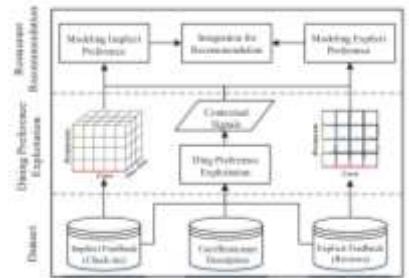
Knowledge Enhanced Recommendation  
KDD 2016



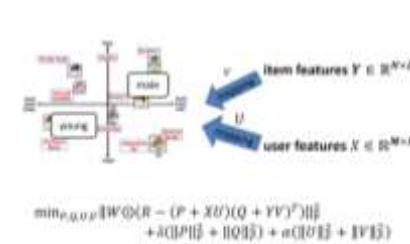
Contextual Intent Tracking  
KDD 2016/best student paper



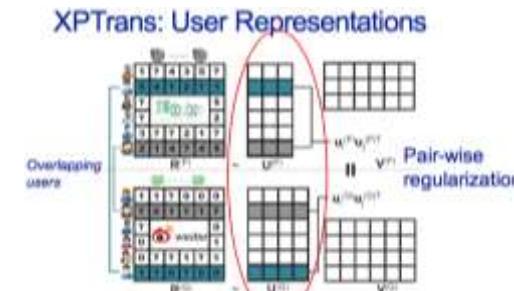
Regularity and Conformity  
KDD 2015



Exploiting Dining Preference  
WWW 2016



Bayesian Content-aware CF  
IJCAI 2016



Cross-Platform Behavior Prediction  
AAAI 2016

$$A = \sum_{i=1}^n U_i \circ C_i \circ V_i \circ S_i \quad \text{s.t. } D(S) = 0$$

App Usage Forecasting  
UbiComp 2016

# LifeSpec: 跨平台用户行为数据集

豆瓣douban

首页 发现小组 发现小站



DerPlough

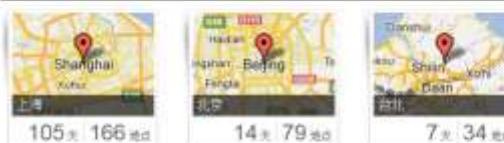
日记 相册 喜欢 广播 二手 豆列 留言板

最近喜欢: Don't stop the Music , You're Beautiful , Free Loop , 爱情转移

最新照片



城市护照



发表的攻略

- Munchies 梦奇屋 (武昌店)
- 大爱土豆丝

DerPlough 2010年12月24日 12:11 生活探秘类影 + 测量 + 回复

http://www.weibo.com/derplough  
上海, 红口区  
简介: 我觉得, 你睡觉... 这一生干嘛只要平凡快乐, 难道这样不懂大呢? 我就活在自己的世界里了! 你  
说到底谁在乎! [更多资料>>](#)

+ 关注 取消关注 推荐给朋友 情感关注

个人照片 查看全部



微博 心情 全部 原创 图片 视频 音乐 标签 高级搜索

火龙果蜂蜜奶昔 + 🍉Berry&Milkshake +



10月9日22:52 来自iPhone客户端 | 报告 转发 收藏 评论(0)

真爱無閑別，極力極外也和是否承認同性婚姻無關！ // @白負責小姐是超不是胖：我最尊  
重愛情了！

@第二类爱情: 【世界，请正视我们吧】有一种爱叫同性恋。中国约有8000万同性恋者，  
然而作为成年人，我们的声音、诉求却被遮蔽，大多数只能生活在柜子里。到目前，全世界已  
有11个国家承认同性婚姻。中国会成为第几个？滴滴转发此微博，让更多的人听到我们的声  
音。（真爱无性别@第二类爱情）

关注他的人同时关注了



茨茨De\_ 加关注



莉Sam... 加关注



美负责 加关注



vicky浅 加关注



吴小白 加关注



Johnn 加关注

他的标签(10)

雨男一宅猪日子

多年前的五月天

帆布鞋控格子控

狮子座月亮双鱼座 撞卯不成卦

双肩包 徒步旅行

过山车控 蔡健雅Tany

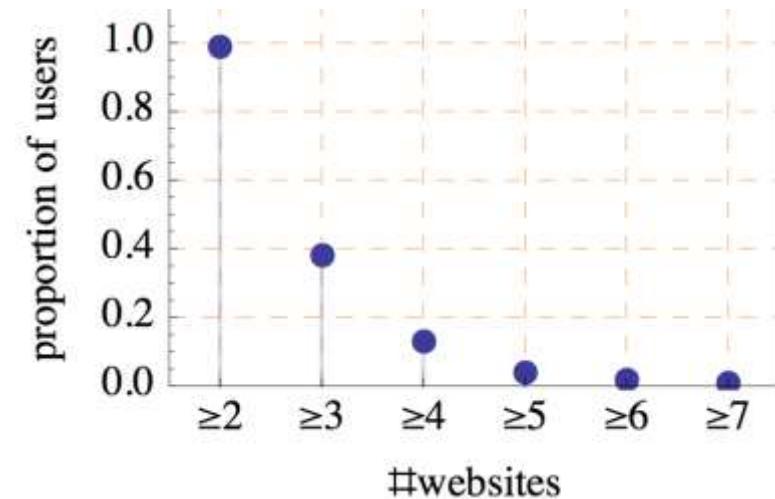
陈奕迅Eason

他关注的微吧(0)



# Data

- 4 (major) networks: Jiepang, Weibo, Douban, Dianping
- 1.4M+ unique (deterministically identified) users accounts
- Heterogeneous footprints: tweets, photos, check-ins, movies, books, music, offline events, online purchase history, etc.
- Rich user profiles integrated from different sites (publicly available)



Age	Gender	Residence	Relationship	Occupation	College	High School	Self description	...

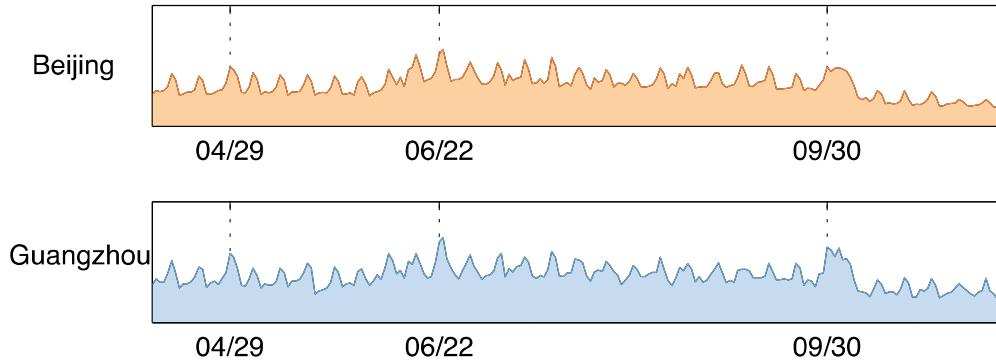
# Data

- 53 million footprints (check-in, movie, music, events, book, etc.)
- 3 million social links
- 39 million check-ins

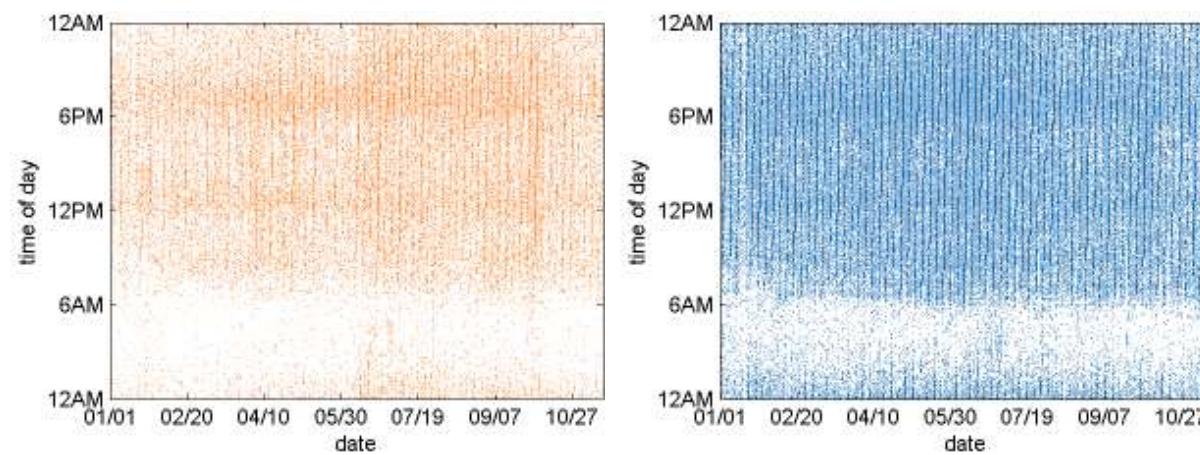
city	Shanghai	Beijing	Guangzhou	Tianjin	Hangzhou	Hongkong	Xiamen	Suzhou	Nanjing	Chengdu	Wuhan	Xian
users	417,681	162,764	53,089	15,490	34,322	12,599	10,123	19,673	21,558	23,372	20,975	15,261
check-in	25,178,189	5,898,447	1,092,138	392,943	619,219	424,650	369,231	560,274	414,202	327,634	321,646	229,678
movie	1,661,214	1,466,479	171,789	118,775	238,721	57,003	70,172	89,706	174,664	191,042	166,337	123,223
music	766,165	737,254	85,953	60,658	103,936	30,313	29,716	39,701	82,513	88,426	76,316	62,876
book	402,318	387,138	51,913	28,188	57,835	18,117	18,516	19,521	44,345	42,241	44,804	28,435
event	609,076	803,158	101,246	52,133	78,587	18,277	20,889	27,400	46,788	66,640	44,764	72,902
total	28,616,962	9,292,476	1,503,039	652,697	1,098,298	548,360	508,524	736,602	762,512	715,983	653,867	517,114

Table 1. Summarization of collected footprints for different cities (partially presented due to page limit).

# Mobility Footprints



**Figure 5. Daily trends of total check-ins in different cities.**

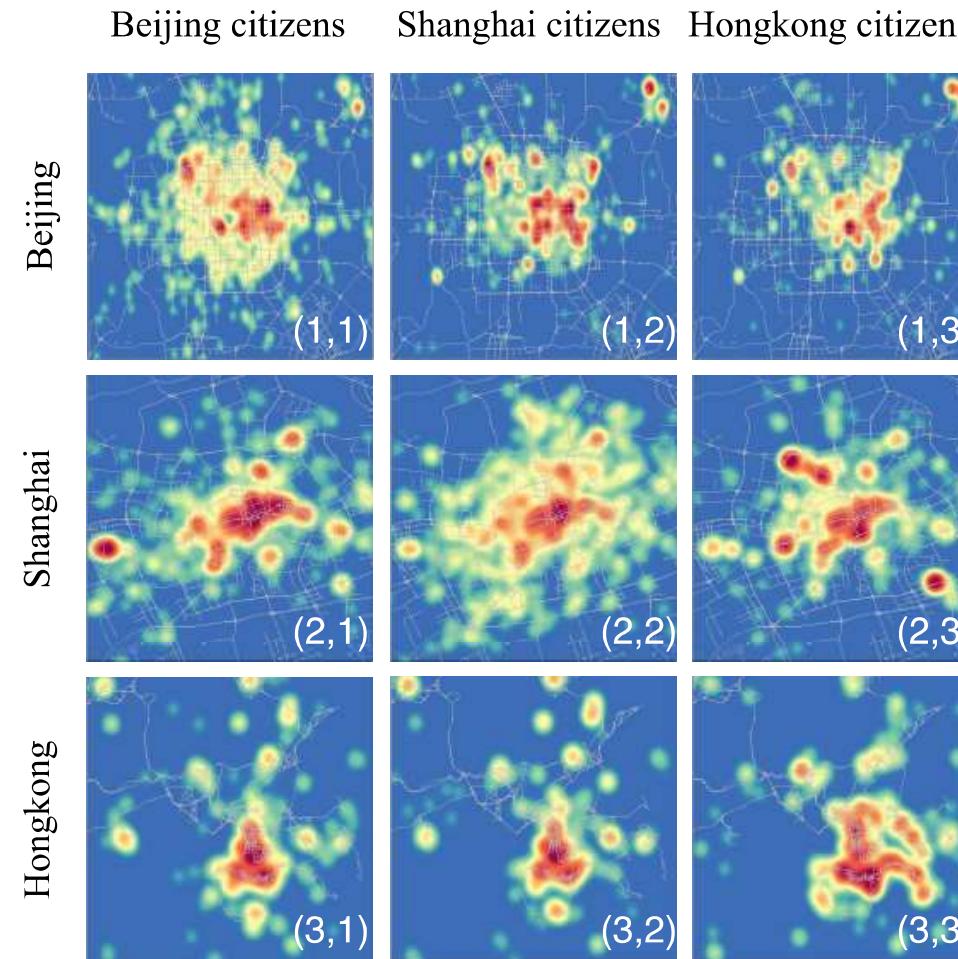


**Figure 6. Diurnal distribution of 1000 sampled users' check-ins.**

Rank	City	Bedtime
排名	城市名称	就寝时间
1	广州	23:08
2	东莞	22:53
3	厦门	22:37
4	西安	22:37
5	天津	22:32
6	长沙	22:29
7	重庆	22:25
8	武汉	22:24
9	郑州	22:19
10	成都	22:18
11	北京	22:15
12	昆明	22:10
13	南昌	22:09
14	淮南	22:08
15	上海	22:08
16	杭州	22:04
17	沈阳	22:04
18	西宁	22:03
19	南京	22:02
20	齐齐哈尔	21:50

Average bedtime of different cities  
(1M respondents, led by Chinese Medical Doctor Association, announced in world sleep day 2013)

# Mobility Footprints



**Figure 7. Check-in density distribution of 3 cities showing where people check-in in each other's cities.**

# Self-Disclosure

## Cross-Domain Posting

The screenshot shows a mobile interface with two feeds:

- Twitter Feed:** A post from user 山子 (@ant\_sz) about a physical check-up at Shanghai Jiaotong University. The URL in the post is 4sq.com/Zlx5ML.
- Foursquare Feed:** A post from user Victor Wong (@vkw) about meeting a founder of Virgin America. The URL in the post is 4sq.com/14d4NBL.
- Bottom Feed:** A post from user Xinxin's ❤️ (@Liongkaibin) about a meal at Taman Sentosa. The URL is 4sq.com/WKBrGL.

A red dashed box highlights the URL in Victor Wong's Foursquare post, and a black arrow points from it to the URL in the bottom Foursquare post.



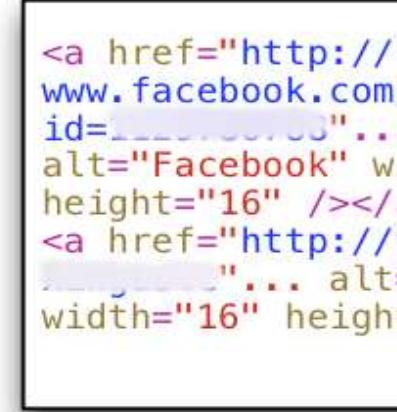
## Hub Sites

The screenshot shows the Foursquare homepage with the following elements:

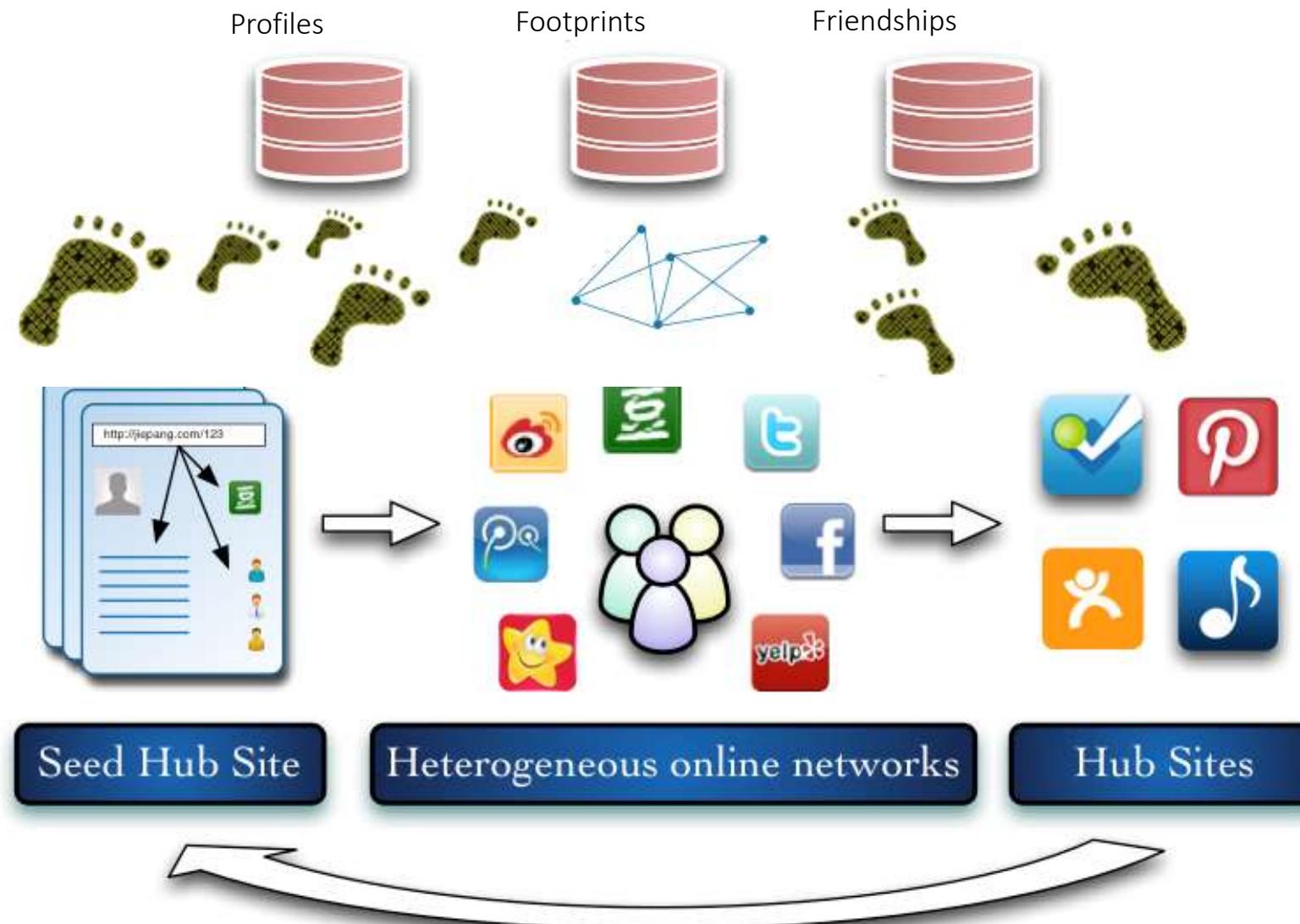
- Header:** "foursquare" logo and search bar "I'm looking for..."
- Middle Section:** "Get personalized recommendations for places near you!"
- Sign Up Options:** "Sign up with Facebook" (using a blue button with a white "f") and "Or take the scenic route: Sign up with Email".
- Bottom Section:** A blurred image of a person holding a smartphone, with a red dashed box highlighting the social media icons (Facebook, Twitter, and a globe) and the text "北京". Below this are the numbers "9 Check-ins" and "8 Days out".

A black arrow points from the "9 Check-ins" text to a red dashed box around the social media icons.

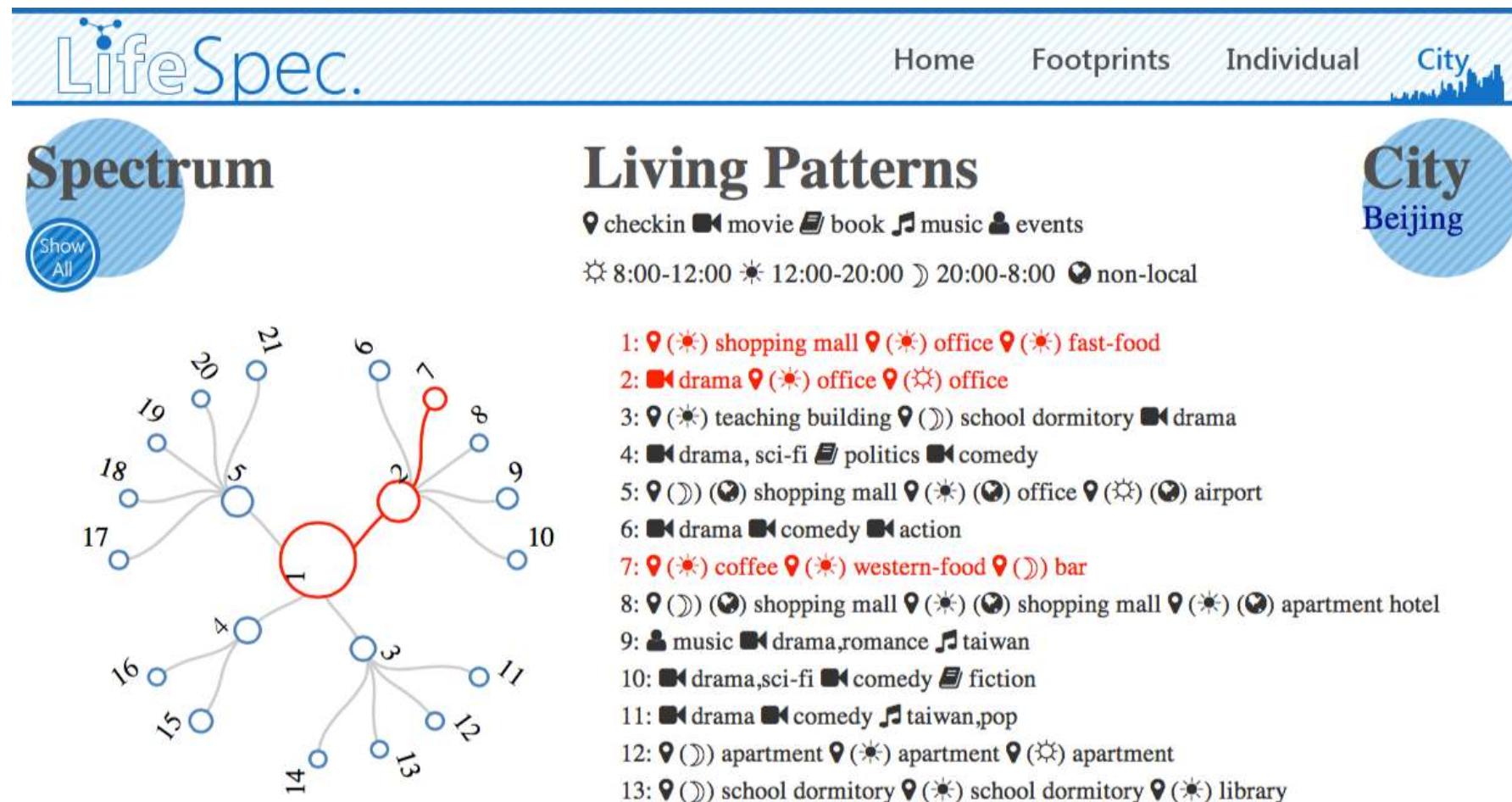
### Tips



# ICONNECT Algorithm



# Lifestyle Spectrum of a Group



# Relational Hierarchical LDA (RH-LDA)

📍 checkin 🎥 movie 📖 book 🎵 music 💬 events  
⌚ 8:00-12:00 ⚡ 12:00-20:00 ☺ 20:00-8:00 🌐 non-local

**footprint (word):** combination of domain specific tags (category)

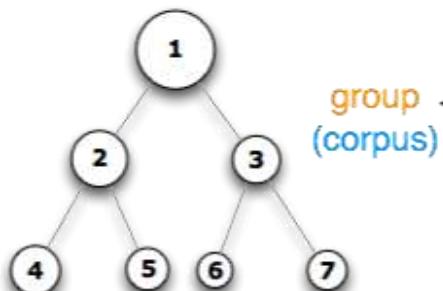
📍 (⌚) shopping mall 🎥 drama, sci-fi 🎵 taiwan,pop 💬 lecture

**living pattern (topic):** frequently co-occurring footprints

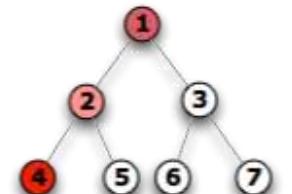
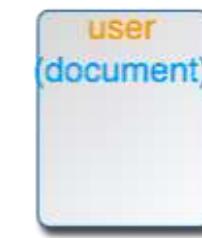
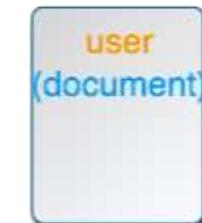
📍 (⌚) shopping mall + 🎵 taiwan,pop + 📍 (☺) bar

**lifestyle spectrum:** tree-structured topic hierarchy

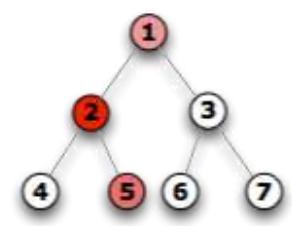
**lifestyle spectrum  
(topic hierarchy)**



