Mining Crowd-Sourced Geo-Tagged Data for Understanding and Sustaining Urban Vibrancy

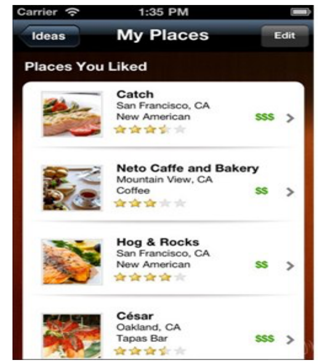
Yanjie Fu

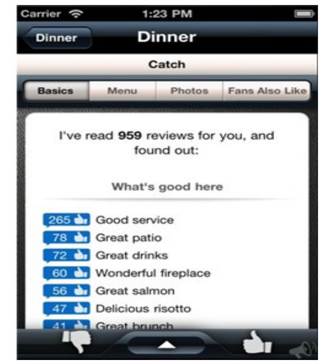
Data Mining in Geo-Mobile Intelligence

Urban Region Level

Spatial and Urban Analytics (transportation analysis, spatial allocation and site selection, etc.)

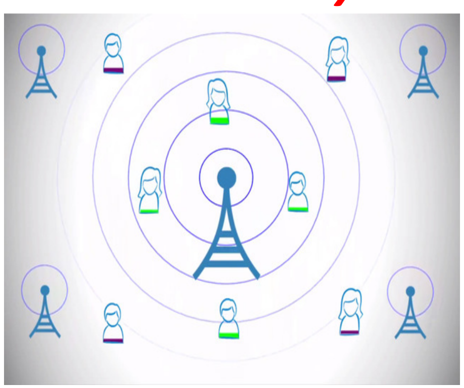
Mobile User Level





Mobile Recommender Systems (restaurant, POI, retweet, etc.)

System and Device Levels



Self-Optimizing Network (SON) and in-App behavior analytics

Outline

- **Background and Motivation**
- Preliminary Analysis
- Modeling Geographic Dependencies
- Exploring Mixed Land Use
- Conclusion and Future Work

The Rise of Consumer Cities

- **Urban Vibrancy: From Production-Centric To Consumer-Centric**
  - Edward L. Glaeser, Urban Economist from Harvard University
  - Glaeser, E. L., Kolko, J., & Saiz, A. (2001). Consumer city. Journal of economic geography, 1(1), 27-50.
  - The future of cities depends demand for density: whether cities are attractive places for consumers to live
  - As firms become more mobile, the success of cities hinges more on cities' role as centers of consumption.

Urban Livability and Vibrancy

- **Construct an urbanity index to measure city amenity and measure willingness to pay for urbanity**
  - Gabriel Ahlfeldt (2013). Urbanity. Working Paper. London School of Economics.
- **Construct a social interaction potential index to measure the face-to-face communication of residents within a community**
  - Steven Farber (2013). Urban sprawl and social interaction potential: an empirical analysis of large metropolitan regions in the United States. Journal of Transport Geography. University of Toronto.

Urban Economics and Social Impact

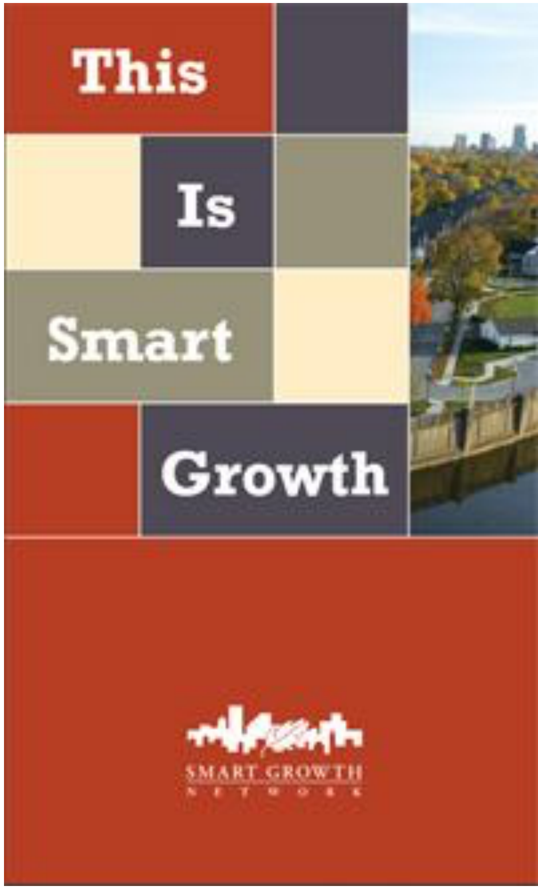
- **Urban structure leads to the spatial concentration of consumer demands and product diversity**
  - Nathan Schiff (2014). "Cities and product variety: evidence from restaurants." Journal of Economic Geography.
- **People are willing to pay higher rents and transport costs for high-density communities with more social interaction and diverse opportunities for consumption**
  - Victor Coutour (2014). Valuing the Consumption Benefits of Urban Density. University of California, Berkeley.