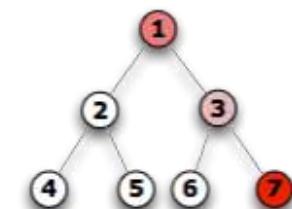
nodes: living patterns  
higher levels: commonalities  
lower levels: variations



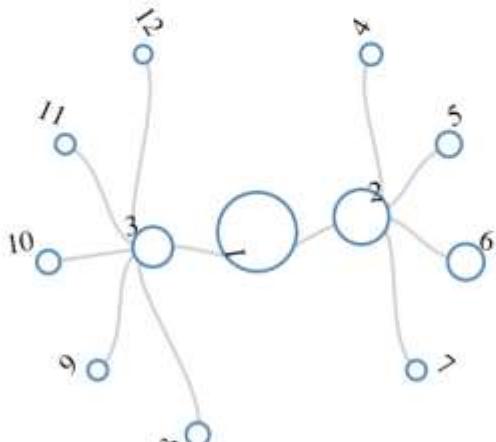
**lifestyle: 1-2-4  
(topic sequence)**



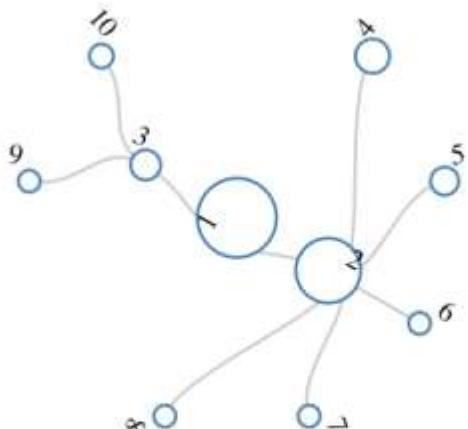
**lifestyle: 1-2-5  
(topic sequence)**



**lifestyle: 1-3-7  
(topic sequence)**



c) Tsinghua students/graduates



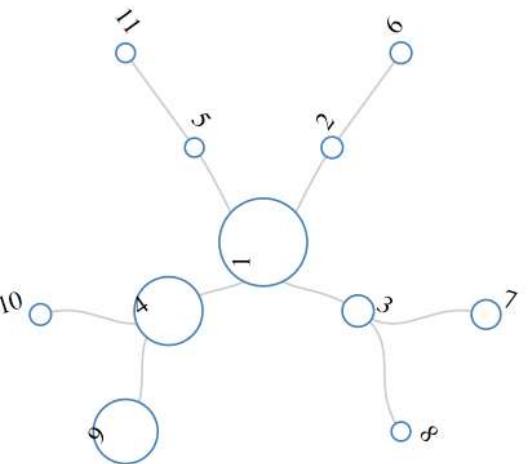
d) BFA students/graduates

- 1: 🎬 drama,romance 🎬 drama,comedy 🎬 drama,action
- 2: 🍔 (●) school canteen 🍔 (●) snack 🍔 (●) train station
- 3: 🏢 (●) office 🏢 (●) apartment 🏢 (●) office
- 4: 🎵 taiwan,pop 🎬 action,sci-fi 🎤 movie
- 5: 🏢 (●) (●) airport 🏢 (●) (●) apartment hotel 🏢 (●) (●) apartment hotel
- 6: 📚 (●) library 📚 (●) school canteen 📚 (●) teaching building
- 7: 🎵 japan,jpop 🎬 mystery,japan 🎵 jpop,japan
- 8: 🎬 drama,romance 🎵 pop,western 🎤 exhibition
- 9: 📖 history,chinesehistory 🎬 mystery,japan 🎬 action,sci-fi
- 10: 🍔 (●) fast-food 🍔 (●) apartment hotel 🏢 (●) institute
- 11: 🎤 music 🎬 investment,finance 🎤 get-together
- 12: 🎵 ost,japan 🎬 cartoon 🎵 folk,inland

c') Tsinghua students/graduates

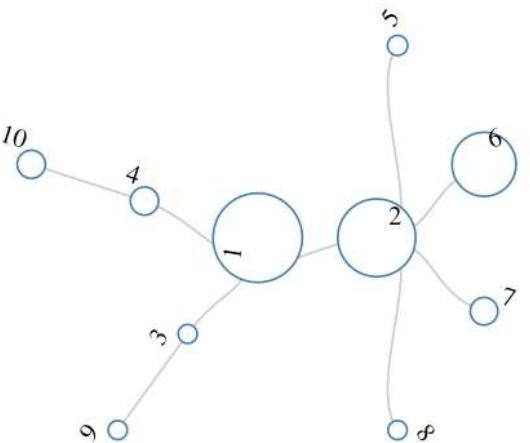
- 1: 🍔 (●) coffee 🍔 (●) western-food 🍔 (●) bar
- 2: 🎬 drama,romance 🎬 drama,comedy 🎵 taiwan,indie
- 3: 🎬 drama,romance 🎬 drama,comedy 🎬 drama,action
- 4: 🎤 music 🎤 movie 🎤 get-together
- 5: 🎬 drama,action 🎬 action,sci-fi 🎬 action,thriller
- 6: 🎵 britpop,uk 🏢 (●) institute 🎵 chineserock,inland
- 7: 📖 fiction,romantic 🎵 jazz,western 🎬 japaneseliterature,japan
- 8: 🎵 folk,inland 🎵 chineserock,rock 🎵 taiwan,pop
- 9: 🏢 (●) freeway 🏢 (●) private place 🏢 (●) freeway
- 10: 🎤 movie 🎬 comedy,romance 🎵 pop,western

d') BFA students/graduates



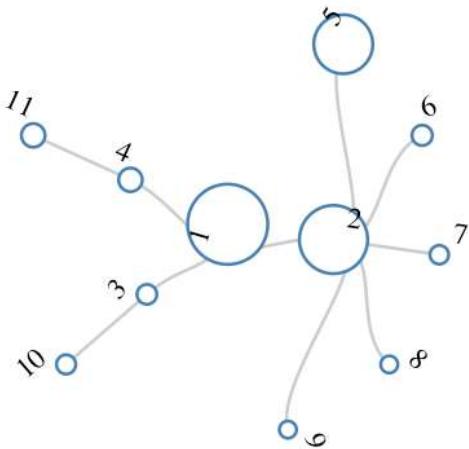
- 1: 📍 (●) office 🎯 (◎) ktv 📍 (◎) office
- 2: 🎵 britpop,indie 🎵 electronic,indie 🎵 japan,jpop
- 3: 📍 (●) (◎) airport 📖 internet,ux 📖 ux,internet
- 4: 🎬 cartoon 📖 programing,computer 📖 algorithm,computer
- 5: 📖 classical literature 📖 sinology 📍 (●) (◎) institute
- 6: 🎵 taiwan,pop 🎵 hongkong,pop 🕸️ music
- 7: 🕸️ music 🕸️ lecture 🕸️ movie
- 8: 🎵 ost,japan 🎵 classical 🎵 classical,piano
- 9: 🎬 drama,comedy 🎬 drama,romance 🎬 drama,action
- 10: 📍 (◎) (◎) apartment 📍 (●) (◎) airport 📍 (◎) private place
- 11: 🎵 taiwan,pop 🎬 drama,romance 📖 fiction,chineseliterature

c') Tencent employees

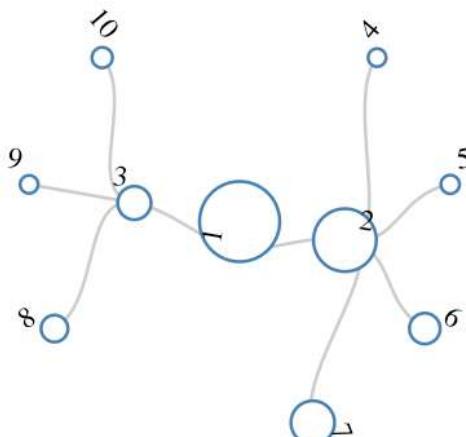


- 1: 📖 computer 📍 (●) coffee 🎵 classical
- 2: 🎵 taiwan,pop 🕸️ music 🎬 drama,romance
- 3: 📖 textbook,politics 📖 thinking,essay 📖 poem,essay
- 4: 🎵 rock 🎵 chineserock 🎵 britpop,uk
- 5: 📍 (●) (◎) office 📍 (●) (◎) jiangzhe cuisine 📍 (●) (◎) shopping mall
- 6: 🎬 drama,action 🎬 drama,comedy 🎬 drama,romance
- 7: 📍 (●) office 🎬 drama,action 📍 (●) western-food
- 8: 📍 (◎) western-food 🎵 hongkong,pop 📍 (●) fashion shop
- 9: 📖 ux,design 📖 history,fiction 📖 design
- 10: 🕸️ lecture 🕸️ music 📖 mystery,japan

d') Baidu employees



a) financial practitioners



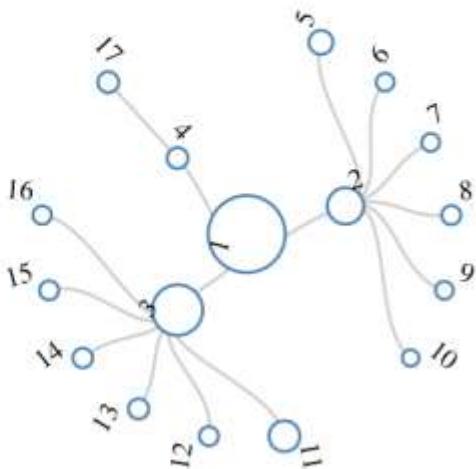
b) software practitioners

- 1: 📖 economics 🏠 (⌚) (🏢) apartment hotel 🏠 (⌚) (🏢) shopping mall  
 2: 🏠 (⌚) japanese cuisine 🏠 (⌚) fast-food 🕸️ lecture  
 3: 🏠 (⌚) hot-pot 🏠 (⌚) bar 🏠 (⌚) snack  
 4: 🏠 (⌚) snack 🏠 (⌚) fast-food 🏠 (⌚) japanese cuisine  
 5: 🎬 drama,romance 🎬 drama,comedy 🎬 drama,action  
 6: 🏠 (⌚) bank 🏠 (⌚) bank 🏠 (⌚) subway  
 7: 📖 fiction,hongkong 📖 fiction,love 📖 mystery,japan  
 8: 🎵 folk,indie 🎵 indie,folk 🎬 drama  
 9: 🏠 (⌚) car-4s 📖 fiction,society 📖 cartoon,philosophy  
 10: 🕸️ music 🎬 drama,romance 🎬 drama,comedy  
 11: 🏠 (⌚) (🏢) scenic 🏠 (⌚) (🏢) airport 🏠 (⌚) (🏢) office

a') financial practitioners

- 1: 📖 computer 📖 programing,computer 🕸️ movie  
 2: 🎬 drama,romance 🎬 drama,comedy 🎵 taiwan,pop  
 3: 📖 ux,design 📖 fiction,foreignliterature 📖 fiction,chineseliterature  
 4: 📖 mystery,japan 🎬 comedy,action 📖 cartoon,mystery  
 5: 🎵 taiwan,pop 🕸️ music 🎵 chineserock,rock  
 6: 🏠 (⌚) apartment 🏠 (⌚) office 🏠 (⌚) (🏢) apartment  
 7: 🎬 drama,romance 🎬 drama,action 🎬 drama,comedy  
 8: 🕸️ lecture 🕸️ music 🕸️ get-together  
 9: 📖 programing,computer 📖 algorithm,computer 🎬 drama,suspense  
 10: 🎬 drama,romance 🕸️ music 🎵 taiwan,pop

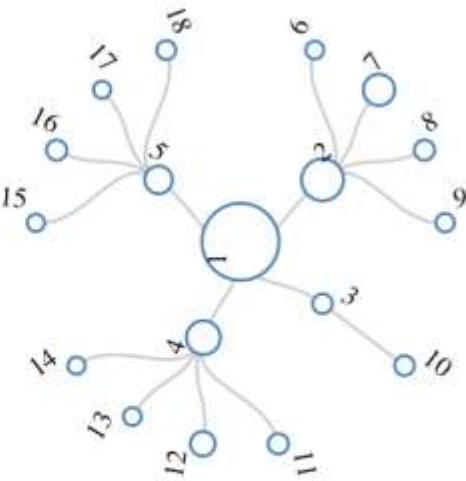
b') software practitioners



c) the post-90s generation

- 1: 🎬 drama,romance 🎬 drama,comedy 🎬 comedy,romance  
 2: 📍 (●) coffee 📍 (●) (●) train station 📍 (○) fast-food  
 3: 🚻 (●) institute 🚻 (●) coffee 🚻 (●) institute  
 4: 📖 hongkong,essay 🚻 (●) (●) coffee 🚻 (●) electric game  
 11: 🚻 (●) subway 🚻 (●) shopping mall 🚻 (●) fast-food  
 12: 🚻 (●) teaching building 🚻 (●) teaching building 💬 music  
 13: 💬 exhibition 📖 fiction,youth 💬 lecture  
 14: 💬 music 🎵 taiwan,indie 🎵 folk,inland  
 15: 🎵 korea,kpop 🎵 taiwan,pop 💬 music

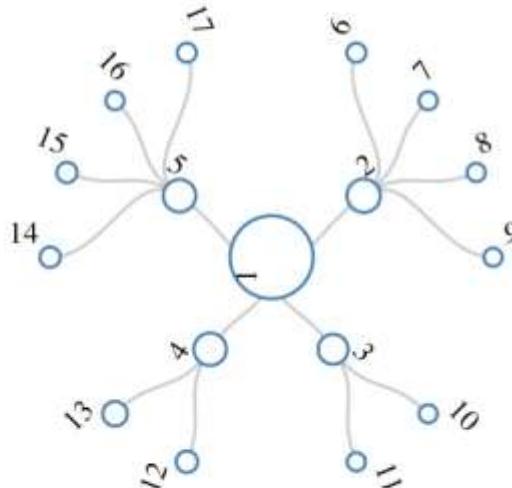
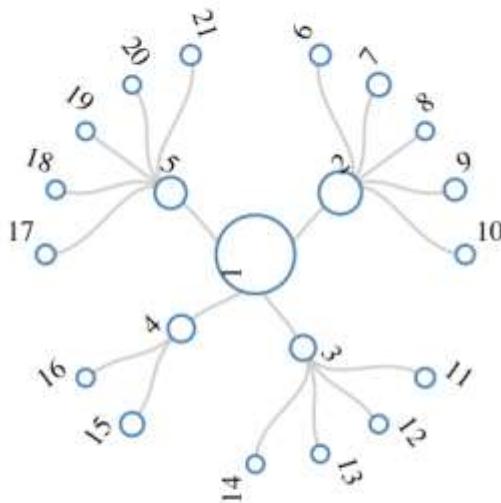
c') the post-90s generation



d) the post-80s generation

- 1: 🎬 drama,romance 🎬 drama,comedy 🎬 drama,action  
 2: 🚻 (●) shopping mall 🚻 (●) hot-pot 🚻 (●) japanese cuisine  
 3: 🚻 (○) shopping mall 🚻 (●) office 🚻 (●) snack  
 4: 💬 get-together 💬 lecture 💬 travel  
 5: 🎬 action,thriller 🚻 (○) apartment 🎬 cartoon,suspense  
 6: 🚻 (○) ktv 🚻 (●) office 🚻 (●) coffee  
 7: 🚻 (●) office 🚻 (●) sichuan cuisine 🚻 (●) office  
 8: 🚻 (○) apartment 🚻 (●) commercial building 🚻 (●) commercial building  
 9: 📖 picturebook,taiwan 📖 picturebook,cartoon 📖 fiction,uk

d') the post-80s generation



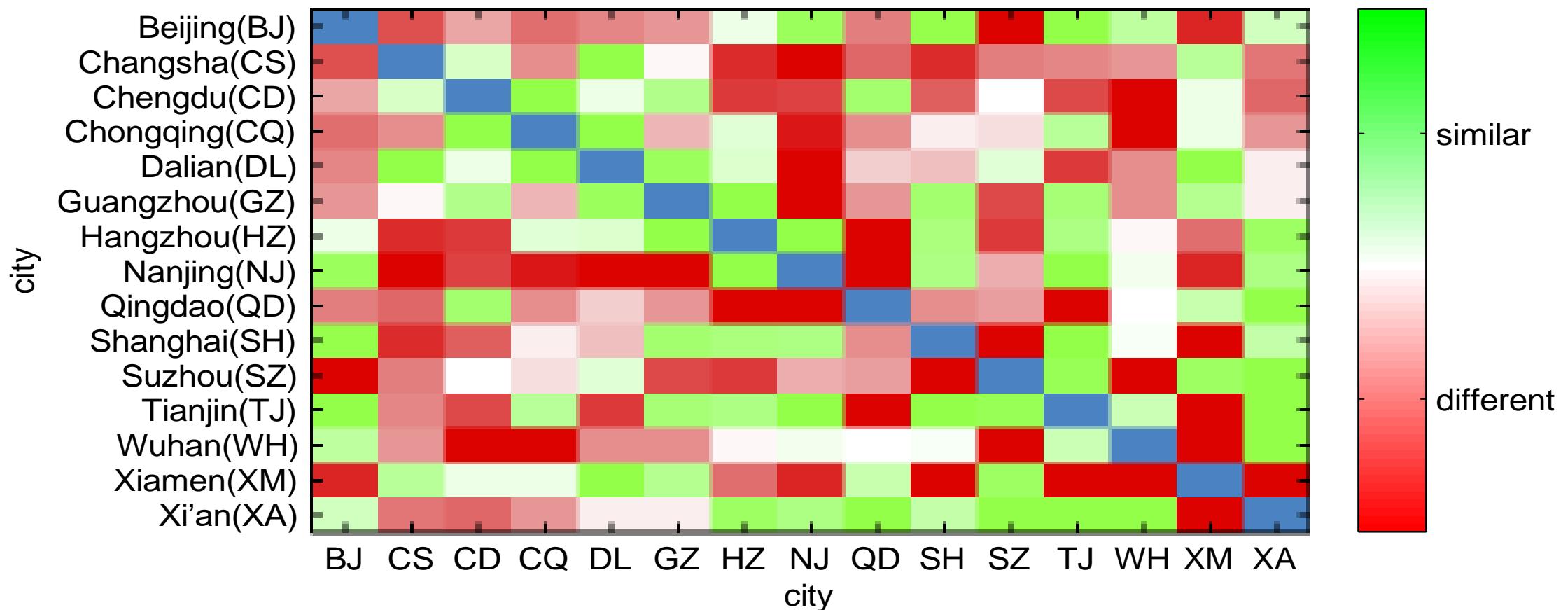
- 1: shopping mall office fast-food
- 2: drama office office
- 3: teaching building school dormitory drama
- 4: drama, sci-fi politics comedy
- 5: shopping mall office airport
- 6: drama comedy action
- 7: coffee western-food bar
- 8: shopping mall shopping mall apartment hotel
- 9: music drama,romance taiwan

a') Beijing

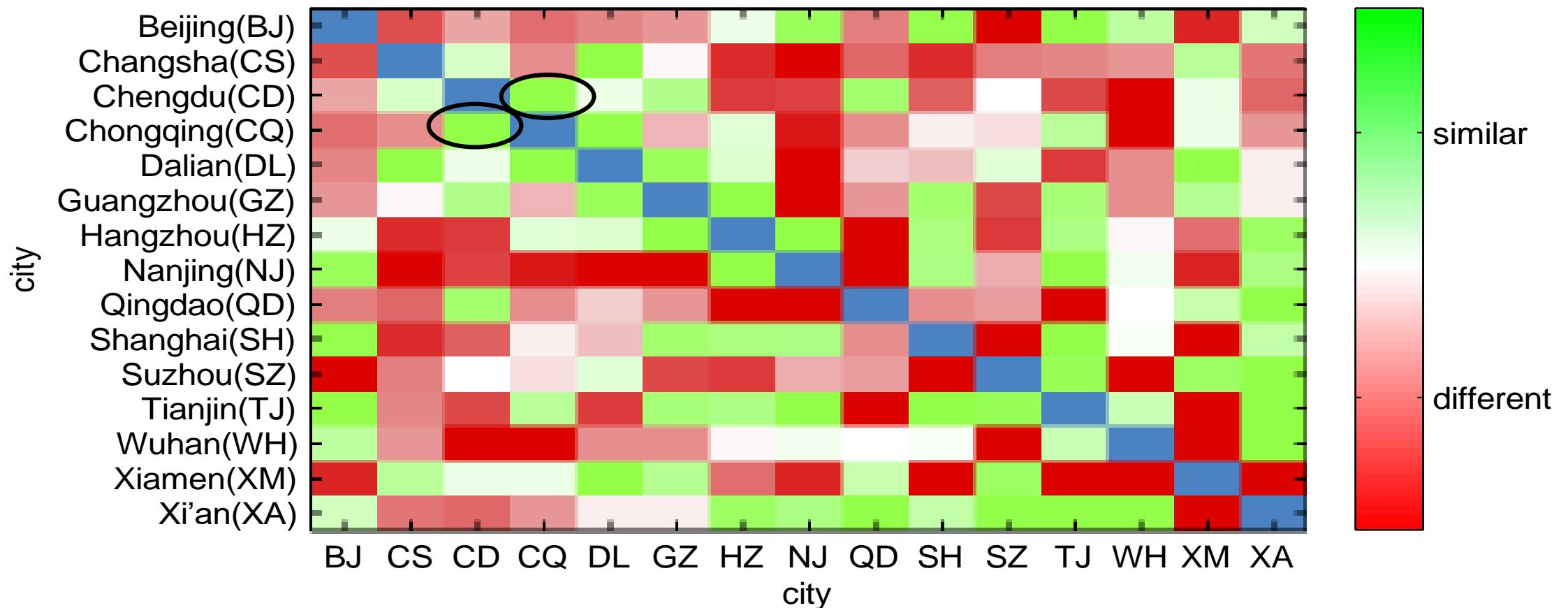
- 1: jiangzhe cuisine fast-food coffee
- 2: drama,comedy coffee western-food
- 3: train station shopping mall outdoors
- 4: drama fiction taiwan
- 5: shopping mall snack coffee
- 6: music get-together drama,romance
- 7: office supermarket fast-food
- 8: drama,action history fiction
- 9: library library scenic

b') Hangzhou

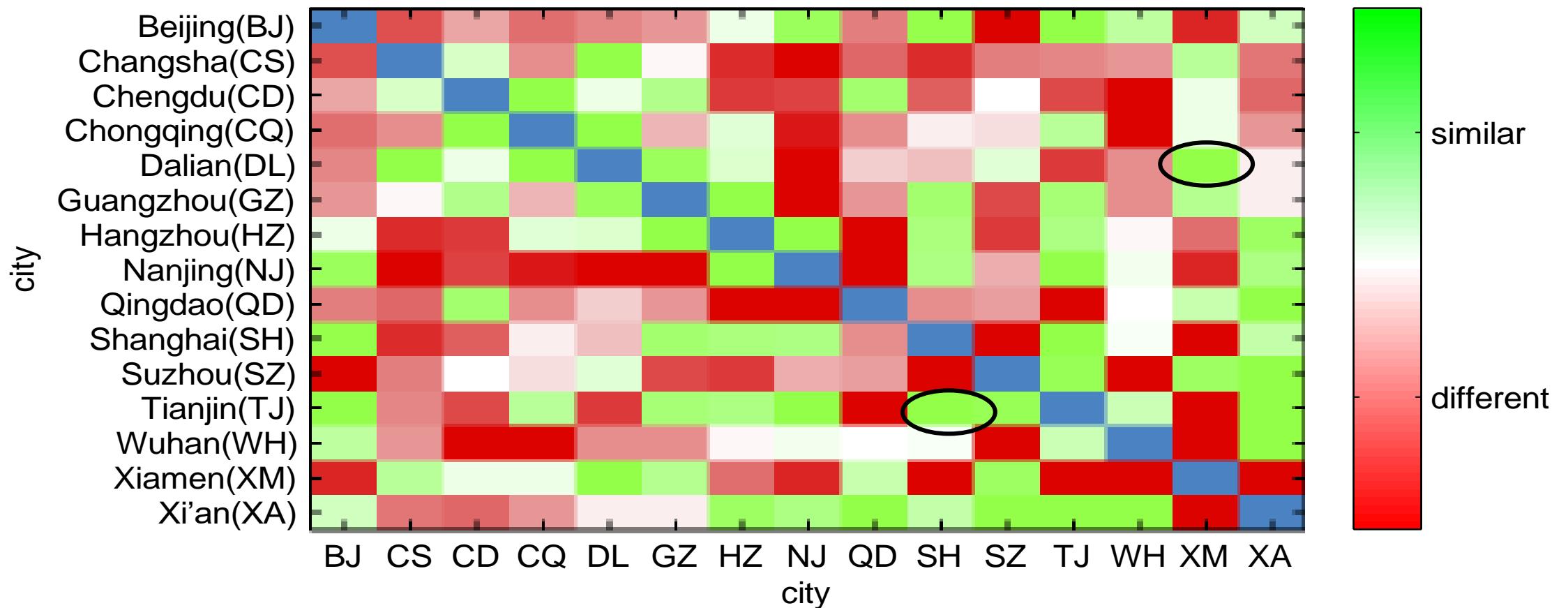
# Similarity Matrix of Lifestyle Spectrum



# Similarity Matrix of Lifestyle Spectrum

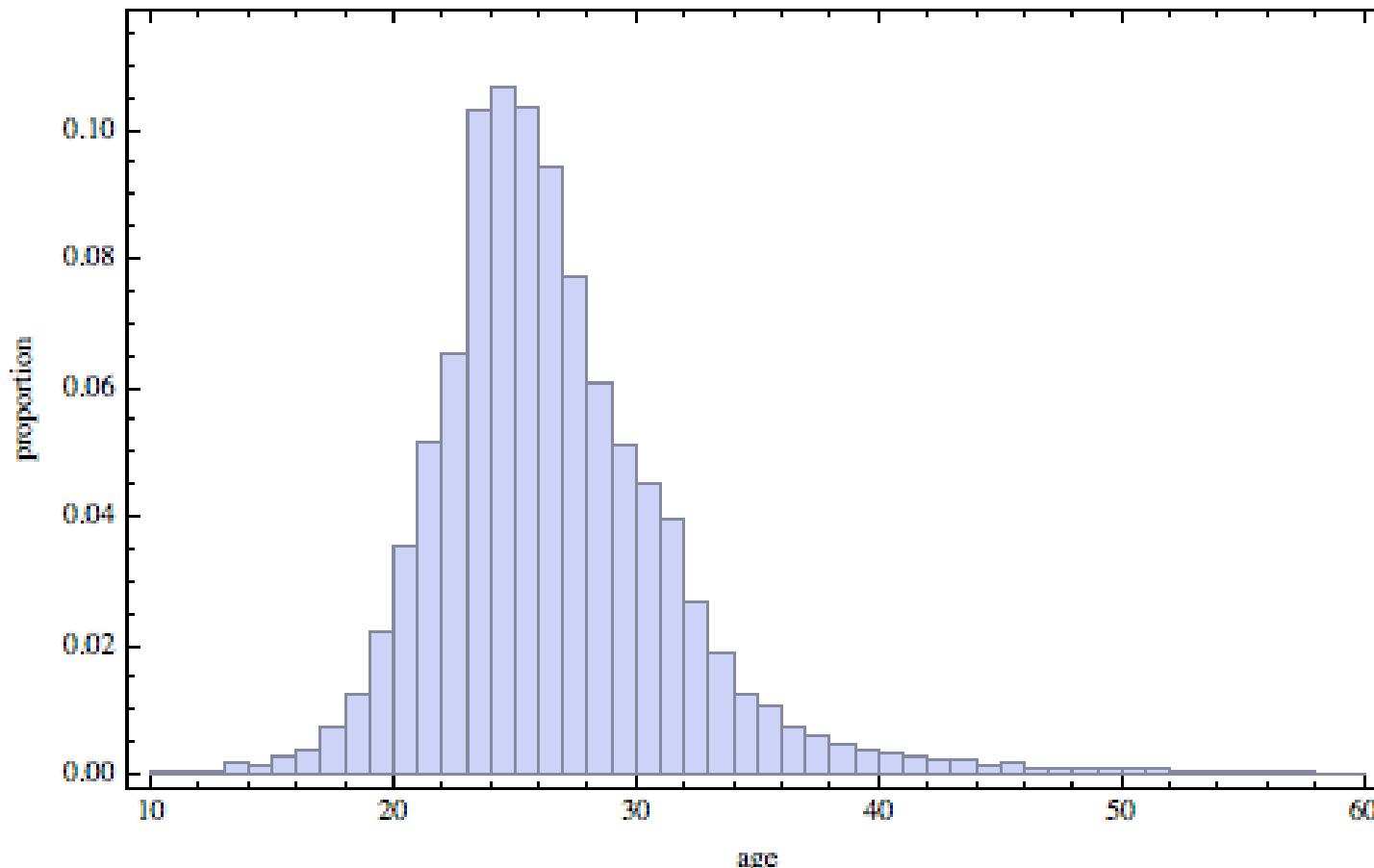


# Similarity Matrix of Lifestyle Spectrum



# Limitations

- Data coverage



# LifeSpec as A Data Repository

- Industry Applications

- Recommendation
- Advertising

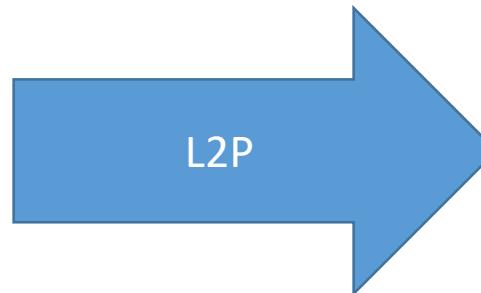
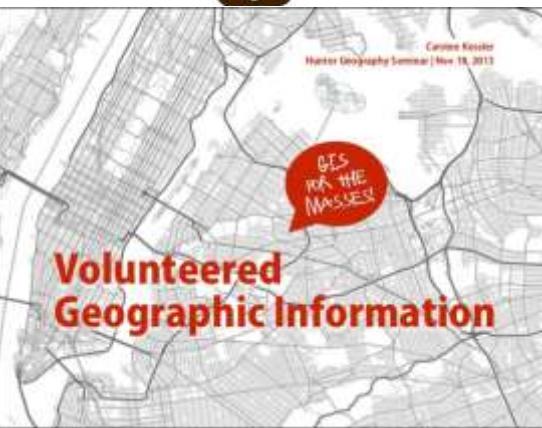
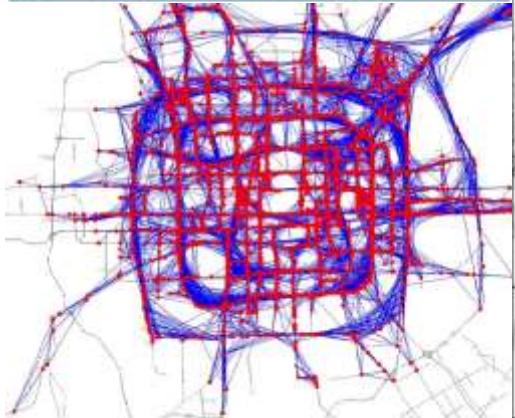
- Scientific Research

- Sociology
- Economics
- Public health

# Summary

- Computational framework for exploring urban lifestyles
- IConnect: identifying connected user accounts based on self-disclosure
- RH-LDA: relational-hierarchical model for summarizing lifestyles
- LifeSpec as a data platform for scientific and industry applications

# 基于用户位置数据的用户画像



★macy's

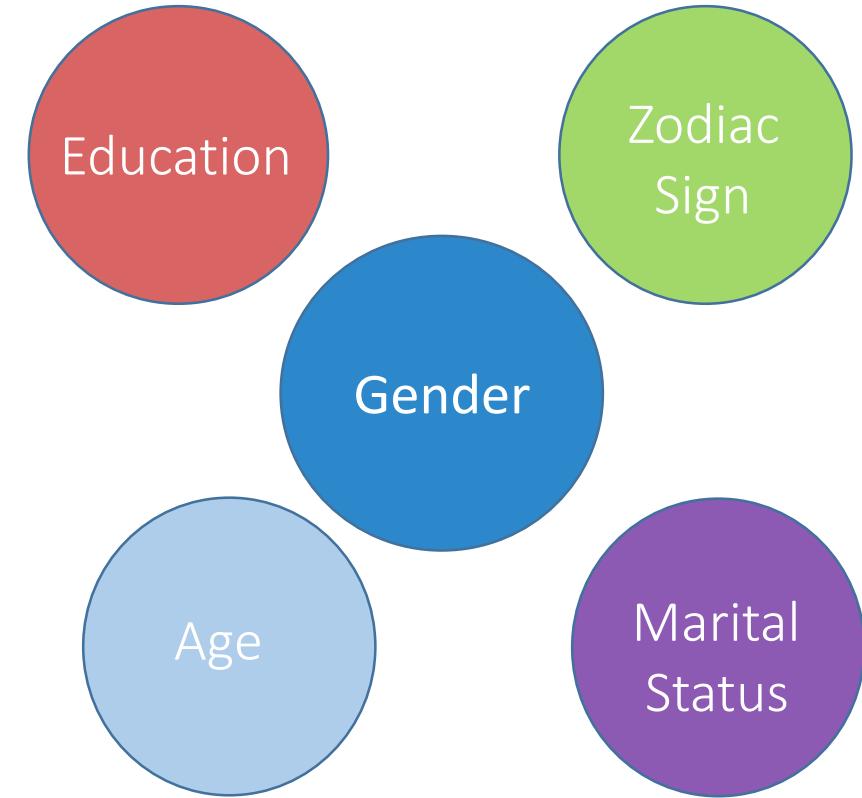
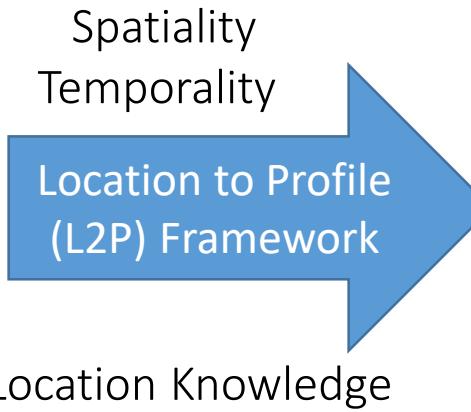
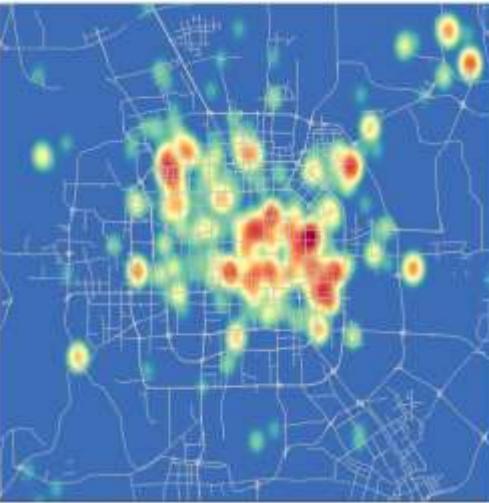
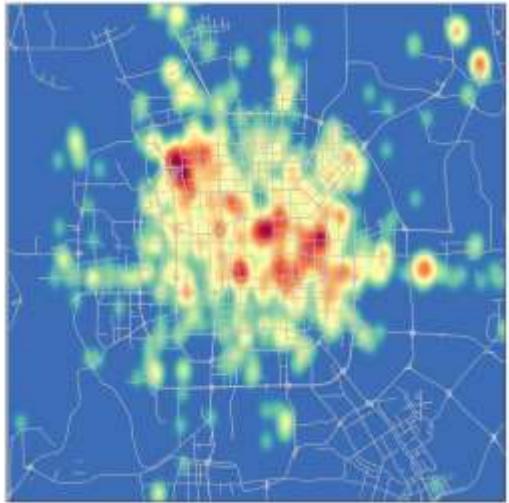


General mobility sequence: LBSN check-in, GPS trajectory, Cellular data, etc.

# Examples



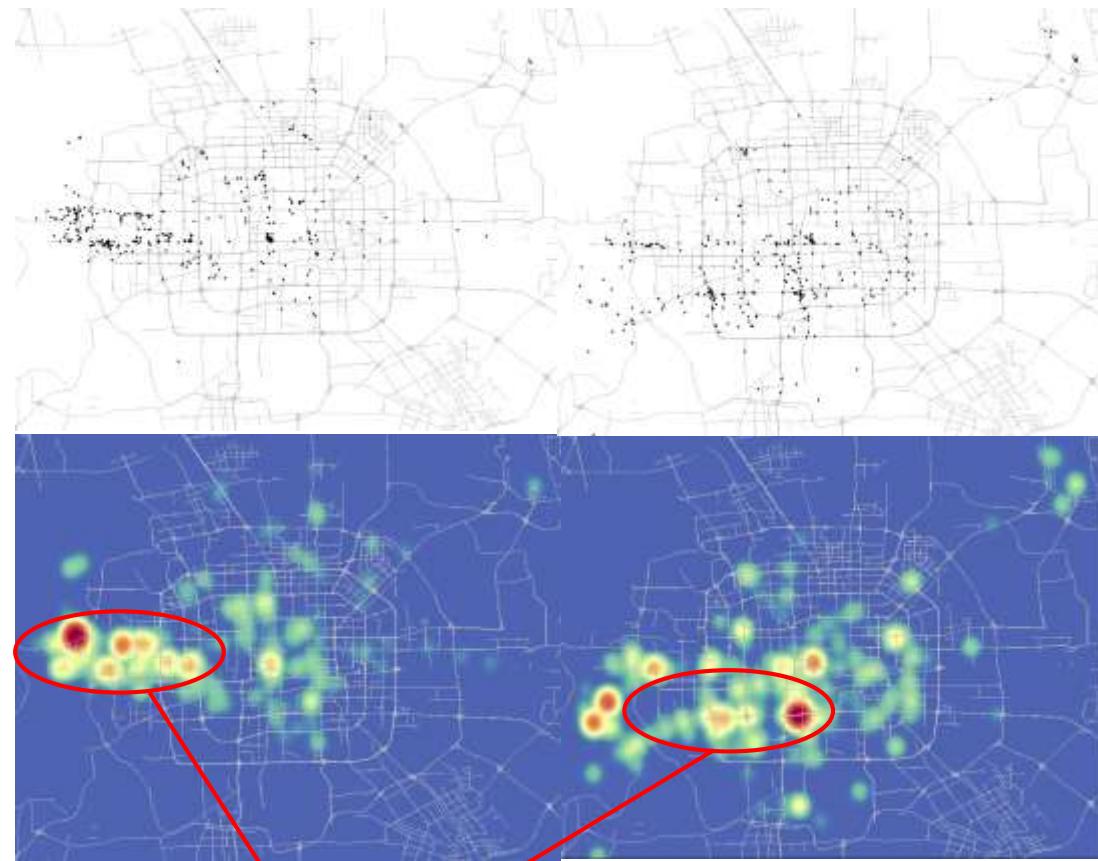
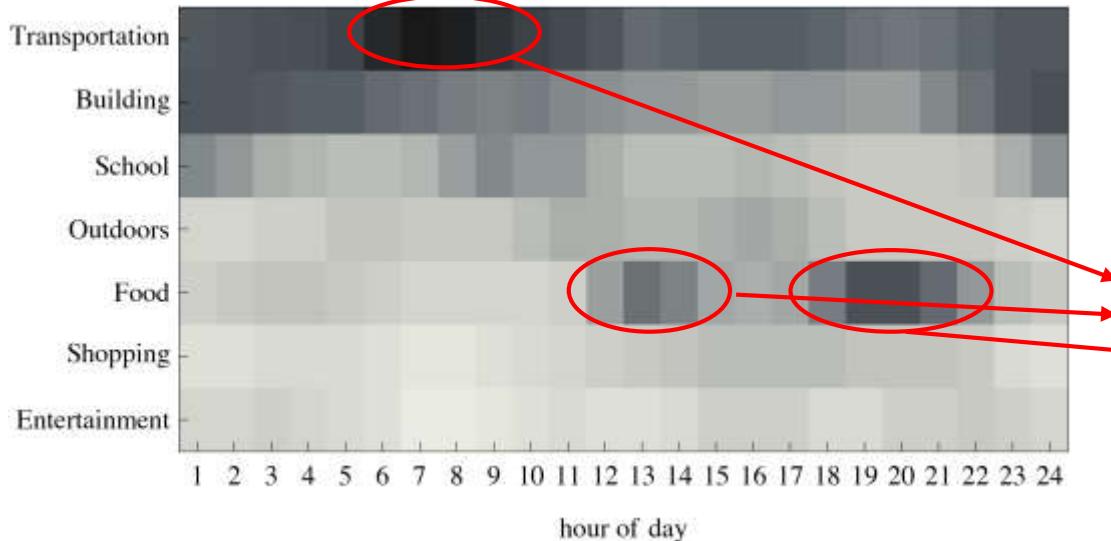
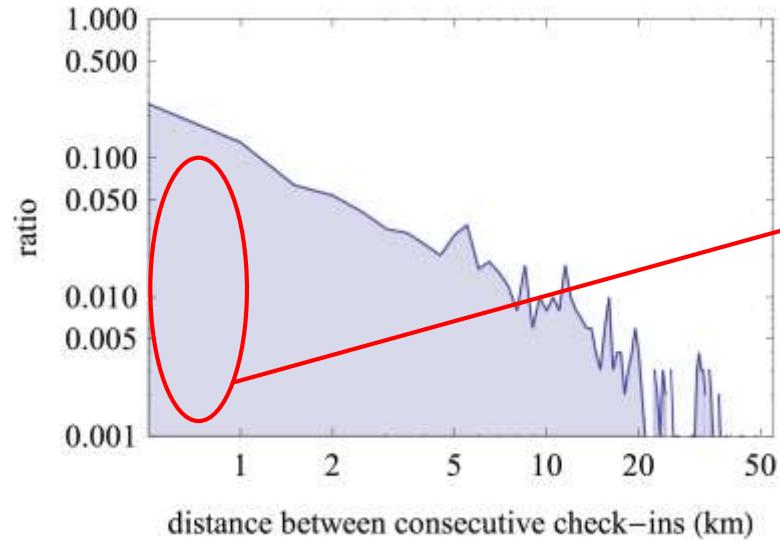
# Profile inference from location check-ins



Location check-ins of Beijing and Shanghai

Users' demographics

# Spatiality and temporality



Temporal & Spatial Regularity

M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453 (7196):779–782, 2008.

Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility[J]. *Science*, 2010, 327(5968): 1018-1021.

# Location knowledge

**POI**

Mexican Grill 2000

5 stars, 94 reviews

\$ Mexican

Address: 703 RT 112, Patchogue, NY 11772 (631) 207-6554 litrade.net/mexicangrill

Map data ©2014 Google

Chicken enchiladas with rice and black beans by Theresa A.

See all 33 photos

Write a Review Add Photo Share Bookmark

"Best, most authentic, and juicy pastor taco I've had on Long Island." in 12 reviews

"The shrimp burrito was literally the size of my forearm..and in this case, size matters." in 8 reviews

Today 12:00 pm - 9:00 pm Closed now

Menu

\$\$\$\$ Price range Under \$10

**Review**

Kathryn W. Long Island, NY Elite '14 60 friends 84 reviews

11/3/2014 AWESOME burritos. My favorite is the chicken fajita burrito packed with delicious grilled seasoned chicken, peppers onions and mushrooms. Priced reasonably, and tastes super fresh and authentic.

Was this review ...?

Useful 1 Funny 1 Cool 1

**Rating**

Eileen L. East Islip, NY 0 friends 9 reviews

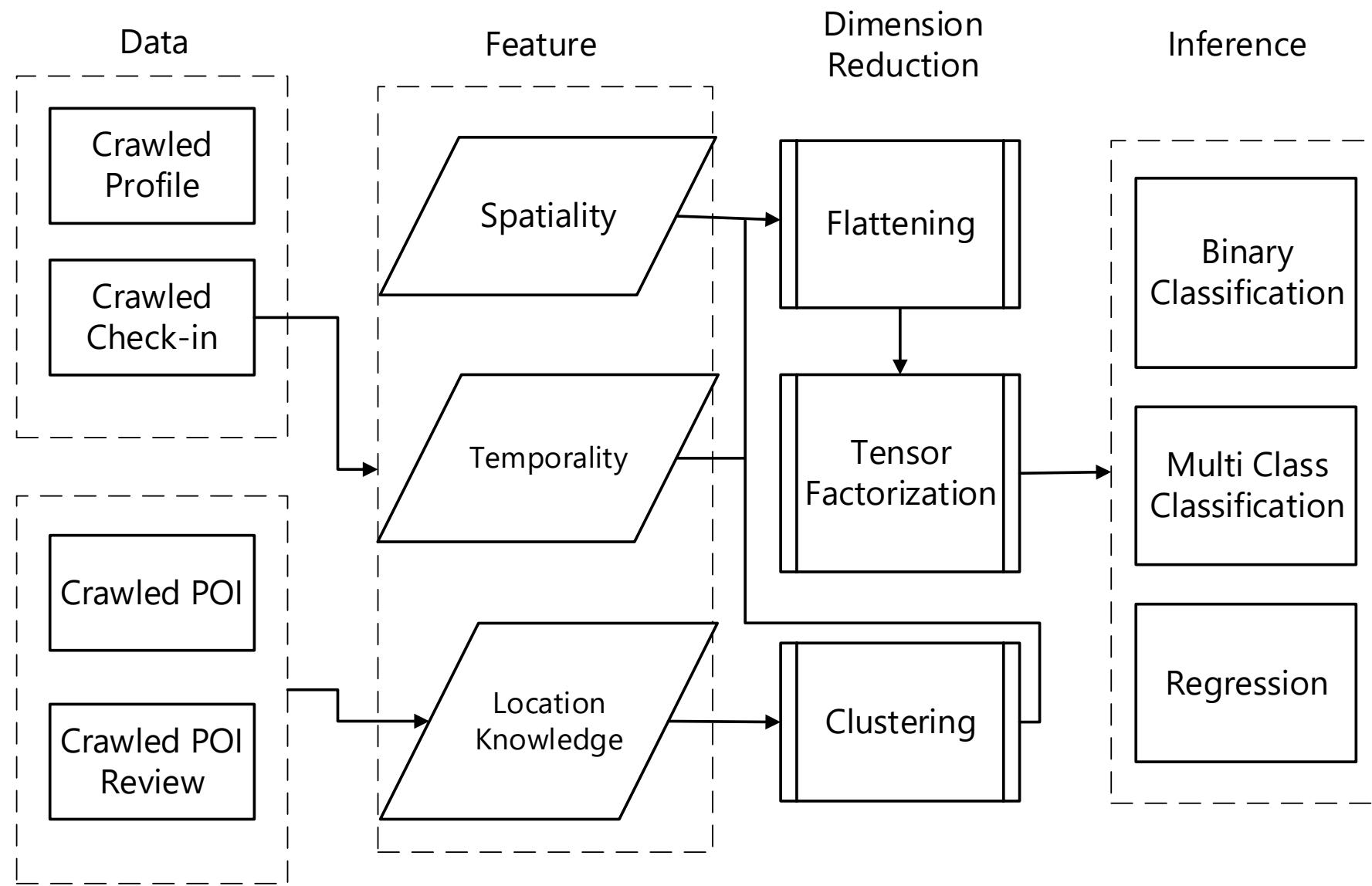
1/29/2014 The best Mexican food...real authentic. The place is certainly nothing to look at you can def roll out of bed and go to pick up or sit at one of the old tables. Trust me you may not like the way the place looks but the food is outstanding!!!!

Was this review ...?