Urban Planning and Governance

- **Live-Work-Play planning strategy can improve business performances of office properties**
  - Yan Song (2014), Does downtown office property perform better in live–work–play centers? UNC Chapel Hill
- **High-density mixed land uses can encourage workability and instant social interaction**
  - Emily Talen (1999). Sense of community and neighborhood form: an assessment of the social doctrine of New Urbanism. Urban Studies.

US Smart Growth White Paper

United States Environmental Protection Agency



What?

"Smart growth" covers a range of development and conservation strategies that help protect our health and natural environment and make our communities more attractive, economically stronger, and more socially diverse.

Why?

EPA works on smart growth issues to help communities develop in ways that are better for health and the environment.

NSF Funded Projects

- **Interaction Potential and the Social and Economic Vibrancy of Metropolitan Regions (2013-2016, PI: Farber S.)**
  - The project's primary goals are to determine which elements of the urban spatial structure restrict or support social interaction potential (SIP) and to quantify the degree to which SIP affects social and economic vibrancy.
  - The researchers will develop a new metric for measuring SIP and will use it to discover how SIP varies within and between metropolitan regions, to determine how spatial structure influences these measurements, and to quantify the intra- and inter-regional socioeconomic outcomes attributable to SIP.
  - This research requires intensive computation and will apply massively parallel computational resources to enhance basic knowledge about urban social and economic processes.

Big Data for Smart Growth

- **Big Crowd-sourced Geo-tagged Data**
  - Mobile devices, e.g., smart phones, POS, wearable devices
  - Vehicles, e.g., taxicabs, buses, subways, city bikes
  - Sensors, e.g., satellite remote sensing
  - Buildings, e.g., banks, shopping malls, restaurants
  - Human in various location based services, e.g., Foursquare.com, Weibo.com, Flickr.com, Tweeter.com, Facebook.com
- **Static and dynamic data**
  - Static Urban Geography
  - Dynamic Human Mobility

Urban Geography Data

- **Urban geography data are a set of geographic characteristics of a city including**
  - road networks, public transportation (bus stops, subways)
  - points of interest (POIs), regional functions

Road Networks

Public Transportation

POI code	POI category	POI code	POI category
1	car service	16	banking and insurance service
2	car sales	17	corporate business
3	car repair	18	street furniture
4	motorcycle service	19	entrance/bridge
5	café/tea Bar	20	public utilities
6	sports/stationery shop	21	chinese restaurant
7	living service	22	foreign restaurant
8	sports	23	fastfood restaurant
9	hospital	24	shopping mall
10	hotel	25	convenience store
11	scenic spot	26	electronic products store
12	residence	27	supermarket
13	governmental agencies and public organizations	28	furniture building materials market
14	science and education	29	pub/bar
15	transportation facilities	30	theaters

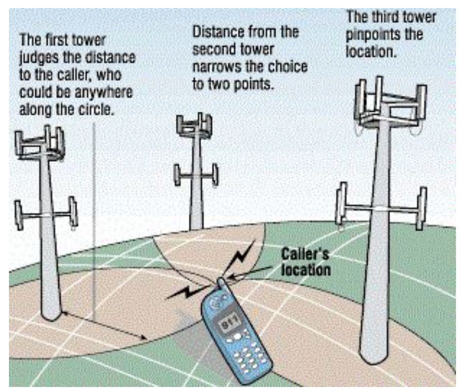
Point of Interests

- **Illustrate the spatial structure (e.g., infrastructure, facilities, geo-presentation) of a community**