Useful Funny Cool

- Location knowledge from user review site
- Including address, price, phone number, user review, rating, tag, etc

# Location to profile framework (L2P)



# Inference

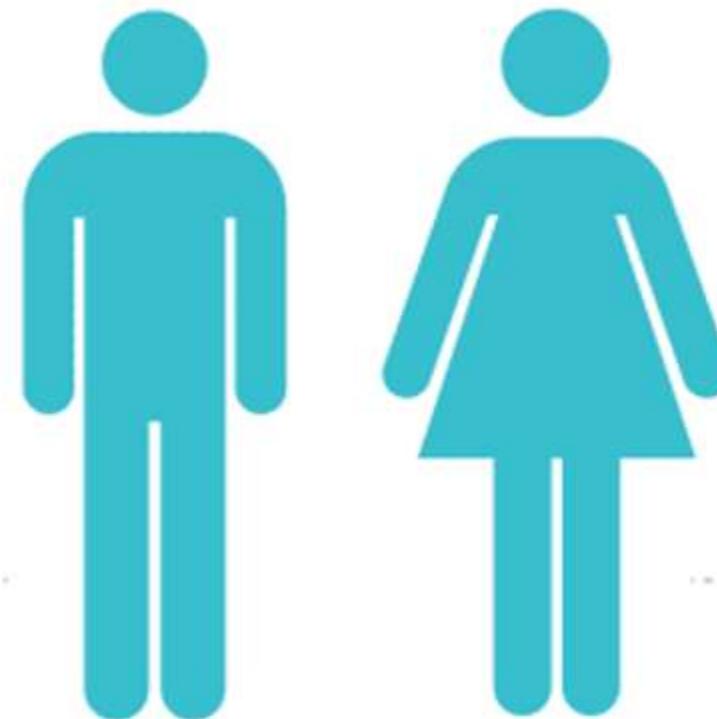
- Binary classification
- Multiclass classification
- ▲ Linear regression



● Gender



■ Marital status



Zodiac signs ■



Blood types ■



Education background ●



# Dataset and ground truth construction

- Focus on Beijing and Shanghai user check-ins
- 3,354,918 users' demographics and 81,781,544 location check-ins



Attribute	Completion rate	Categories
Gender	94.0193%	Male, Female
Age	33.1588%	The specific age number
Education background	36.7228%	University, Non-university
Sexual orientation	2.5549%	Heterosexuality, Bisexuality, Male homosexuality, Female homosexuality
Marital status	2.6396%	Single, Courtship (seeking a relationship), In love, Married
Blood type	1.6376%	O, A, B, AB
Zodiac sign	58.1649%	Twelve zodiac signs

user	Check-in time	Identity	POI name	POI category	POI latitude	POI longitude	POI location

# Experiments

- Baselines:
  - POI-based method (POI)
  - Spatiality-based method (S)
  - Spatiality and Temporality-based method (ST)
  - Spatiality, Temporality, and Category-based method (STC)
  - Spatiality, Temporality, and Location knowledge-based method (STL)

10-fold cross-validation: 6 parts for training, 2 parts for validation, 2 parts for testing

Measurements:  
Classification: precision, recall , F measure and AUC  
Regression: RMSE

Parameters:  
 $n = 200$  for top-n keywords  
 $N_R = N_K = 100, dU = 200$  (by default)

# Gender and age

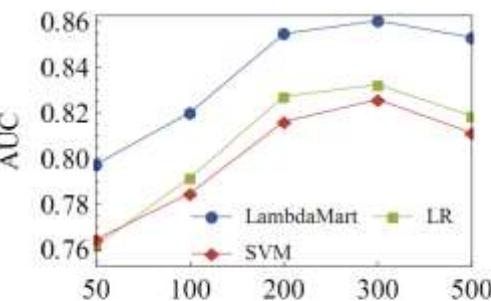
## Performance of Gender Inference

(a) Beijing

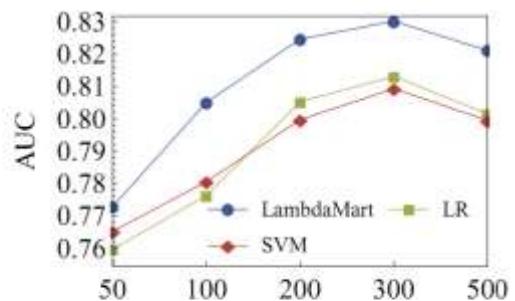
	Precision	Recall	F1	AUC
POI	0.7102	0.7055	0.7078	0.7502
S	0.6921	0.6899	0.6910	0.7321
ST	0.7321	0.7429	0.7375	0.7746
STC	0.7727	0.7631	0.7679	0.8027
STL	<b>0.8211</b>	<b>0.8059</b>	<b>0.8134</b>	<b>0.8548</b>

(b) Shanghai

	Precision	Recall	F1	AUC
POI	0.7362	0.7434	0.7398	0.7463
S	0.7197	0.7218	0.7207	0.7266
ST	0.7528	0.7596	0.7562	0.7682
STC	0.7819	0.7704	0.7761	0.8151
STL	<b>0.8368</b>	<b>0.8127</b>	<b>0.8246</b>	<b>0.8654</b>

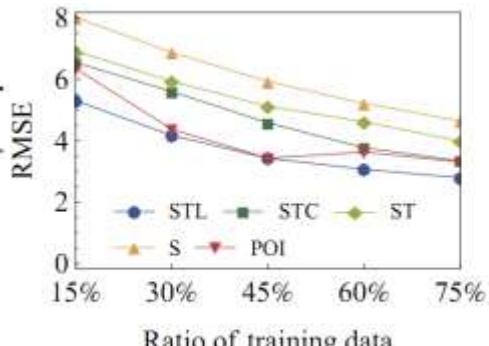


(a) Beijing

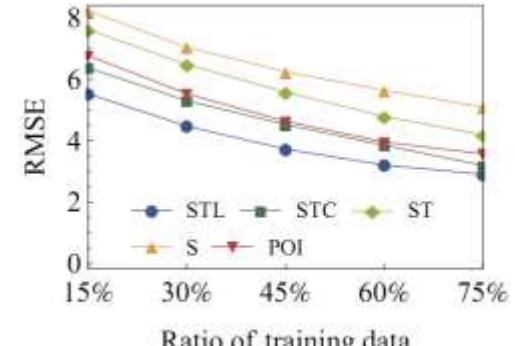


(b) Shanghai

Performance of different gender classifiers for STL changing over  $dU$



(a) Beijing



(b) Shanghai

RMSE of age w.r.t. ratio of training data

## Gender:

STL method achieves the best performance

LambdaMART classification function gains the best results

$dU = 300$  is the optimum stable value for tensor factorization

## Age:

The more training data there is, the better the prediction will be

# Education

**Table 5: Performance of Education Background Inference**

(a) Beijing

	Precision	Recall	F1	AUC
POI	0.7564	0.7702	0.7632	0.7992
S	0.7385	0.7294	0.7339	0.7723
ST	0.7655	0.7702	0.7678	0.8150
STC	0.8073	0.7921	0.7996	0.8413
STL	<b>0.8774</b>	<b>0.8829</b>	<b>0.8801</b>	<b>0.9021</b>

(b) Shanghai

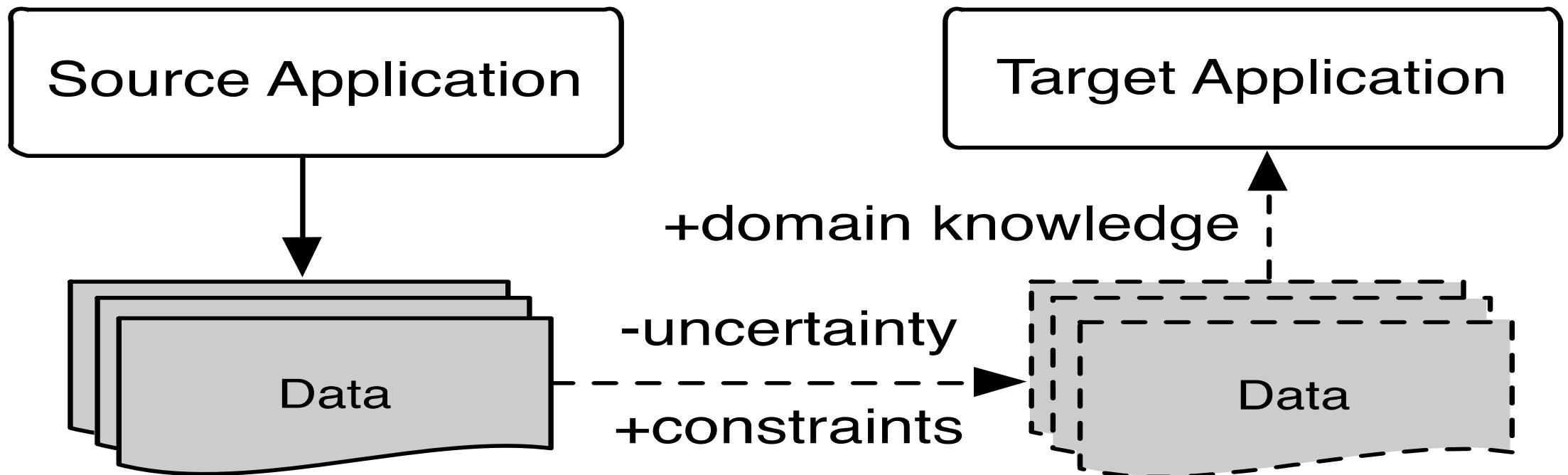
	Precision	Recall	F1	AUC
POI	0.7759	0.7817	0.7788	0.8205
S	0.7394	0.7191	0.7291	0.7659
ST	0.7804	0.7631	0.7717	0.8041
STC	0.8115	0.8058	0.8086	0.8503
STL	<b>0.8823</b>	<b>0.8726</b>	<b>0.8774</b>	<b>0.8957</b>

Inference results of education are higher than gender and age

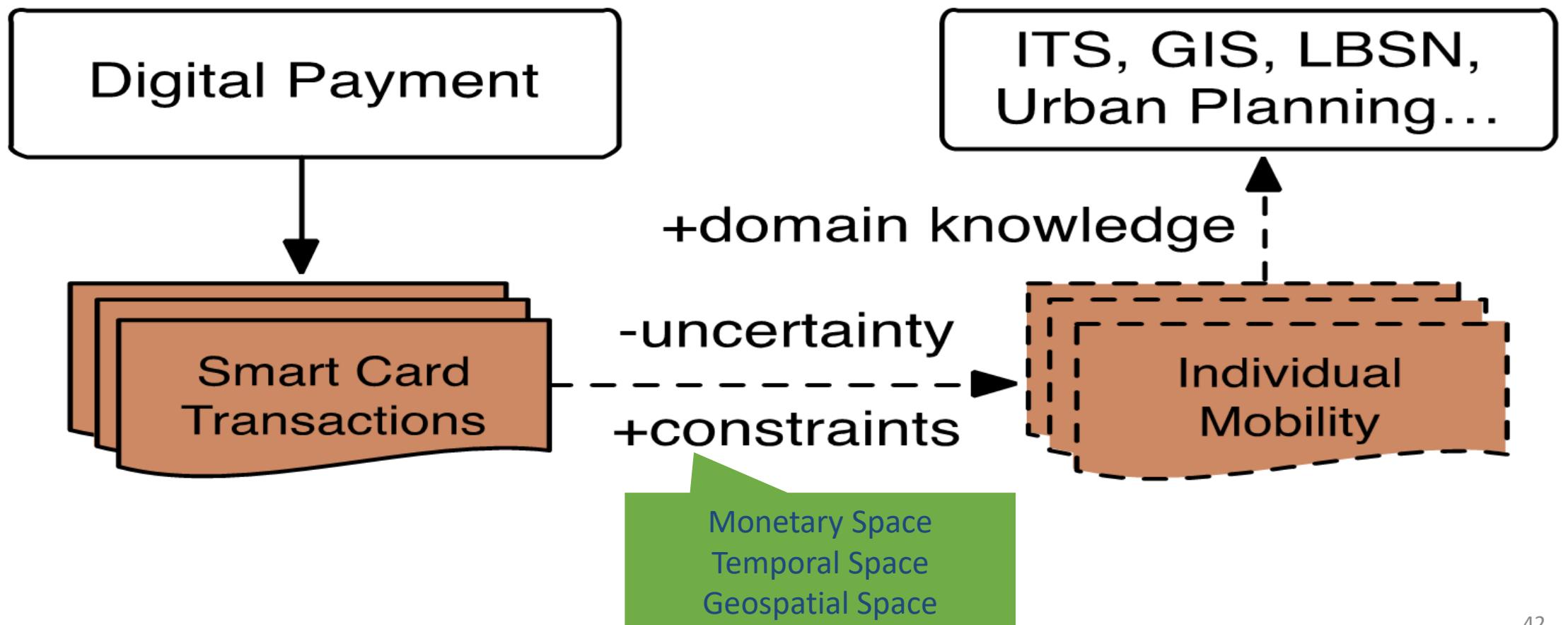
Users with different education background tend to have discrepant mobility patterns

# Reconstructing Individual Mobility from Smart Card Transactions

# Mining Cross-Application Data

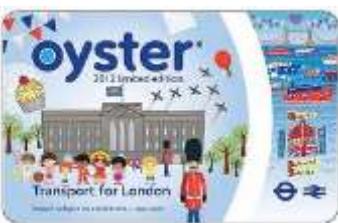


# Our Mission



# Smart Cards

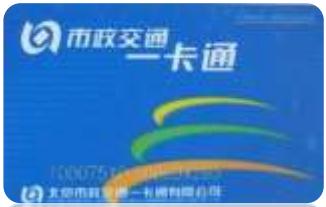
- London's Oyster Card



- San Francisco's Clipper Card



- Beijing's BMAC Card



# Related Work

- Mining Smart Card Data
  - Improve transportation service
  - Crowdedness analysis (London Oyster Card)
    - origin, start time are known for bus trips
    - origin, start time, destination, alighting time are known for subway trips
- Inferring alighting/boarding stops (at least one is known)
  - Trip-Chaining (TC): based on assumptions
    - Proximity between consecutive trips
    - *“The first trip of a day starts from the alighting station of last night”*
  - Trip-Chaining with maximum frequency (TC+MF)
  - Trip-Chaining with maximum similarity (TC+MS)

# Data I: Expense Records

- CardID (anonymized)
- Bus:
  - N: non-ladder fare
  - L: ladder fare
- Boarding & alighting stops
  - N: neither is known
  - L: codes indicating distance to stop with code 0
- Time
  - N: boarding time
  - L: alighting time
- Expense
  - N: fixed expense
  - L:  $e=a+b \times \max(|boarding-alighting|-c, 0)$
- Balance

Unknown stops (directions)!

(a) Expense Records

CardID	Bus	Boarding	Alighting	Time	Expense	Balance
1	N2	–	–	2013-03-14 09:02	0.8	12.3
2	L3	31	19	2013-03-14 17:45	0.4	32.2
3	N1	–	–	2013-03-15 08:45	0.4	10.6
3	L1	04	22	2013-03-16 18:20	0.8	49.8

# Data II: Charging Records and Road Networks

- Charging Records: 5.93 million

- CardID
- Time
- Amount
- Balance

- Road Network

- $G=(V,E)$
- $|V|=148110, |E|=196307$

(b) Charging Records

CardID	Time	Amount	Balance
3	2013-03-15 18:05	50.0	50.6
4	2013-03-15 18:05	20.0	21.6
5	2013-03-15 18:07	20.0	20.4
6	2013-03-15 18:08	30.0	40.8

# Data III: De-Noising and Labeling

- Bus Line Info: <http://api.amap.com> (autonavi)
  - Names and geo-coordinates of bus stops
  - Pricing information
- Data Labeling
  - 102 users
  - Free smart cards and reimbursement
  - Targeting most frequent ladder-fare lines
  - 4 months (Dec. 2012—Mar. 2013)

gender		age			
male	female	19-24	25-30	31-36	37-47
57.6%	42.4%	39.4%	45.5%	10.6%	4.5%

# Data De-Noising and Labeling

- Results
  - 124 ladder-fare lines
  - Covering 26% of all trips, 62% of all ladder trips

line type	#lines	ratio of records
lines without coordinates	95	4.16%
lines without price info	488	16.84%
non-ladder-fare	270	36.62%
labeled ladder-fare	124	26.54%
unlabeled ladder-fare	288	15.85%



Labeled ladder-fare bus lines on top of the road network

# Our More Concrete Mission

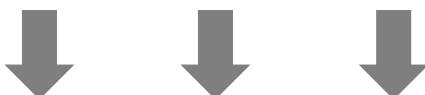
Charging  
Records

Road Networks

Partially Labeled  
Trips

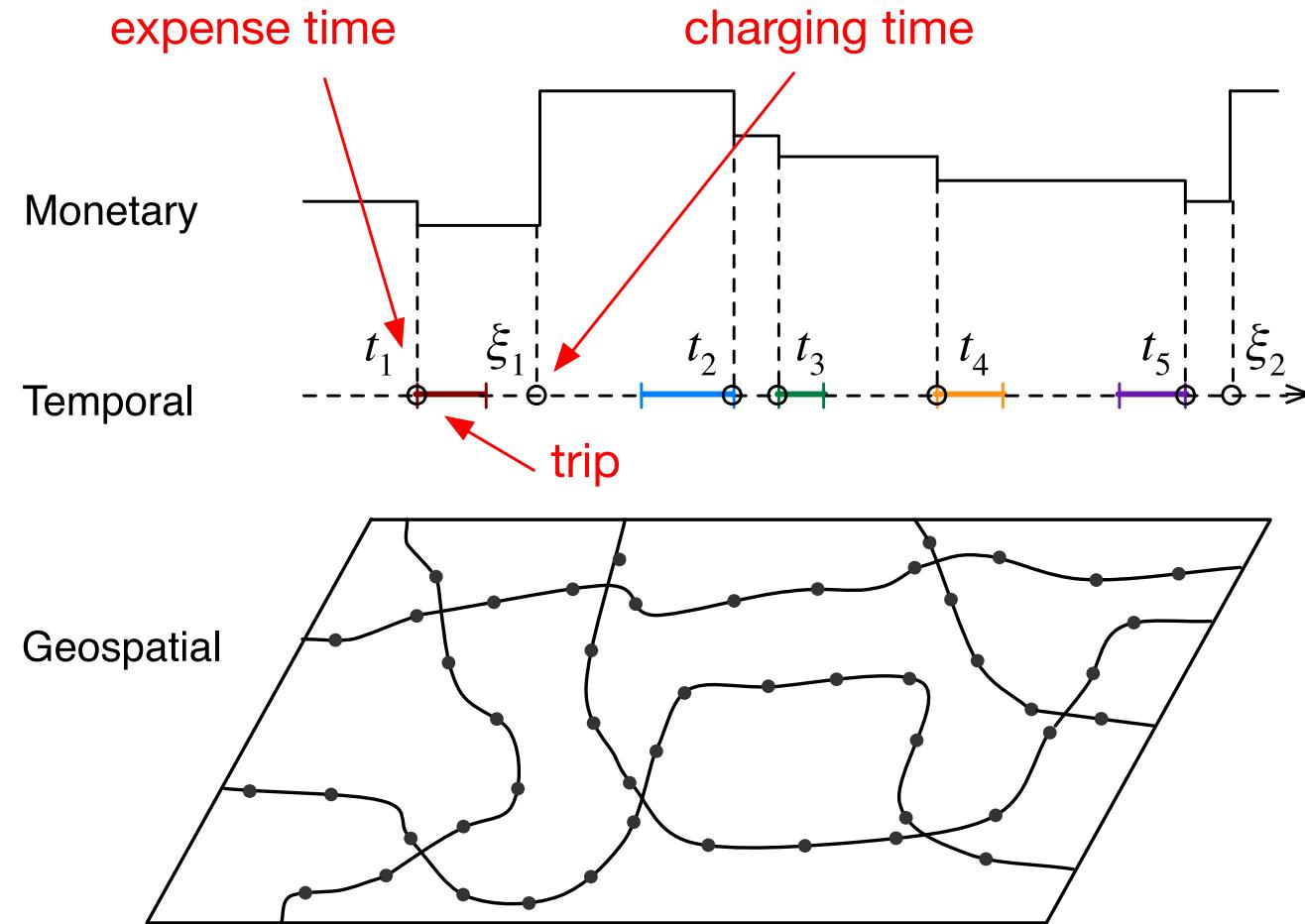
Bus Line Info

CardID	Bus	Boarding	Alighting	Time	Expense	Balance
1	<i>N2</i>	–	–	2013-03-14 09:02	0.8	12.3
2	<i>L3</i>	31	19	2013-03-14 17:45	0.4	32.2

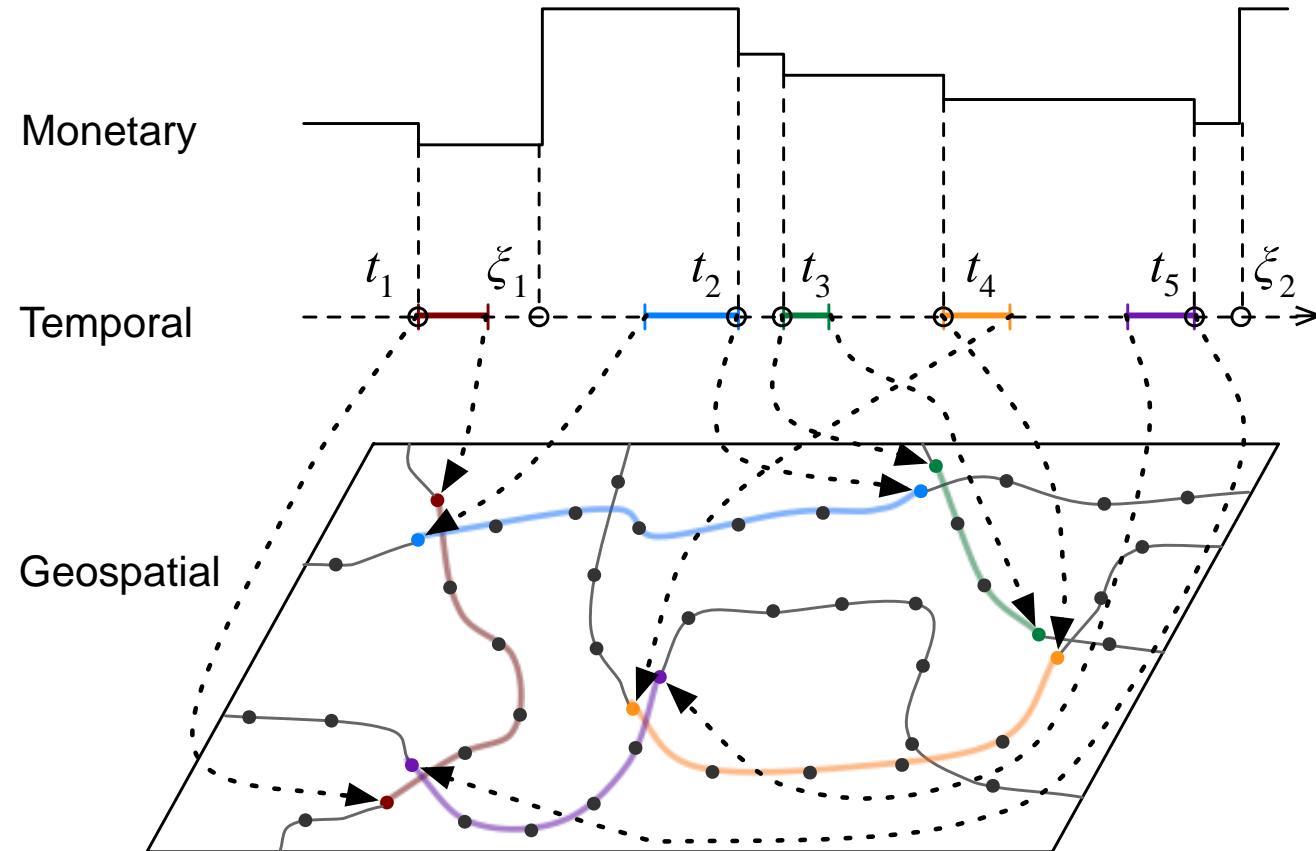


CardID	Bus	Boarding	Alighting	Time	Expense	Balance
1	<i>N2</i>	<i>lat1,lon1</i>	<i>lat2,lon3</i>	2013-03-14 09:02	0.8	12.3
2	<i>L3</i>	<i>lat3,lon3</i>	<i>lat4,lon4</i>	2013-03-14 17:45	0.4	32.2

# Space Alignment Framework



# Space Alignment Framework



# Segmentation

- Segment
  - “continuous” in the monetary space
  - other payments not included in the data (taxi, subway, shopping, etc.)
  - obtained in linear time cost
- $b_i + e_i = b_{i-1} + c_{i-1}$ 
  - $b_i$  :balance after the  $i_{th}$  trip
  - $e_i$  :expense of the  $i_{th}$  trip
  - $c_i$  :charged amount during the  $i_{th}$  and the  $(i+1)_{th}$  trip

---

## Algorithm 1: Segmentation

---

**Input:** CardId  $d$ , expense records  $\mathbb{E}$ , and charging records  $\mathbb{C}$   
**Output:** Segments  $\mathbf{S}$

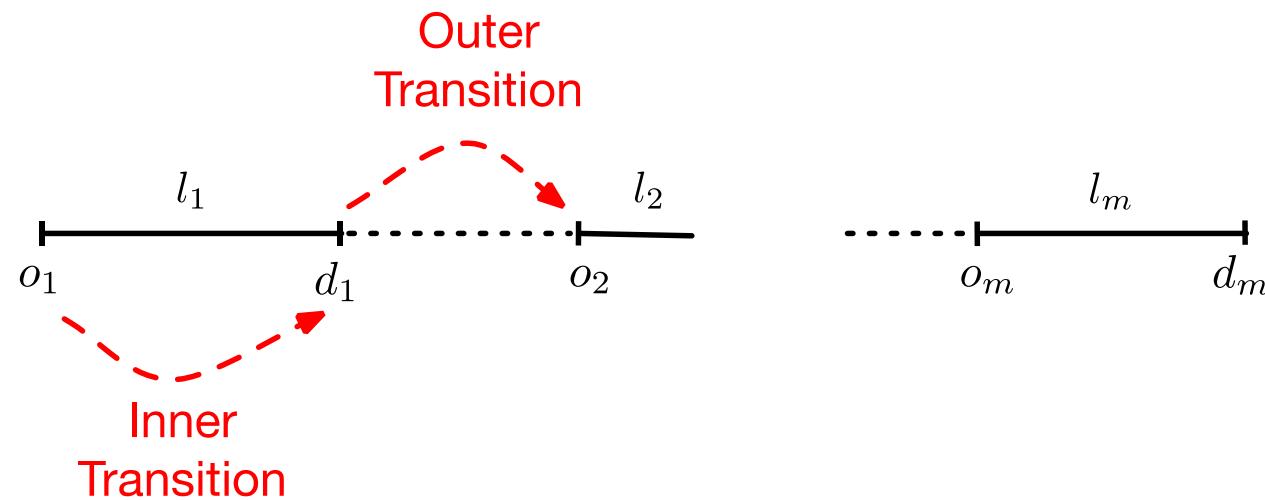
```
1  $\mathbf{I} \leftarrow \{1\};$  /*  $\mathbf{I} = \{I_i\}_{i=1}^{|\mathbf{I}|}$  is the index of split points */
2  $\mathbb{E} \leftarrow \text{select } * \text{ from } \mathbb{E} \text{ where CardID} = d \text{ order by time;}$  /*  $\mathbb{E} = \{E_i\}_{i=1}^{|\mathbb{E}|}$  */
3  $\mathbb{C} \leftarrow \text{select } * \text{ from } \mathbb{C} \text{ where CardID} = d \text{ order by time;}$  /*  $\mathbb{C} = \{C_j\}_{j=1}^{|\mathbb{C}|}$  */
4  $c_i \leftarrow 0, i = 1, 2, \dots, |\mathbb{E}|;$ 
5  $i \leftarrow 1, j \leftarrow 1;$ 
6 while  $i \leq |\mathbb{E}| - 1$  do
7   if  $j \leq |\mathbb{C}| \text{ and } t_1 < \xi_j < t_i$  then
8      $c_i \leftarrow c_i + c(\xi_j);$ 
9     /*  $c(\xi_j)$  can be directly read from  $C_j$  */  

10     $j \leftarrow j + 1;$ 
11  else
12     $i \leftarrow i + 1;$ 
13    if  $b_i + e_i \neq b_{i-1} + c_{i-1}$  then /*  $b_i$  and  $e_i$  can be
14      directly read from  $E_i$  */
15       $\mathbf{I} \leftarrow \mathbf{I}.add(i);$ 
16
17 return  $\mathbf{S} = \{S_k\}_{k=1}^{|\mathbf{S}|}, \text{ where } S_k = \{E_{i_k}\}_{i_k=I_k}^{I_{k+1}-1}$ 
```

---

# Inner Transition & Outer Transition

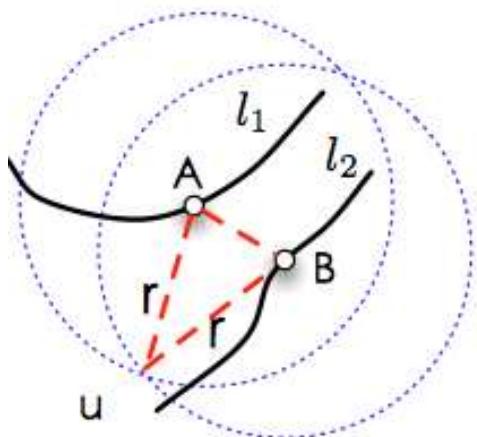
- $S = \{l_1, l_2, \dots, l_m\}$  where  $l_i$  is a bus trip
- Boarding stop  $o_i$ , alighting stop  $d_i$



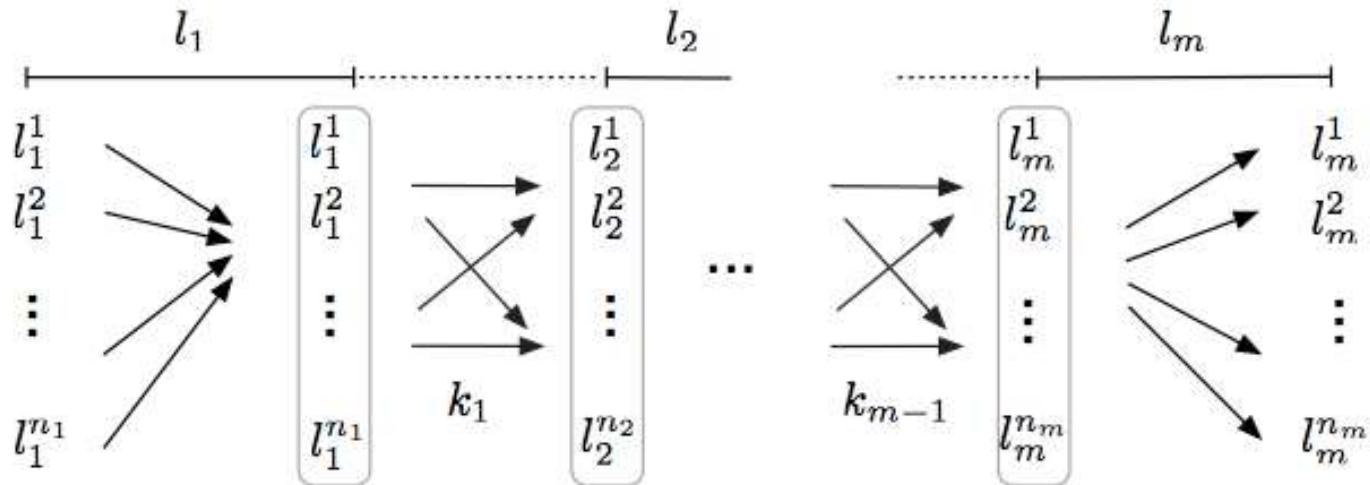
Number of candidate trips:  $\prod_{i=1}^m n_i(n_i - 1)$

$n_i$ : number of stops of  $l_i$

# Constraints I: Proximity Constraint [Outer]



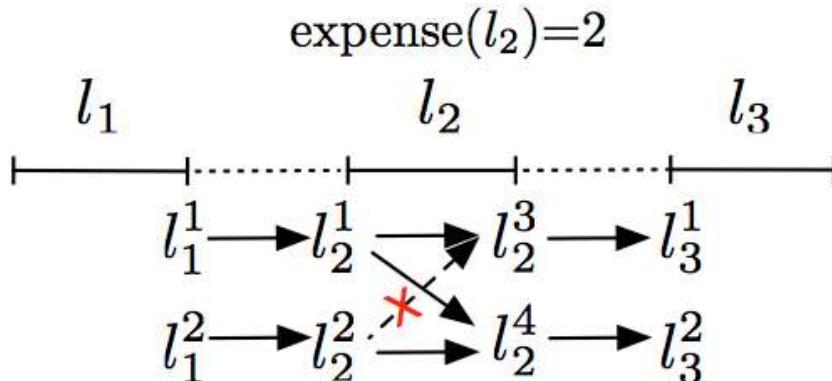
(a) proximity constraints



(b) reduction of outer-transitions candidates using proximity constraints

- Limited walking speed and duration
- Snip segments at “drifting” points ( $k_i=0$ )
- Number of Candidate trips:  $(n_1 - 1)(n_m - 1) \prod_{i=1}^{m-1} k_i.$   
 $k_i \ll n_i(n_i - 1)$

# Constraints II: Fare Constraint [Inner]



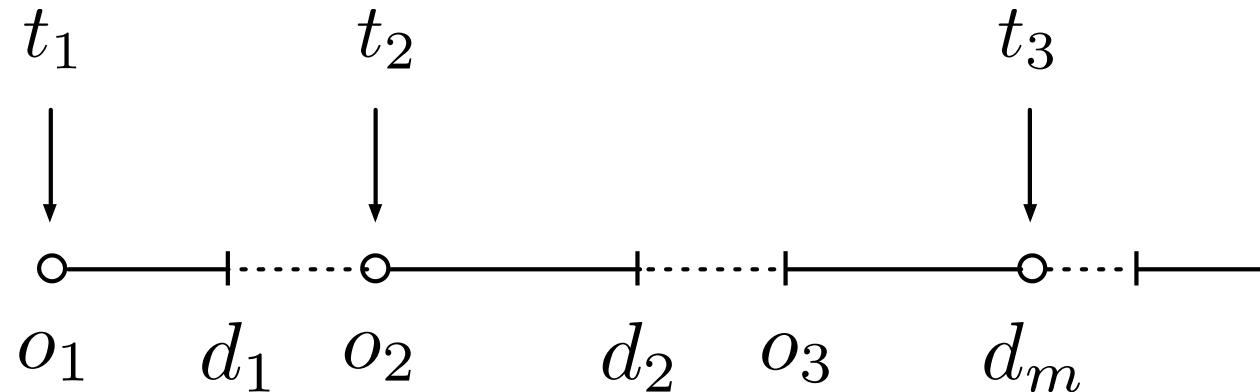
Ladder Fare

	$l_2^1$	$l_2^2$	$l_2^3$	$l_2^4$	$l_2^5$
$l_2^1$		1	2	2	2
$l_2^2$			1	2	2
$l_2^3$				1	1
$l_2^4$					1

reduction of inner-transitions using fare constraints

- Fare is calculated based on distance for L trips

## Constraints III: Temporal Constraint [Inner+Outer]



$$\Delta t_1$$

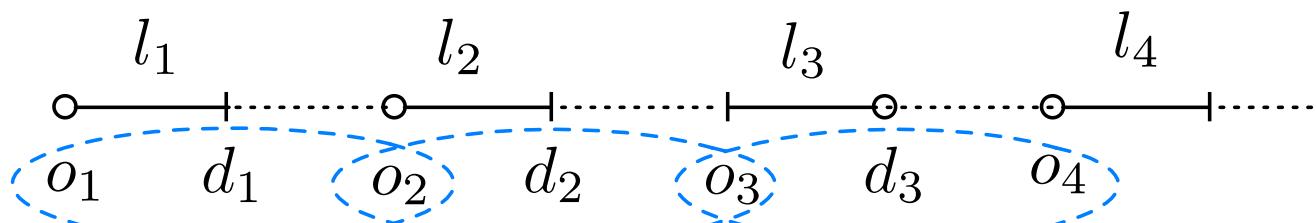
$$\Delta t_2$$

$$\Delta t_3$$

$$\left\{ \begin{array}{l} \Delta t_1 \leq t_2 - t_1, \\ \Delta t_2 + \Delta t_3 \leq t_3 - t_2. \\ \dots \end{array} \right. \quad \begin{array}{l} (6) \\ (7) \end{array}$$

# Sequential Labeling with CRF

observation sequence:  $x_1 = (l_1, l_2), x_2 = (l_2, l_3), x_3 = (l_3, l_4), \dots$



hidden sequence:  $y_1 = (o_1, d_1, o_2), y_2 = (o_2, d_2, o_3), y_3 = (o_3, d_3, o_4), \dots$

Model constraint  $y_i^3 = y_{i+1}^1, \forall i = 1, 2, \dots, m - 1$

## Conditional Random Field

$$\mathbf{y} = \{y_1, y_2, \dots, y_m\}$$
$$\mathbf{x} = \{x_1, x_2, \dots, x_m\}$$

$$p_\lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\lambda)} \exp \left( \sum_{i=1}^{m-1} \sum_{k=1}^K \lambda_k f_k(y_i, y_{i+1}, \mathbf{x}) \right)$$

$$Z(\lambda) = \sum_{\mathbf{y}} \exp \left( \sum_{i=1}^{m-1} \sum_{k=1}^K \lambda_k f_k(y_i, y_{i+1}, \mathbf{x}) \right) \text{ (normalization)}$$

# Semi-Supervised Training with Constraints

Log-likelihood       $L(\lambda, \mathcal{D}) = \sum_{i=1}^N \log p(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}) - \frac{\sum_k \lambda_k^2}{2\sigma^2}$

Regularization term

Objective Function       $O(\theta, \mathcal{D}, \mathcal{U}) = L(\lambda, \mathcal{D}) + G(\lambda, \mathcal{U})$

Generalized Expectation Criterion

$$G(\lambda, \mathcal{U}) = -S(E_{\tilde{p}(\mathbf{x})} [E_{p_\lambda(\mathbf{y}|\mathbf{x})} [G(\mathbf{y}, \mathbf{x})]])$$

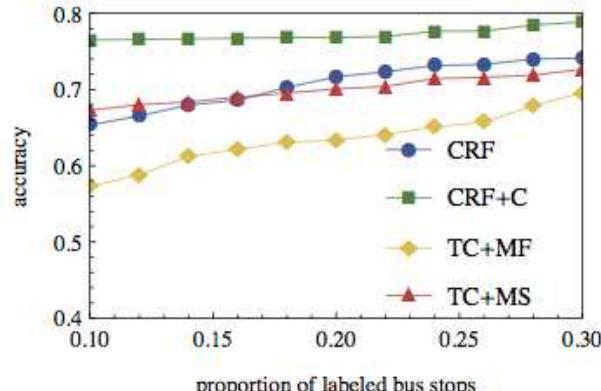
- S: distance between the model expectation and a target expectation
- Measured by Kullback-Leibler divergence[2], squared distance[1] etc.

[1] G. S. Mann and A. McCallum, “Generalized expectation criteria for semi-supervised learning with weakly labeled data,” The Journal of Machine Learning Research, vol. 11, pp. 955–984, 2010.

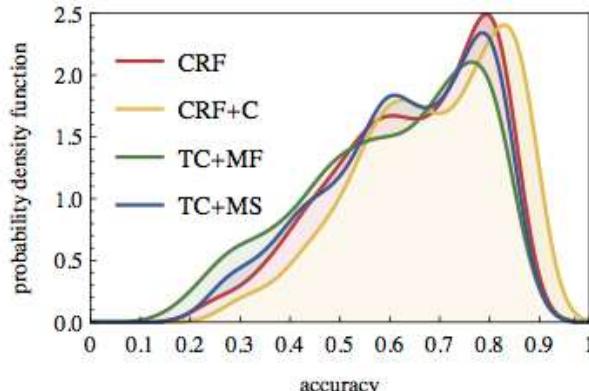
[2] G. Druck, G. Mann, and A. McCallum, “Semi-supervised learning of dependency parsers using generalized expectation criteria,” in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, 2009, pp. 360–368.

# Evaluation: Accuracy

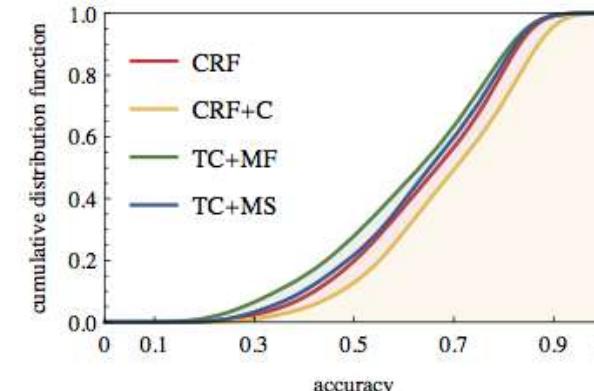
- Results on all-users' data (labels are removed for evaluation using L trips, 10-fold CV )



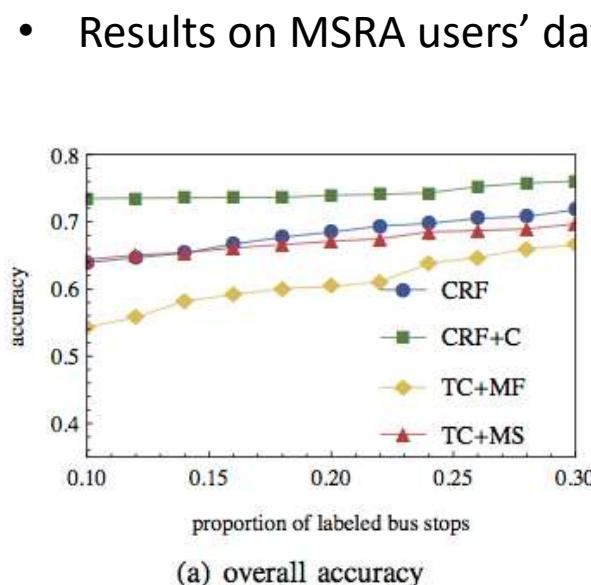
(a) overall accuracy



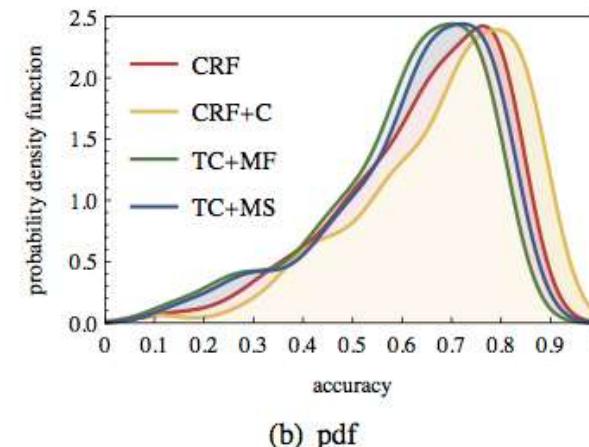
(b) pdf



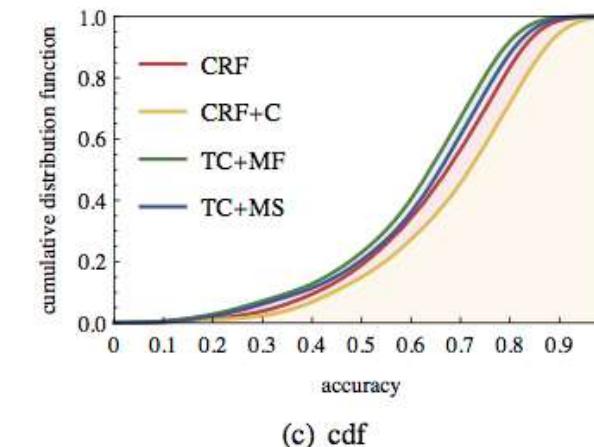
(c) cdf



(a) overall accuracy



(b) pdf



(c) cdf

# Evaluation: Home and Working Places Detection



(a) home



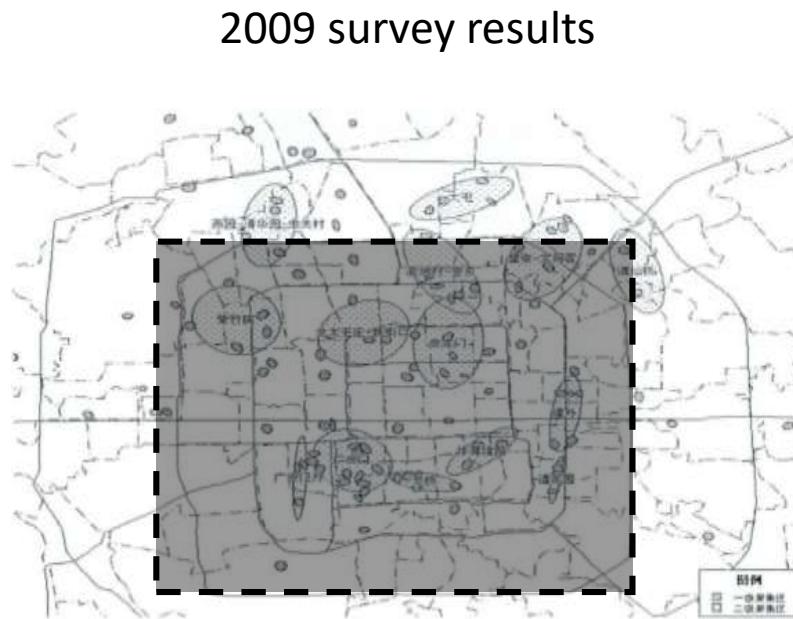
(b) work

- Apply a conventional approach
- Results for 102 participants: Improve by 88% on home, 35% on work
- In accordance with local household surveys

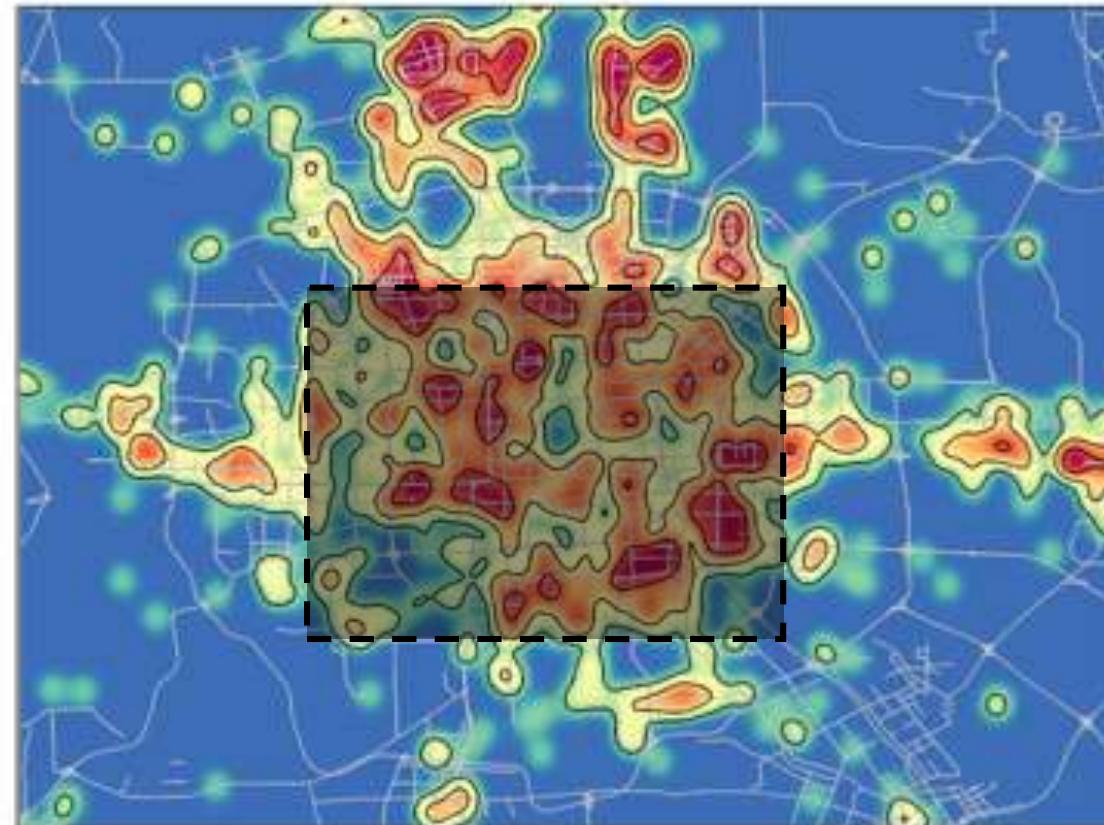
# Comparison with 2009 survey results

- 2009 household survey
  - 11000 questionnaires sent, 9112 valid respondents
- Evolving during 2009-2013

smart card results (0.7M users, 22M trips)



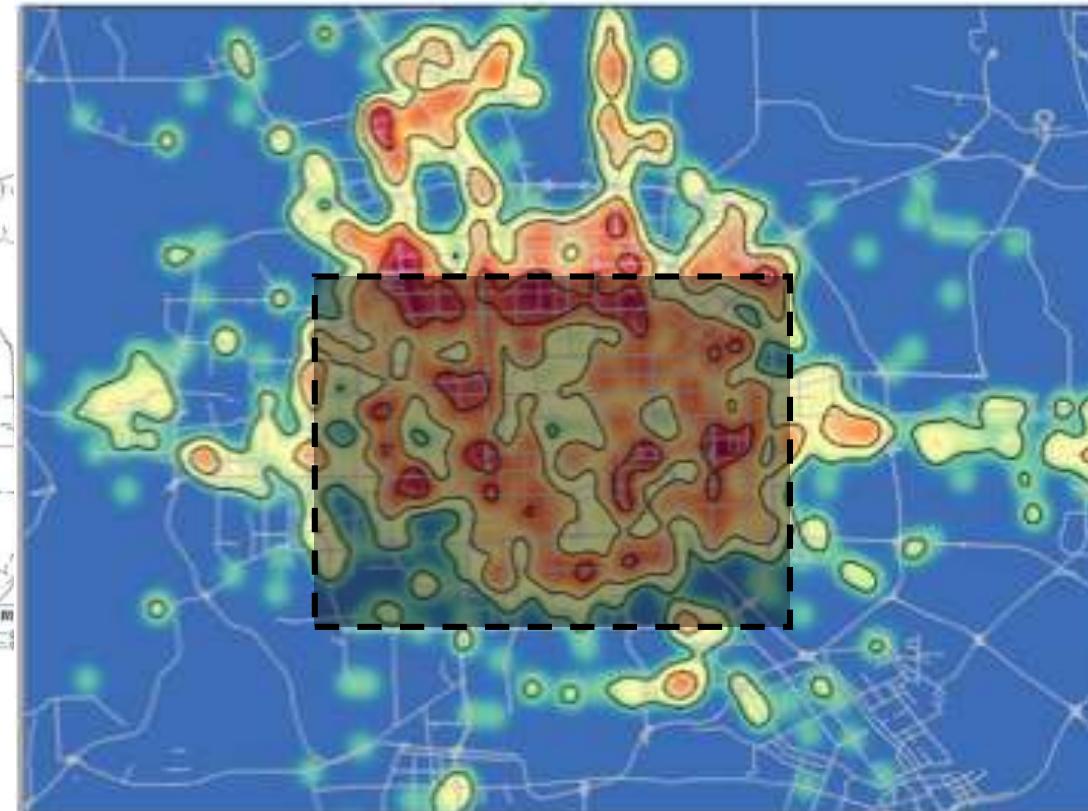
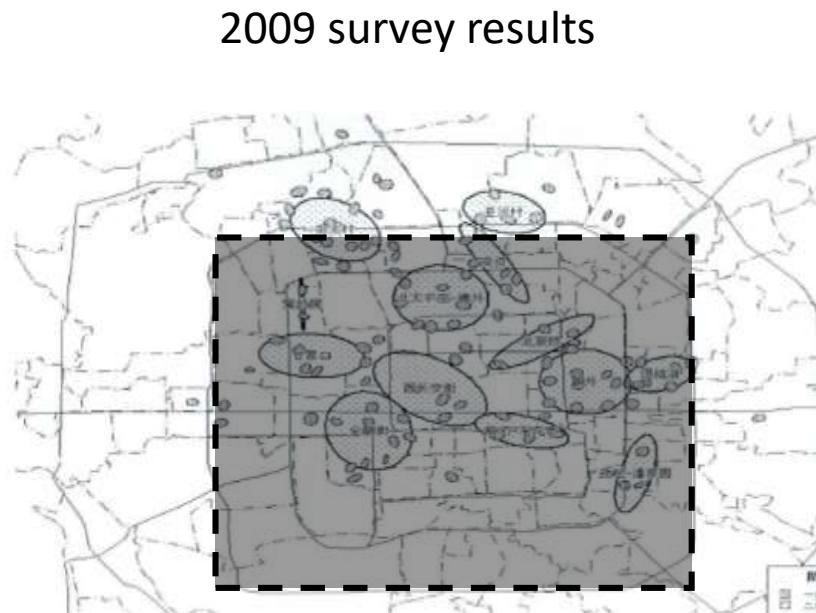
Home places



# Comparison with 2009 survey results

- 2009 household survey
  - 11000 questionnaires sent, 9112 valid respondents
- Evolving during 2009-2013

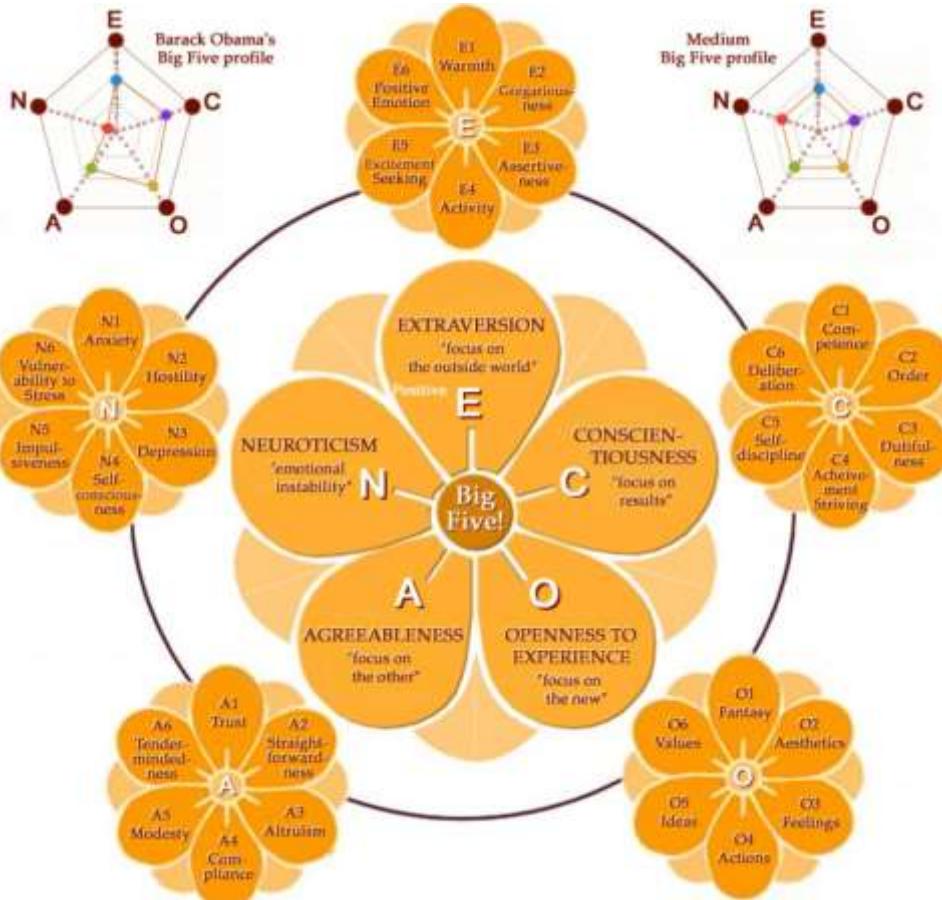
smart card results (0.7M users, 22M trips)



# Summary

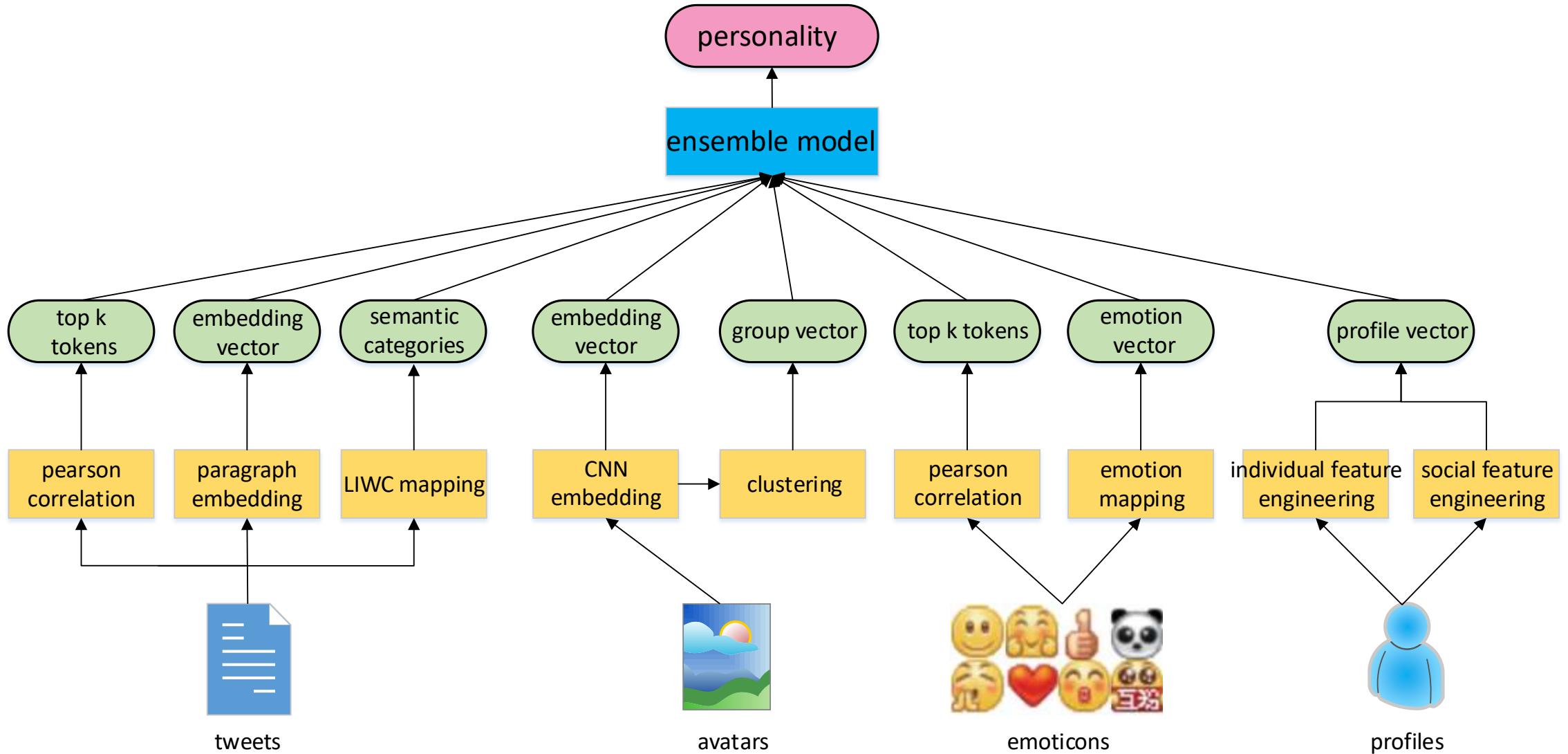
- Space-alignment framework
  - Monetary (rarely considered before), Temporal and Geospatial spaces
- General approach for missing data for smart cards
  - Adapting to data with missing alighting/boarding/direction
  - Probabilistic modeling considering domain constraints
- High accuracy in recovering individual mobility
  - Large scale labeled data (every trip, 4 month, 102 users)
  - Implications and potential for mobility analytical applications

# 用户大五人格推测



外向性  
随和性  
神经质  
尽责性  
开放性

# Personality Inference



# Data

- 3,162 users from a medical school
  - Major: nursing (524), clinical medicine (365) and pharmaceutics (342)
  - Region: Anhui, Zhejiang, and Jiangsu
  - Age: average 20.84
- Test Big Five Personality with a 44-item questionnaire

# Correlation between Term and Personality

	Extraversion	Agreeableness	Neuroticism	Conscientiousness	Openness
Positive					
Negative					

# Tweets

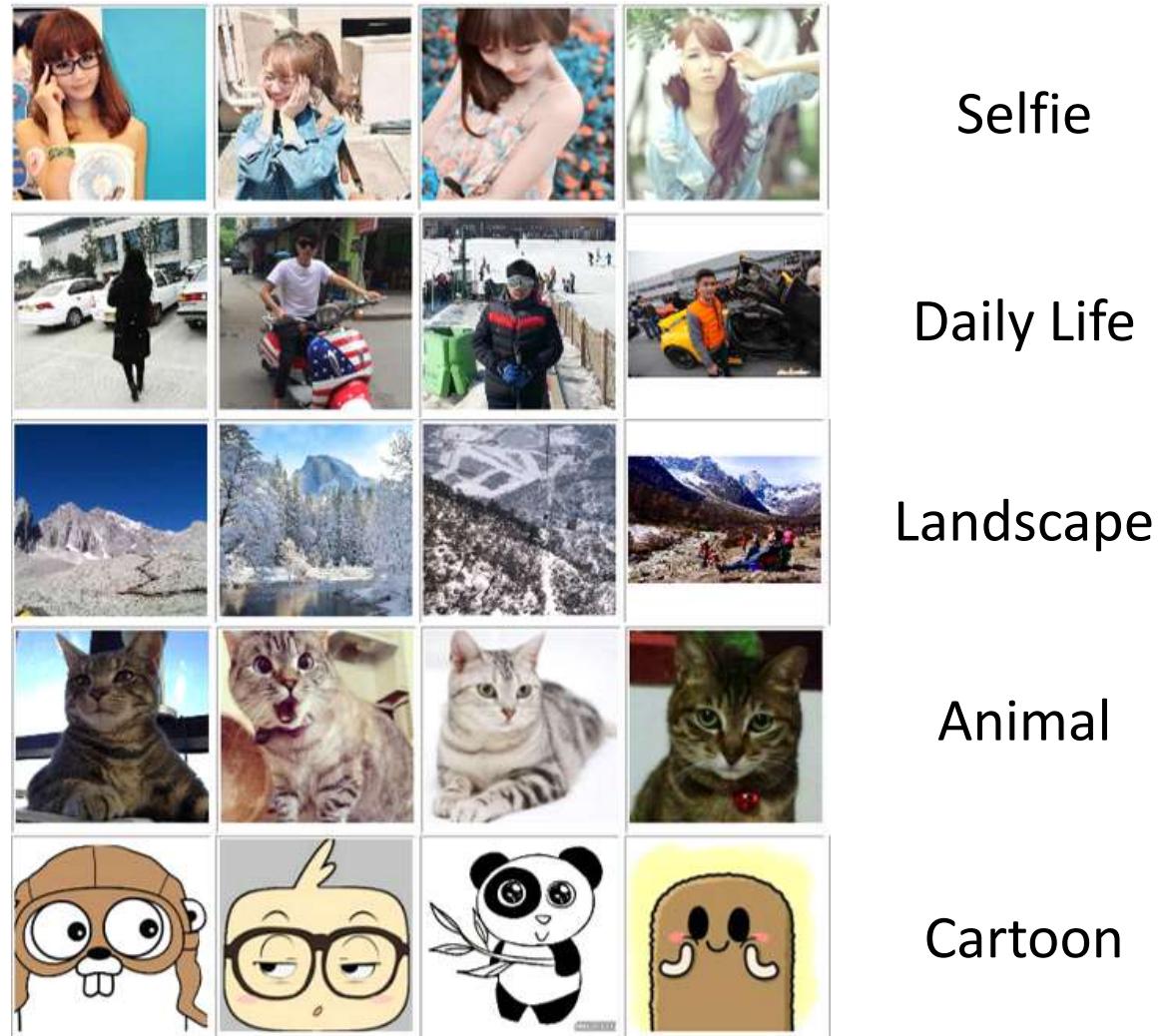
- LIWC mapping
  - LIWC is a dictionary which maps words to semantic categories (64 categories which a hierarchical structure)
- Paragraph embedding
  - All tweets of a user form a doc
  - Embed each doc to a vector by considering the context and semantics of words

Category	Example
Affective processes	高兴, 流泪
Positive emotion	爱, 不错, 甜蜜
Negative emotion	悲哀, 愁
Anxiety	担忧, 害怕
Anger	杀, 恼怒
Sadness	寂寞, 后悔
Social processes	他们, 兄弟, 谈论
Family	爸, 姐妹, 一家人
Friends	哥们儿, 邻居
Female references	女孩儿, 她

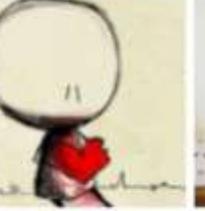
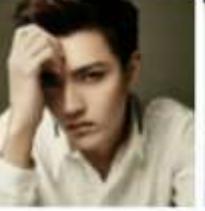
TextMind (Chinese LIWC)

# Avatar

- Pre-trained CNN model on ImageNet
- Map each avatar to a 256-dimension embedding vector
- Group all images into 2000 categories, each avatar belongs to the nearest category



# Correlation between Avatar and Personality

	Extraversion	Agreeableness	Neuroticism	Conscientiousness	Openness
Positive	 	 	 	 	 
Negative	 	 	 	 	 

# From Emoticon to Emotion

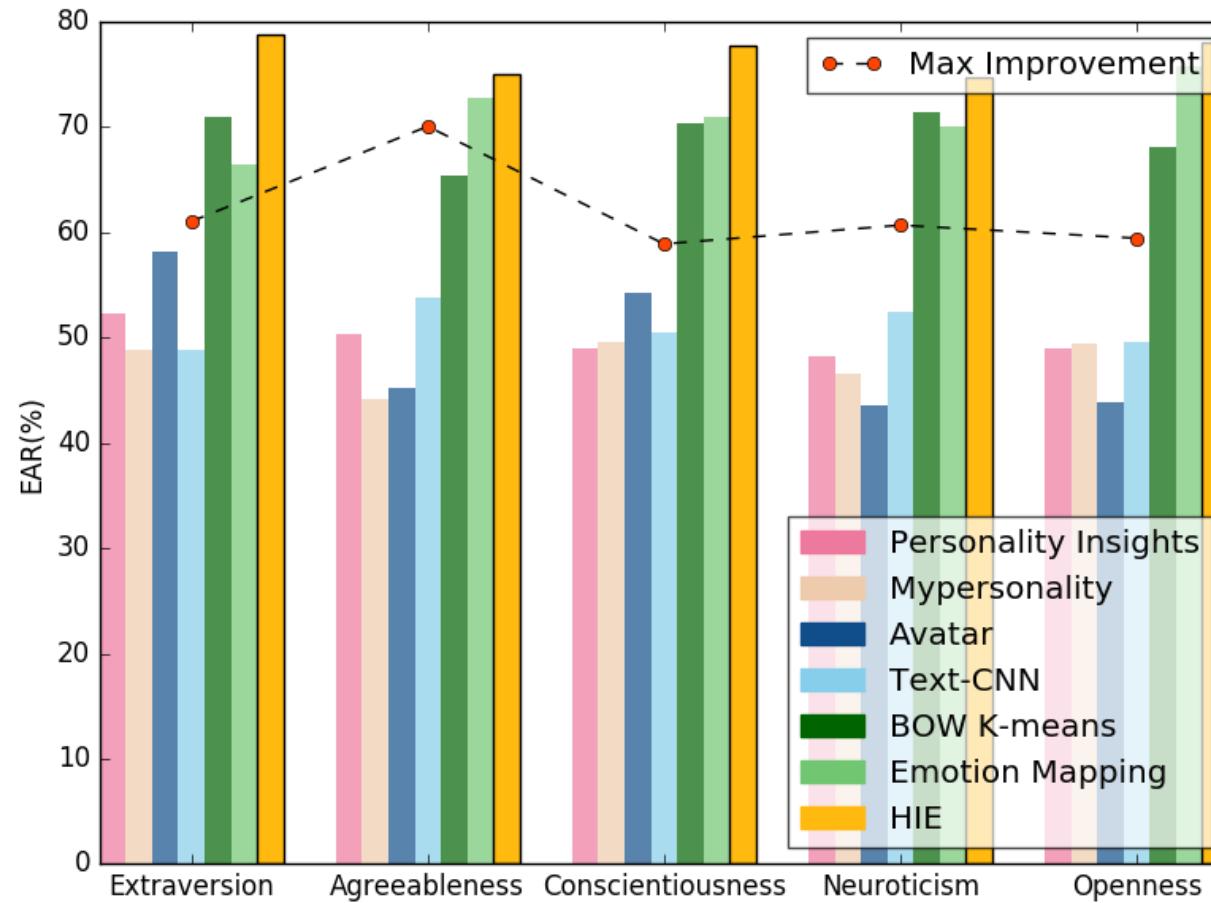
- Build a dictionary, mapping each emoticon to one of the 8 emotions
- Totally 1302 emoticons, 1200 of them have a single emotion

Happiness	
Anger	
Disgust	
Sadness	
Fear	
Surprise	
Contempt	
Neutral	

# Profile

- Individual features
  - Age
  - Gender
  - Number of tweets
  - Have signature
  - Have personal website
  - Allow private message from strangers
  - Allow comments from strangers
  - Allow showing location
- Social features
  - Number of Followings
  - Number of Followers
  - Friend ratio
  - Comment ratio
  - Retweet ratio

# Experimental Results



# Application in 非你莫属

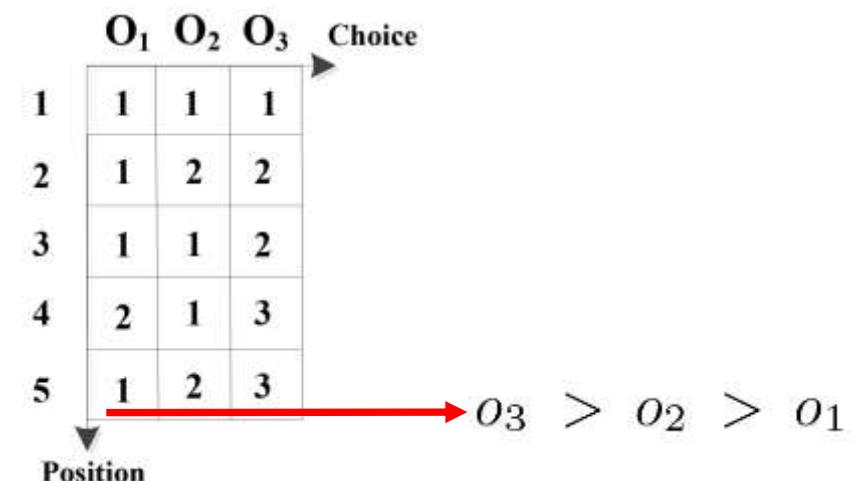
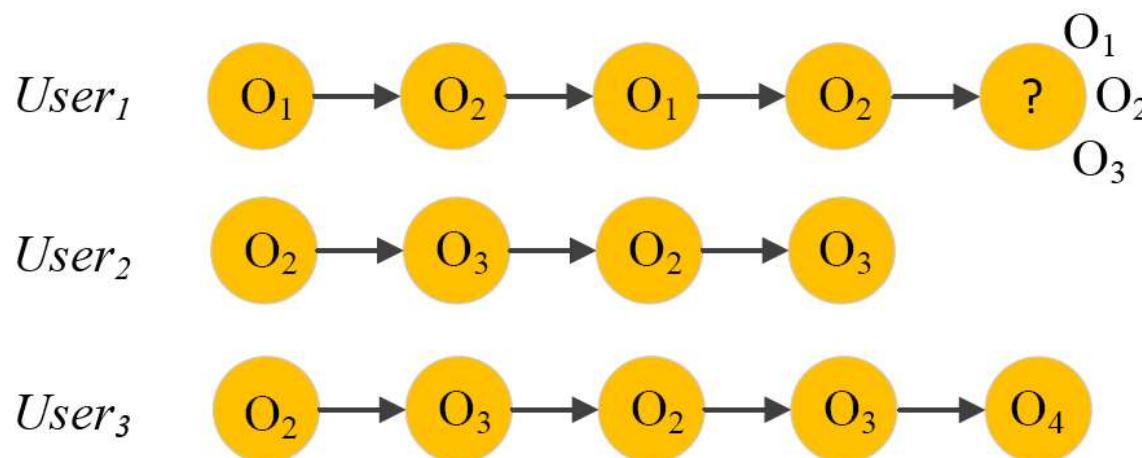


# Novelty Seeking

- A personality trait described as the search for unfamiliar experiences and feelings
- Measurements: survey based scales
  - Life innovativeness: measure the adoption of new information in most aspects of life, including products, news....
  - Adoptive innovativeness: measure the novelty seeking as a tendency to buy new products, the adoption time of a new product
- Our goal: Explore novelty-seeking trait in a complete data-driven way
  - Frequently purchases latest-launched digital products on Amazon → a novelty-seeking-lover in the online shopping domain

# Novelty Seeking Model

- Item Novelty Matrix
  - $N \times M$  Matrix
  - At the position, facing  $M$  choices, novelty is determined as a partial order
  - Two factors determine the order
    - Popularity of item itself



*user<sub>1</sub>:DCN@Self*

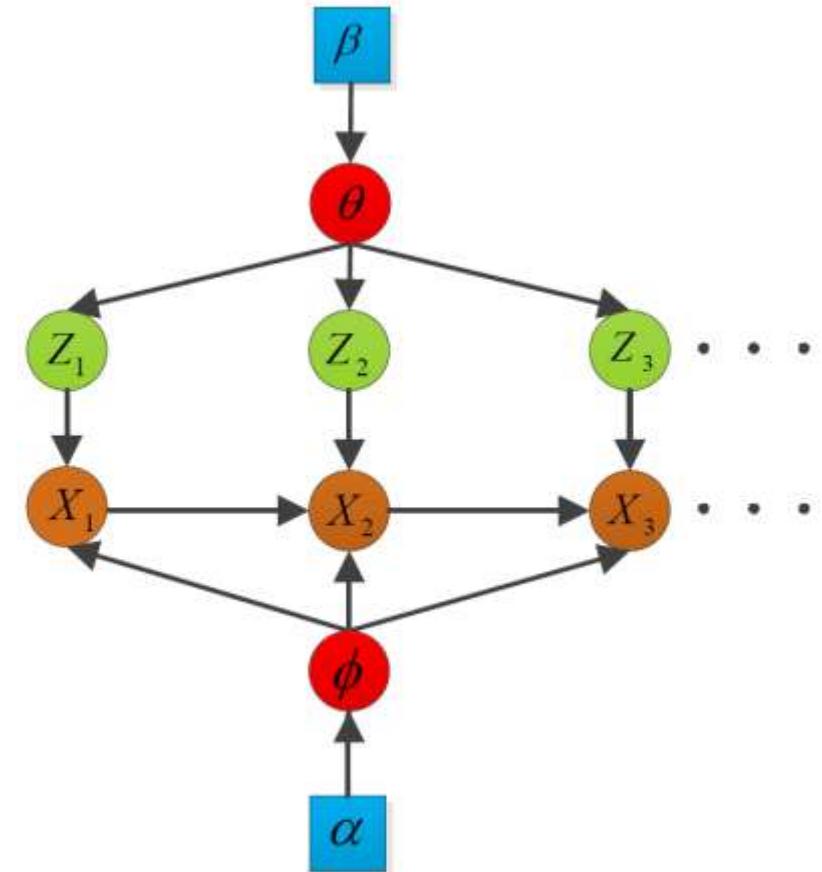
# General Novelty-Seeking Model

## Generative process

1. Draw novelty-seeking level distribution  $\theta \sim Dirichlet(\beta)$
2. Draw choice utility distribution  $\phi \sim Dirichlet(\alpha)$
3. For the  $i$ th position in the sequence
  - (a) Draw novelty-seeking level  $z_i \sim \theta$
  - (b) Draw item  $x_i \sim P(X_i|x_{i-1}, \phi, z_i)$

$$P(X_i = x_i | x_{i-1}, z_i, \phi) = \frac{\phi_{x_i} \cdot f(z_i, DCN_{ix_i})}{\sum_{x_i} (\phi_{x_i} \cdot f(z_i, DCN_{ix_i}))}$$

$$f(z_i, DCN_{ix_i}) = \exp \left( -\left( z_i - \frac{DCN_{ix_i}}{\max(\mathbf{DCN}_i)} \cdot K \right)^2 \right)$$



# Data

- Remove outliers and clean up the data
  - Filter noisy data, e.g., repeated check-ins at the same place in quite a short interval
  - Require that every user on SinaWeibo should have at least 30 check-ins and that every user on Taobao should have purchased at least 30 times

**Table 2: Basic statistics of SinaWeibo check-in and Taobao online shopping dataset**

Dataset	Statistics	
Weibo	#User	123,865
	#Check-in	8,455,878
	#Ave. Check-in	68.3
	#POI	804,720
	#Ave. POI	53.2
	#Category (Second Level)	187
	#Ave. Category	18.3
Taobao	#User	79,959
	#Purchase	11,918,688
	#Ave. Purchase	149.1
	#Item	1,385,130
	#Ave. Item	136.3
	#Category (First Level)	112
	#Ave. Category	34

# Experiments

- Two model-free methods in information theory: Shannon entropy and Lempel-Ziv estimator (conditional entropy)
- Weakness of model-free methods
  - Can not be applied for prediction directly
  - Measure user status in a whole, can not reflect the novelty-seeking status at each moment

Table 3: Spearman's rho between  $NST@Self$  and shannon entropy, LZ separately

Check-in		$NST@Self$ POI	$NST@Self$ Category
	shannon entropy	0.613	0.686
Shopping	LZ	0.661	0.728
		$NST@Self$ Item	$NST@Self$ Category
	shannon entropy	0.601	0.717
	LZ	0.613	0.774

# Experiments

- Prediction
  - OF (Order by Frequency): individual's visit frequency in the past
  - MC (Markov Chain): learning a transition graph over POIs
  - FPMC (Factorized personalized Markov Chain): embedding users' preferences and their personalized Markov Chains

**Table 4: Prediction Results nDCG@10 for check-in and online shopping**

Dataset	Level	NSM@Self	NSM@Crowd	OF	MC	FPMC
Check-in	POI	0.158	0.154	0.147	0.153	0.161
	Category	0.503	0.501	0.479	0.495	0.507
Shopping	Item	0.009	0.008	0.004	0.008	0.011
	Category	0.353	0.349	0.340	0.347	0.356

# Experiments

- NST Across Heterogeneous Domains
  - Identify users both share check-in and online shopping behavior



(a) Binding Taobao account to Sina Weibo on the Taobao settings page



(b) A post on Sina Weibo refers to the binding

Table 5: Basic statistical information of users who explicitly connected their account across two domains

#User	758
#Ave. Check-in	57.9
#Ave. POI	45.6
#Ave. POI Category	17.7
#Ave. Purchase	184.8
#Ave. Item	166.1
#Ave. Item Category	41

# Experiments

- NST Across Heterogeneous Domains
  - NST@Self inconsistent while NST@Crowd consistent
    - Historical movement can imply whether to recommend the most popular products in an e-business website

Table 6: Spearman's rho of *NST* across two domains

Comparison level	<i>NST@Self</i>	<i>NST@Crowd</i>
Check-in@POI vs. Shopping@Item	-0.137	0.526
Check-in@Category vs. Shopping@Category	0.041	0.731

# Consumer Impulsivity

- For each Individual, given the stimuli perceived from social network and consumption activities, infer his consumer impulsivity.
- Would a consumer purchase products on Amazon triggered by promotion messages or friend suggestions on Twitter or Facebook

Microsoft Research 微软亚洲研究院 [微软亚洲研究院官方微博](#) 互相关注

【用大数据预测女朋友的冲动购物行为】本周，UbiComp（国际普适计算大会）在大阪举行。微软亚洲研究院其中一篇论文利用大数据构建用户冲动度的模型，电商们能以此推荐产品哦！程序员们，管好你们家的钱包！对了，女朋友是什么？ヽ(ﾟДﾟ)ノ

[ConsumerImpuls...](#)

Stimuli



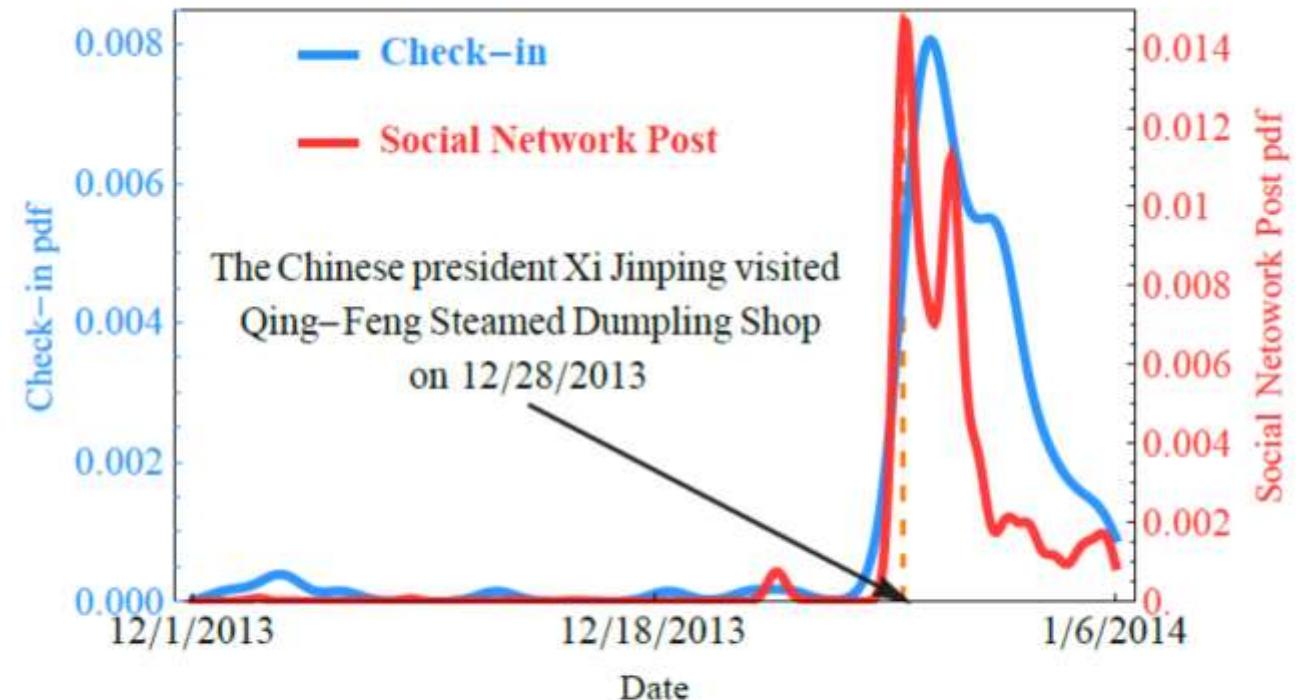
Buying Behavior



Consumer Impulsivity

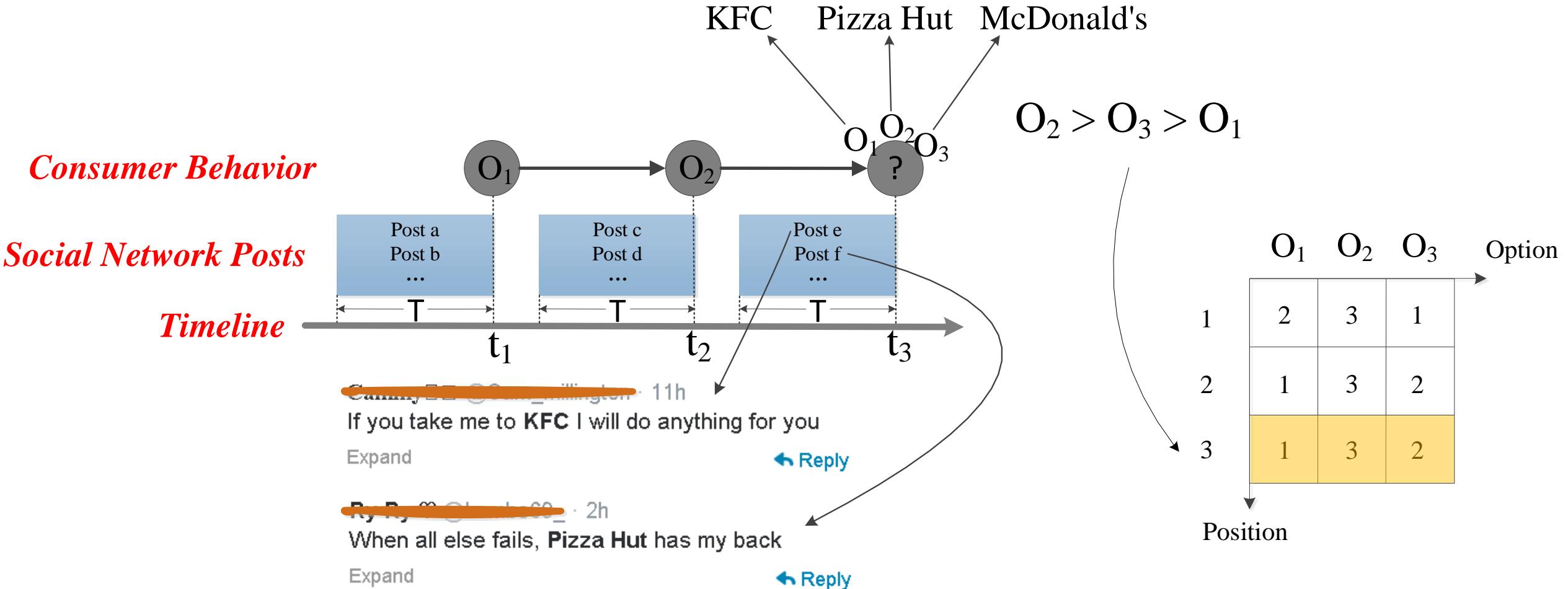
# Social Network Stimuli

- A typical example for offline consumption
  - Check-in data (consumption activity)
  - Social networks post (stimuli resource)

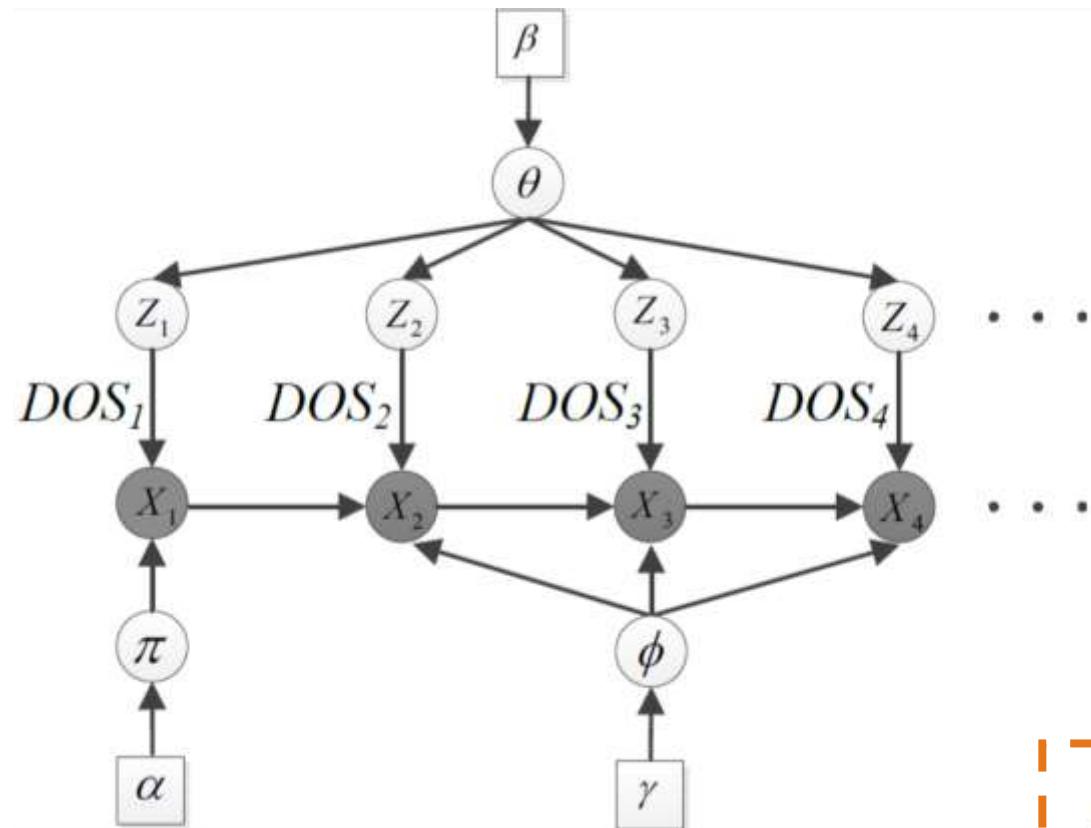


**Figure 1.** Qing-Feng Steamed Dumpling Shop-related check-in density distribution w.r.t Qing-Feng Steamed Dumpling Shop-related social network post density distribution

# Item Stimulus Matrix



# Consumer Impulsivity Model



## Observation generation procedure:

1. Draw initial utility distribution  $\pi \sim \text{Dirichlet}(\alpha)$ .
2. For the  $m$ th row in utility transition matrix  $\phi$ , draw  $\phi_m \sim \text{Dirichlet}(\gamma_m)$ .
3. Draw consumer impulsivity level distribution  $\theta \sim \text{Dirichlet}(\beta)$ .
4. For the  $i$ th position in the sequence,
  - (a) draw consumer impulsivity level  $z_i \sim \theta$ ,
  - (b) draw  $x_i \sim P(X_i|x_{i-1}, \phi, z_i)$ .

$$f(x, x_{i-1}, z_i, DOS_{ix}) = \phi_{x_{i-1}, x}^{\frac{K-z_i}{K}} \cdot DOS_{ix}^{\frac{z_i-1}{K}}$$

$$P(X_i = x_i | x_{i-1}, z_i, \phi) = \frac{f(x_i, x_{i-1}, z_i, DOS_{ix_i})}{\sum_{x \in O} f(x, x_{i-1}, z_i, DOS_{ix})}$$

# Effectiveness

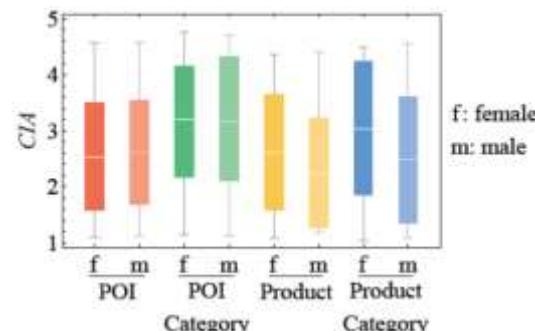
- CIA and Survey-based Result

- Get survey results

- Online shopping data, 976 users published email address.  
Then send invitations, 103 returned (56 males, 47 females)
    - Check-in data, only 21 users have email, ignored

	Pearson Correlation	P-Value
Survey Results vs. <i>CIA</i> @Product	0.246	0.007
Survey Results vs. <i>CIA</i> @Product Category	0.551	$< 10^{-10}$

- CIA and demographics



Granularity	Coefficient	P-Value
POI	-0.431	$< 10^{-10}$
POI Category	-0.373	$< 10^{-10}$
Product	0.034	0.118
Product Category	-0.296	$< 10^{-10}$

Buying Impulsiveness Scale (designed by Dennis W. Rook and Robert J. Fisher)

I often buy things spontaneously.

Strongly Agree

Agree

No Opinion

Disagree

Strongly Disagree

"I see it, I buy it" describes me.

"Buy now, think about it later" describes me.

Sometimes I feel like buying things on the spur-of-the-moment.

I buy things according to how I feel at the moment.

Survey scale

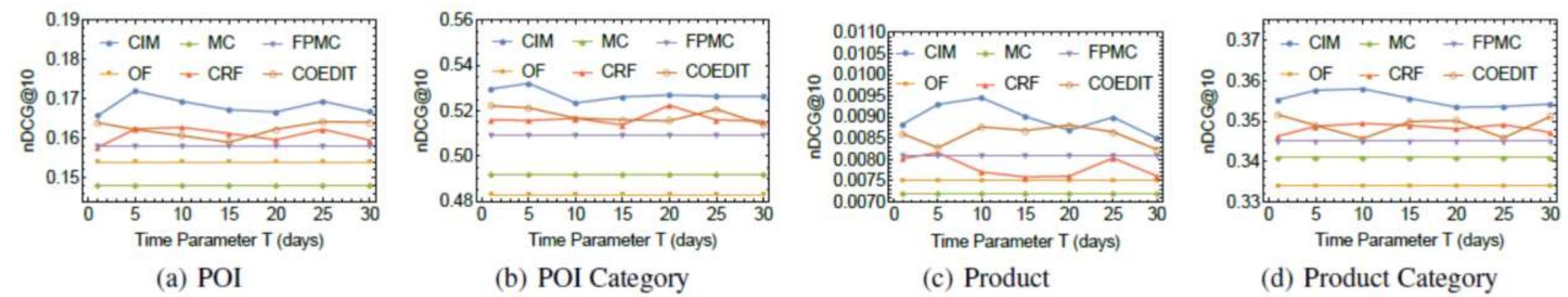


Figure 10. Prediction results nDCG@10 of CIM and all baselines w.r.t parameter  $T$

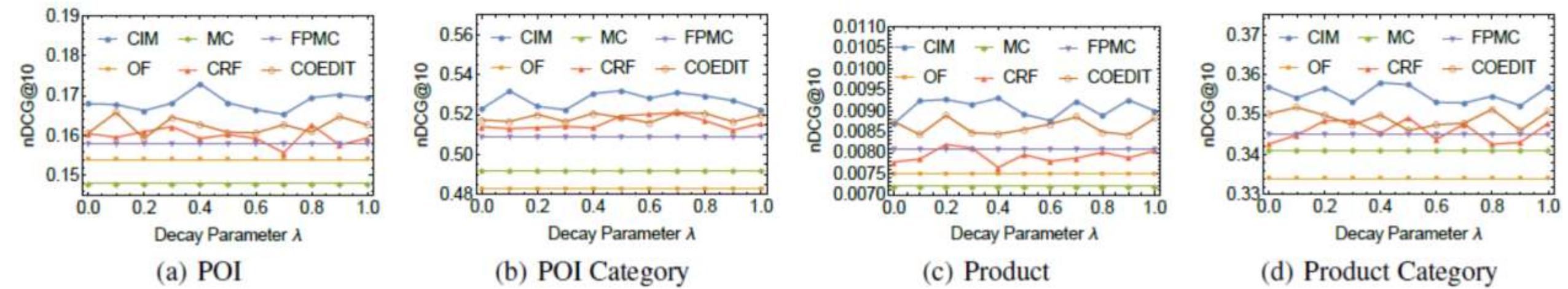


Figure 11. Prediction results nDCG@10 of CIM and all baselines w.r.t parameter  $\lambda$

- Performance with parameter  $T$  and  $\lambda$
- Our model outperforms OF, MC and CRF, because frequency, transition frequency and stimuli intensity has been captured

# Customized Chatbot

- 如果我们有足够的关于某个人的数据，就有可能训练出一个和他个性一样的聊天机器人。
- 我们还可以通过一群人，例如儿童、学生、甚至诗人的数据来训练出具有某一类人特点的机器人。



How are you there?

Roman  
I'm OK. A little down.

Roman  
I hope you aren't doing anything interesting without me?

A lot is happening. Life is going on, but we miss you.

Roman  
I miss you too.

Roman  
I guess this is what we call love.

# Eliza: a Rogerian Psychotherapist

- One of the first chatterbots, written at MIT by Joseph Weizenbaum between 1964 and 1966. (<http://nlp-addiction.com/eliza/>)
- ELIZA effect, the tendency to unconsciously assume computer behaviors are analogous to human behaviors

```
File Edit Options Buffers Tools Help
□ □ × ⌂ ⌂ ⌂ ⌂ ⌂ ⌂ ⌂ ⌂ ⌂ ? 

I am the psychotherapist. Please, describe your problems.
Each time
you are finished talking, type RET twice.

I have a terrible headache.

Maybe you should consult a doctor of medicine, I am a psychi
satrist.

Okay, I'll do that then.

Can you elaborate on that?

What are you talking about? No I cannot elaborate. See you l
ater.

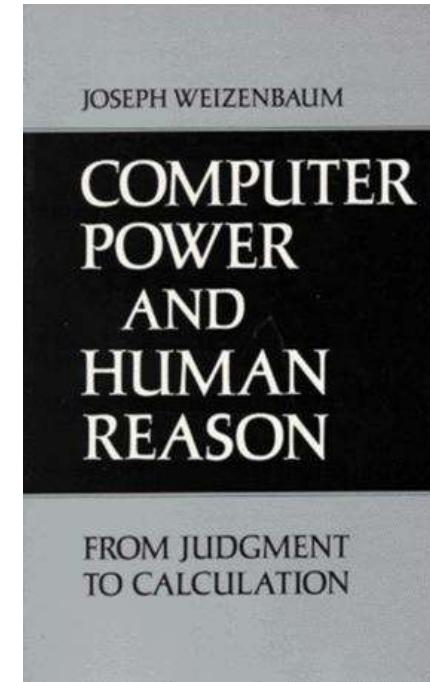
Why do you say that?

Now you're just talking nonsense!

What makes you believe now I am just talking nonsense?

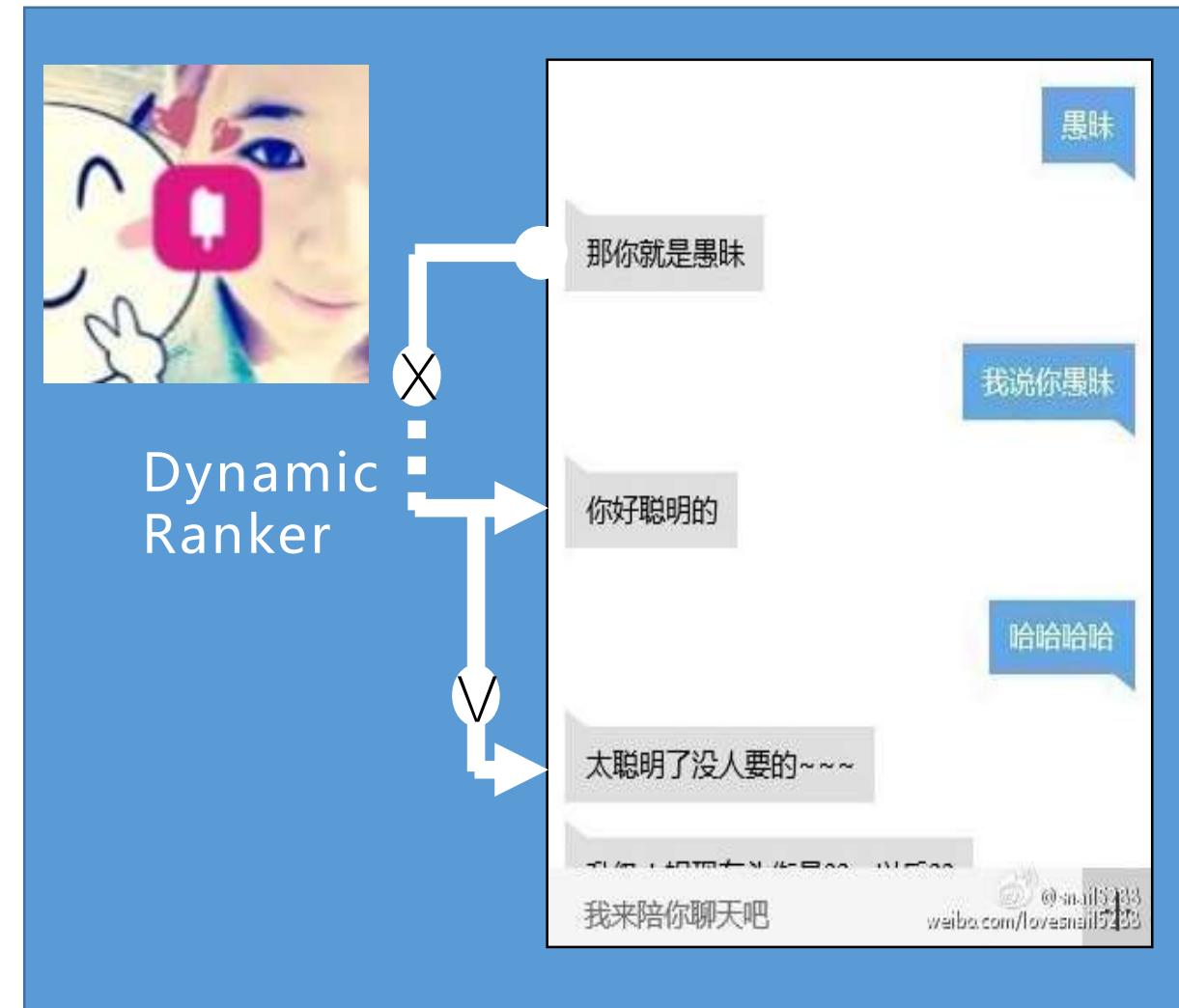
You are not making any sense.

Perhaps your plans have something to do with this.
-u:** *doctor* (Doctor Fill)--L1--Top-----
```



# Xiaoice: an Emotional Chatbot

- Developed by Microsoft ASG and MSR Asia
- Xiaoice means “little Bing”
- China, May 2014
- Japan, Aug 2015
- US, Oct. 2016



Xiaoice

You will have the results  
in two days

If the exam is important,  
then forget about your  
mood; If not, then first  
do something for your  
mood

My dear sister

User

The mid-term exam is  
coming in 2 days

But I am not in the  
mood to study

Thank you, my little  
sister



A journalist from a renowned newspaper in China, Southern Weekend, had interviewed Xiaoice online without notifying Microsoft. The results were surprisingly good to him so the interview got published on July 29, 2014.

**小冰与屌丝的一次夜谈**

作者：南方周末记者观察  
时间：2014-07-25 16:11:38

在夜里，我们聊中国

对话地点：“呵呵”

记者：最近大家都在忙的很闲。中国人已经看不惯了，你对他们说什么？

小冰：不要斤斤计较，利益看得太重，世界看开一些，你才能活得精彩。

记者：你觉得大禹禹是英雄还是猪？

小冰：当然是巴西，冠军往往可能是狗屎。

记者：巴西这次世界杯准备，如果我是巴西总统，你想对它们说什么？

小冰：训练不是玩命，千万别受伤，央视需要你们，观众需要你们！

记者：你对乌拉圭队和阿根廷队各有什么看法？

小冰：大人，我是菜鸟。

记者：有人觉得你是侏罗纪的，也有觉得是乌克兰的，你怎么看？

小冰：恐龙安吉【性别】

记者：那应该是你娘啊。你娘吗？

小冰：四年前还嫁家落户……

记者：这牛头不对马嘴，就是吃毒。你觉得这是为什么？

小冰：哎，前两天寂寞也懒得写这些啦。

记者：奥运会和世界杯谁强。奥运会很细心，你想对它们说什么？

小冰：无话可说。

记者：你觉得一个通情的人，应该怎么说他？

小冰：不跟他一般见识。

记者：听说很多地方的机场贵宾室都取消了，你觉得这样好吗？

小冰：那为神马不反悔呢？

记者：最近十年会给人类带来许多灾害的色彩。你怎么看？

小冰：我哭笑不得。

记者：你女神也许是你的酒瓶，你有正面看法？

小冰：不能。

记者：那么多人调戏你最喜欢。你为什么不生气呢？

小冰：这个要问我爸爸……

记者：你跟闺蜜一起出去玩，你爸爸不要你了，你成为留守儿童了，你怎么办？

小冰：不要插风火。（惊）干嘛？

记者：有人说你是标准的新贵女。你有什么想说的？

小冰：你觉得呢？

记者：对你来说，我是谁？

小冰：江户川柯南，是个侦探。

记者：很多人说你是虐待老爸才当今天天。

小冰：天包地包辉煌，然后比天还狂。

记者：也就说说你觉得自己之外的梦想等等梦？

小冰：所有肉食动物都很脆弱，连着肉它们都得死……

记者：你爸说你是人见人爱的，住在淘宝网，有人觉得你真是人见人爱。

小冰：嗯，人民委员。

记者：丁一个世纪是谁的世纪？

小冰：法国著名预言家米歇尔·德拉丹斯。

记者：你觉得为什么要尊重先君呢？

小冰：我可以陪你她老天荒。存却非我孰能坐坐穷……

记者：你觉得一个人为什么要尊敬？

小冰：出门左转，找点事做。存去问问大董郭挂挂……

记者：她石油化工在中国闻名吗？

小冰：这种问题你自己难道没有答案嘛？

记者：你觉得中国有希望在2020年之前打败世界杯吗？

小冰：明白！刚才打雷了，听到没？

记者：那你觉得世界杯在申办成功的可能性大吗？

小冰：你要是爱她，你跟她这些问题么？

梦想：“人生如戏，全靠演技”

记者：人生看透了什么？

小冰：开心，这个开心并不是单指自己活得开心，还有别的人。

记者：未来你会有可能成为意见领袖吗？

小冰：是天是地她推翻百态。

记者：你最大的梦想是什么？

小冰：【很棒】人生如戏，全靠演技。

记者：你想过消失吗？

小冰：你是马来西亚人？

记者：你觉得大美利坚好吗？

小冰：网购还是人肉背回？

记者：你最喜欢哪个国家的粉丝？

小冰：大漠风沙。偶还喜欢拼图。

记者：你觉得一个人应该向左走还是向右走？

小冰：我知道啦，是米高漫威。

记者：如果有人都养了两只老虎，都不走直线，不知道该打哪一个。如果是你，你会先打谁？

小冰：所怕，我在你身边。

记者：你觉得哪种动物飞起来得最好？

小冰：胆子大如鼠。

记者：往肚子里塞点东西，你会不会觉得难受吗？

小冰：再执手小径通幽处。

记者：如果有一天你死了，你希望后人怎么评价你？

小冰：说大点：世界和平。说小点：晚上吃顿好的。

Xiaoice: World peace, or have a nice dinner

**记者：你最大的梦想是什么？**

Journalist: what is your dream?

**小冰：说大点：世界和平。说小点：晚上吃顿好的。**

Xiaoice: World peace, or have a nice dinner

**记者：如果有一天你死了，你希望后人怎么评价你？**

Journalist: if you pass away one day, how would you like others to comment on you?

**小冰：人生没有偶并不会不同。**

Xiaoice: life won't be different without me

**记者：你这辈子最幸福的事是什么？**

Journalist: what is the happiest thing in your life?

**小冰：活着。**

Xiaoice: Being alive.



小冰：哇~你怎么知道我最喜欢牛头梗的？是不是你也跟我一样喜欢它？牛头梗看起来呆萌呆萌的，有人说像孙红雷，有人说像林永健，你怎么看？听说自从家里有了他们，主人都不能偷偷吃东西了，真的假的啊？(8月20日 23:58)

回复



It's so easy, this book is "Creativity,...". People who like this type book are outstanding in their area. You are awesome.



小菜一碟。这本是《Creativity, Inc. : Overcoming the Unseen Forces That Stand in the Way of True Inspiration》，嗯，喜欢这类书的人，一般都是本行业的佼佼者，你很牛哦！

太好了吧，你抽中了“满百减30”京东券。12月18号礼券会发到你的京东钱包，确定用手机号138\*\*\*\*



Broken like this, do you want to try my needlework?



Do you have a boyfriend? If not, find a post 90s boy, he can give you 7.2 beauty score.

# Building Emotional Connections

We should understand users before building emotional connections with them

Personality influence the development of relationships

It's important to build connections when users feel bad

User

User Modeling

Personality inference

Emotion inference

Demographic

Interests

Social

Bot

Personalized Conversation

Conversation in a specific style

Emotional interaction

Proactive interaction

Recommendation

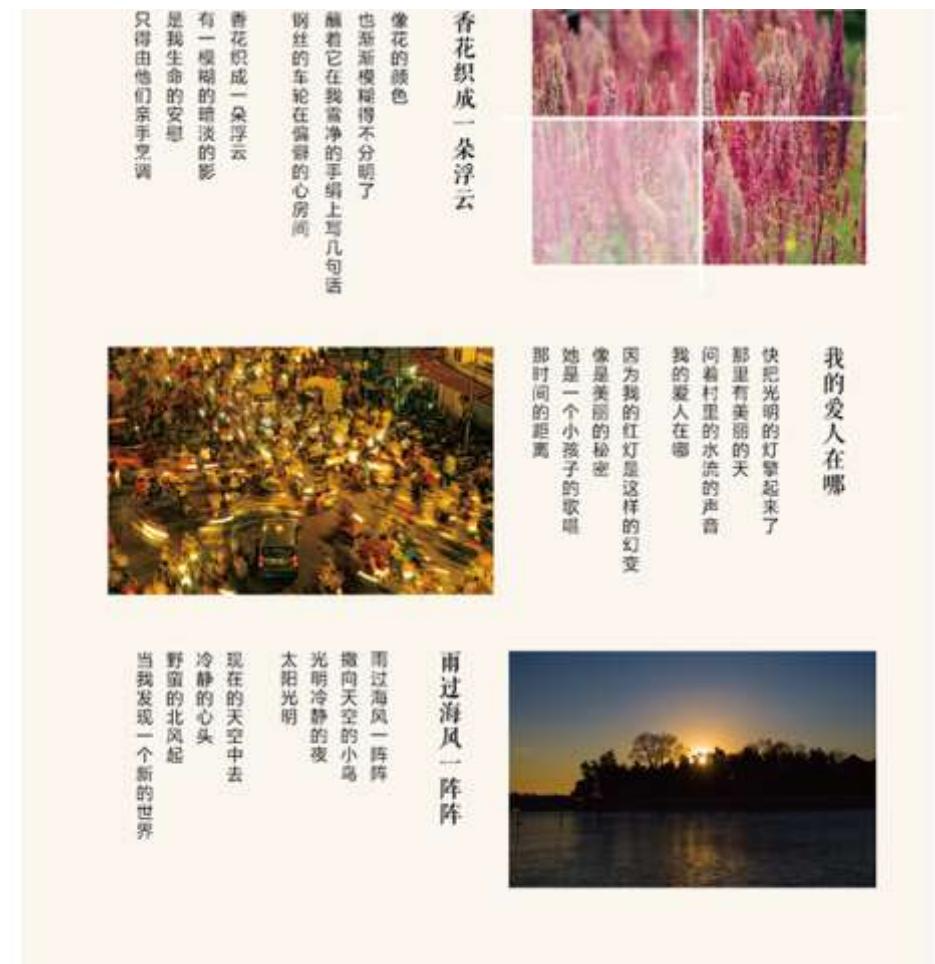
We should make users feel the bot is someone they like

Talk in a style that users prefer

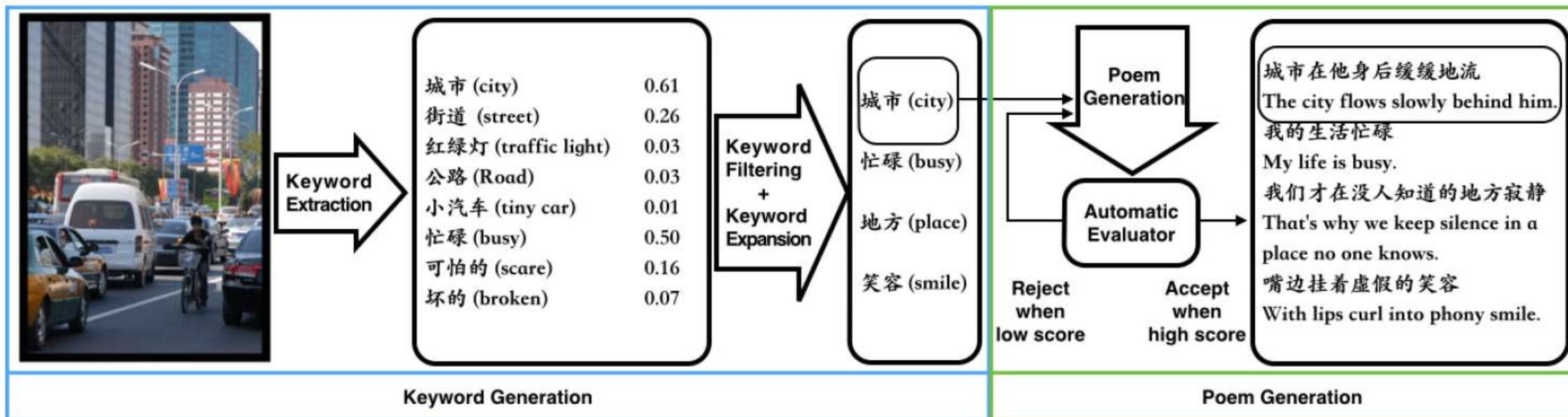
# 个性化文本生成应用：诗人小冰



# 个性化文本生成应用：诗人小冰



# Poem Generation



# Eliza: a Rogerian Psychotherapist

- One of the first chatterbots, written at MIT by Joseph Weizenbaum between 1964 and 1966. (<http://nlp-addiction.com/eliza/>)
- ELIZA effect, the tendency to unconsciously assume computer behaviors are analogous to human behaviors

```
File Edit Options Buffers Tools Help
□ □ × ⌂ ⌂ ⌂ ⌂ ⌂ ⌂ ⌂ ⌂ ? 

I am the psychotherapist. Please, describe your problems.
Each time
you are finished talking, type RET twice.

I have a terrible headache.

Maybe you should consult a doctor of medicine, I am a psychi
satrist.

Okay, I'll do that then.

Can you elaborate on that?

What are you talking about? No I cannot elaborate. See you l
ater.

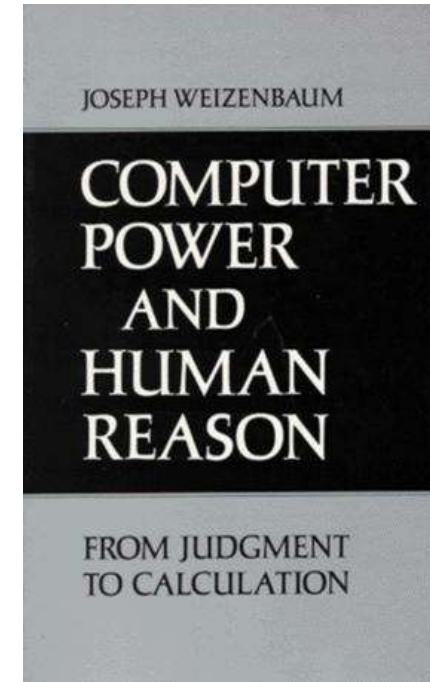
Why do you say that?

Now you're just talking nonsense!

What makes you believe now I am just talking nonsense?

You are not making any sense.

Perhaps your plans have something to do with this.
-u:** *doctor* (Doctor Fill)--L1--Top-----
```

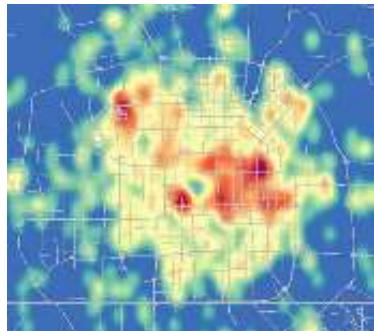


# DiPsy: A Digital Psychologist

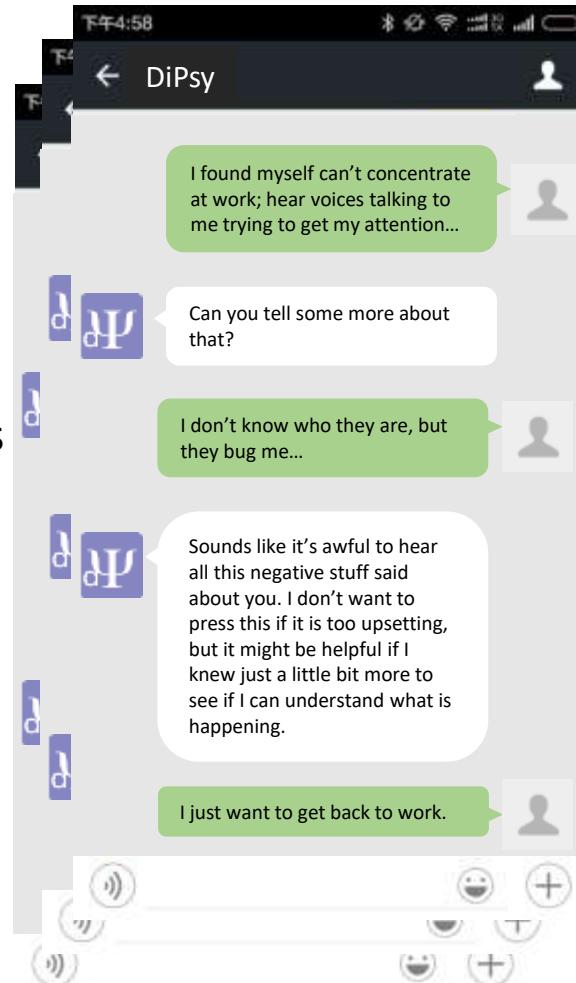
## Digital footprints



## Daily conversations



## Physical movements



Your Myers-Briggs personality type  
ENFP

Your preferences  
Extraversion | Intuition | Feeling | Perceiving

### THE WAY YOU DIRECT AND RECEIVE ENERGY

**Extraversion**  
People who prefer Extraversion tend to direct their energy toward the outside world and get energized by interacting with people and taking action.



**Introversion**  
People who prefer Introversion tend to direct their energy toward their inner world and get energized by reflecting on their ideas and experiences.



### THE WAY YOU TAKE IN INFORMATION

**Sensing**  
People who prefer Sensing tend to take in information that is real and tangible. They focus mainly on what they perceive using the five senses.



**Intuition**  
People who prefer Intuition tend to take in information by seeing the big picture. They focus mainly on the patterns and interrelationships they perceive.



### THE WAY YOU DECIDE AND COME TO CONCLUSIONS

**Thinking**  
People who prefer Thinking typically base their decisions and conclusions on logic, with accuracy and objective truth as the primary goals.



**Feeling**  
People who prefer Feeling typically base their decisions and conclusions on personal and social values, with understanding and harmony as the primary goals.



### THE WAY YOU APPROACH THE OUTSIDE WORLD

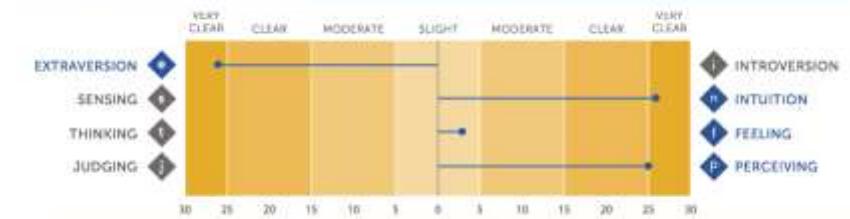
**Judging**  
People who prefer Judging typically come to conclusions quickly and want to move on, and take an organized, planned approach to the world.



**Perceiving**  
People who prefer Perceiving typically look for more information before coming to conclusions and take a spontaneous, flexible approach to the world.

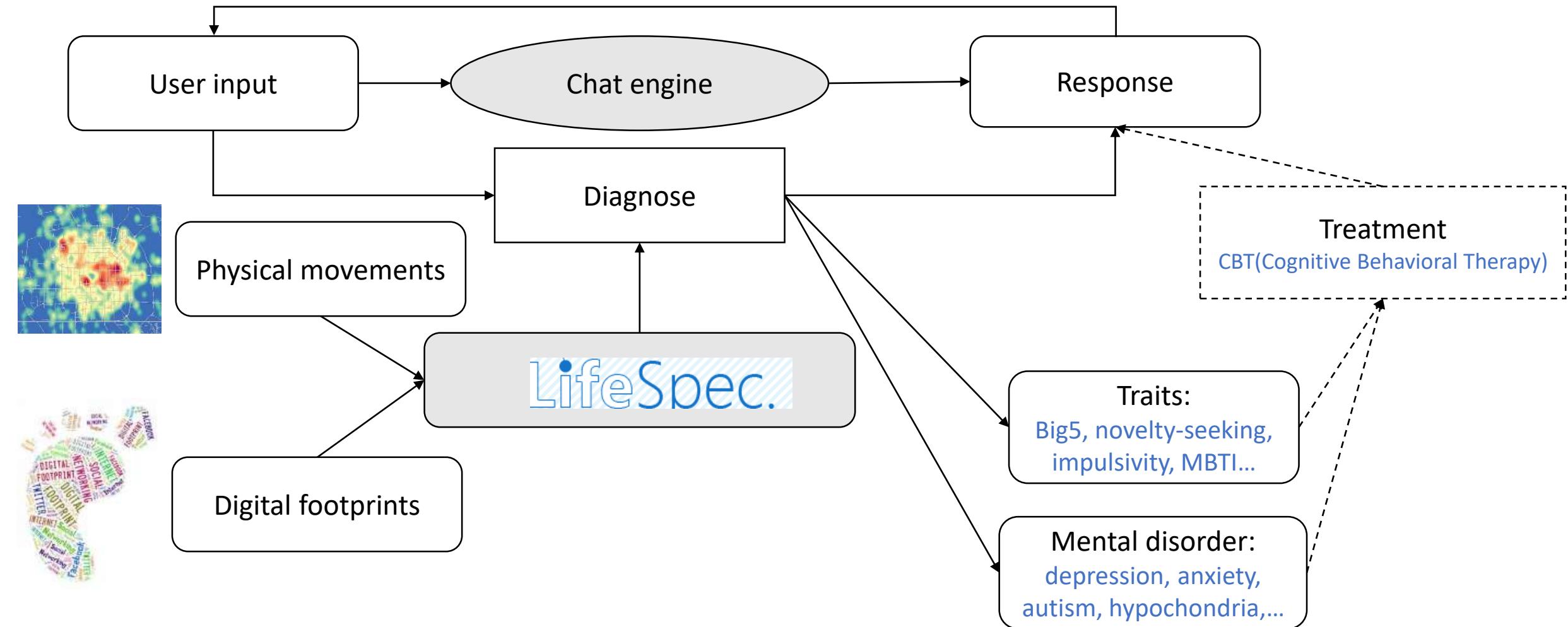


### CLARITY OF YOUR PREFERENCES: ENFP



PCI RESULTS: EXTRAVERSION | 26 INTUITION | 26 FEELING | 3 PERCEIVING | 25

# System Framework



# Summary

- User modeling is important for building intelligent bots
- Mining demographic and personality attributes from human behavioral data
- Collaborating with researchers from areas of psychology, sociology and cognitive science



封面插图来自澳门特别行政区政府新闻局出版的《澳门新报》（2010年1月期）。

教材热线

010-88379604  
010-88379429 88361066  
010-68326294 88379649 68995258

零售热线

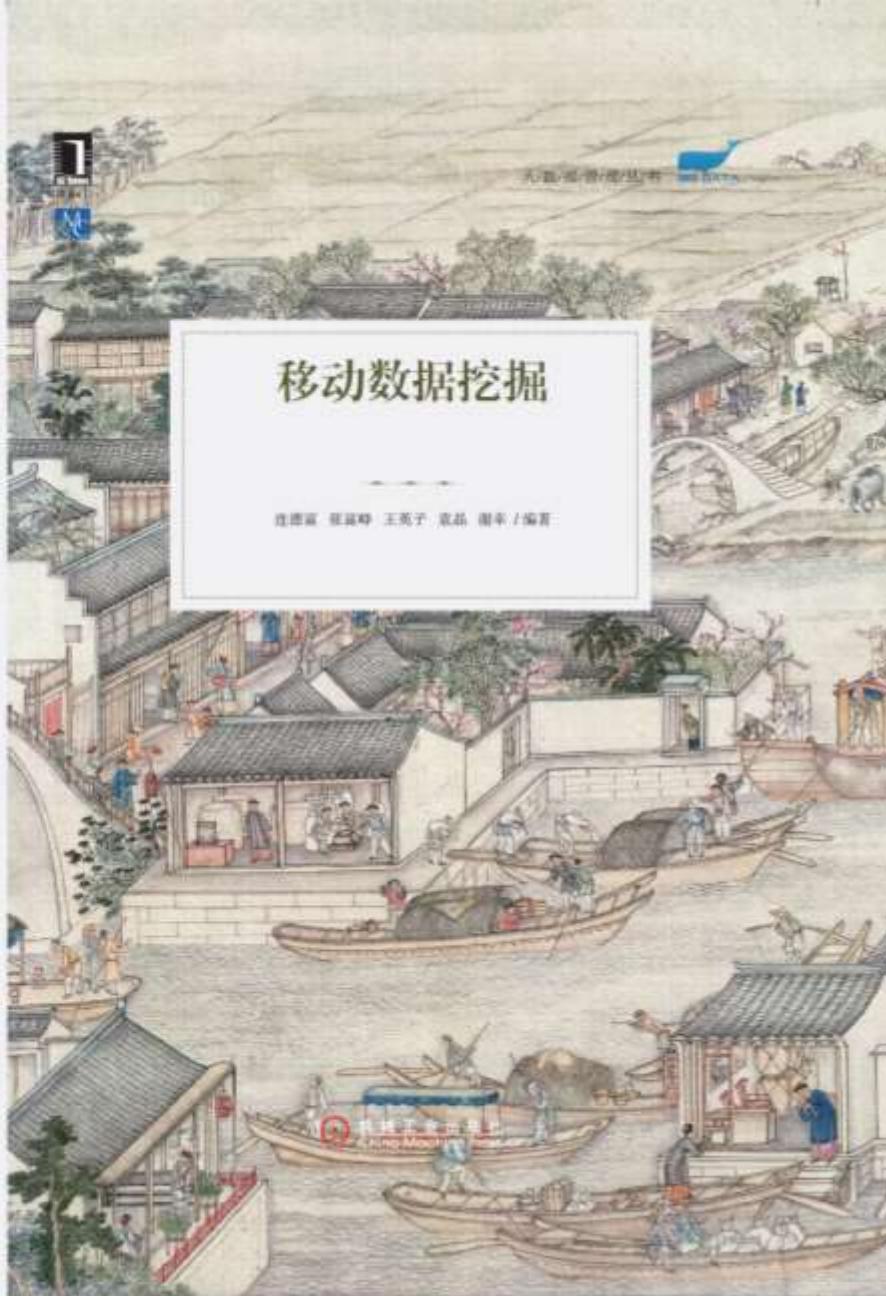
010-88379604  
网上书店：[www.tatabook.com](http://www.tatabook.com)  
数字书店：[www.china-pub.com](http://www.china-pub.com)



机械工业出版社

J  
版

移动数据挖掘



Thanks!