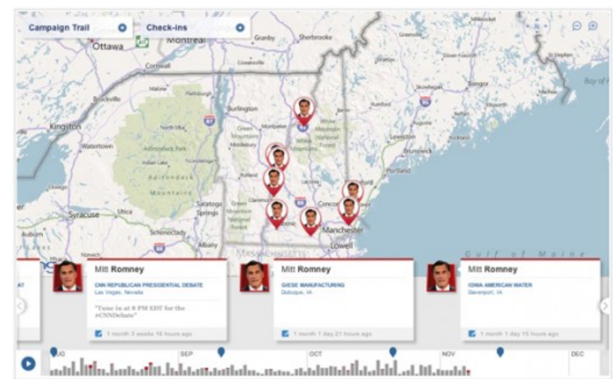
Human Mobility Data

- Human mobility data are people's movement trajectories which can be
  - phone traces or trajectories of driving routes (taxicab, bus)
  - a sequences of posts (like geo-tweets, geo-tagged photos, or check-ins)

Taxicab GPS Traces



Phone Traces



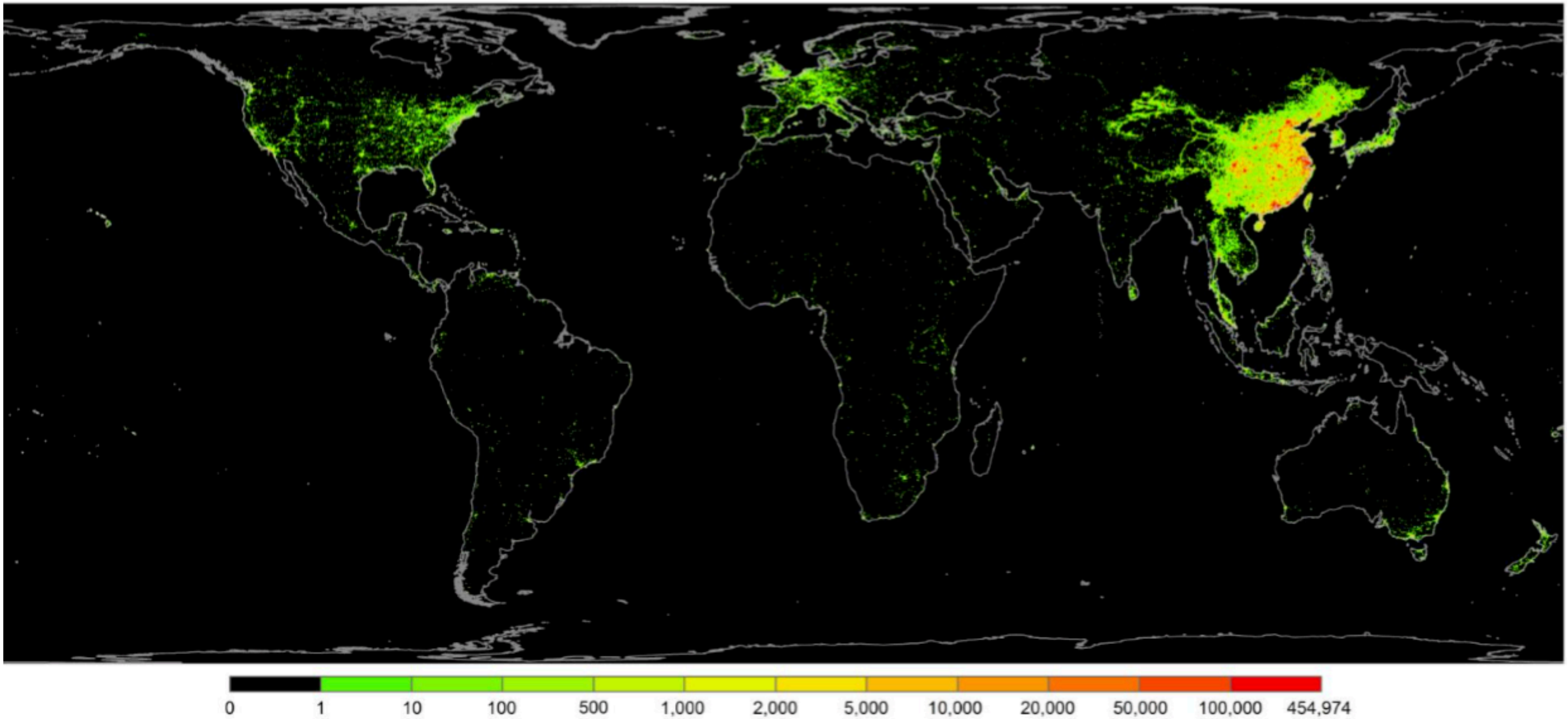
Mobile Checkins

- Encode the social interaction within a community and across communities

Urban Vibrancy in World: Facebook

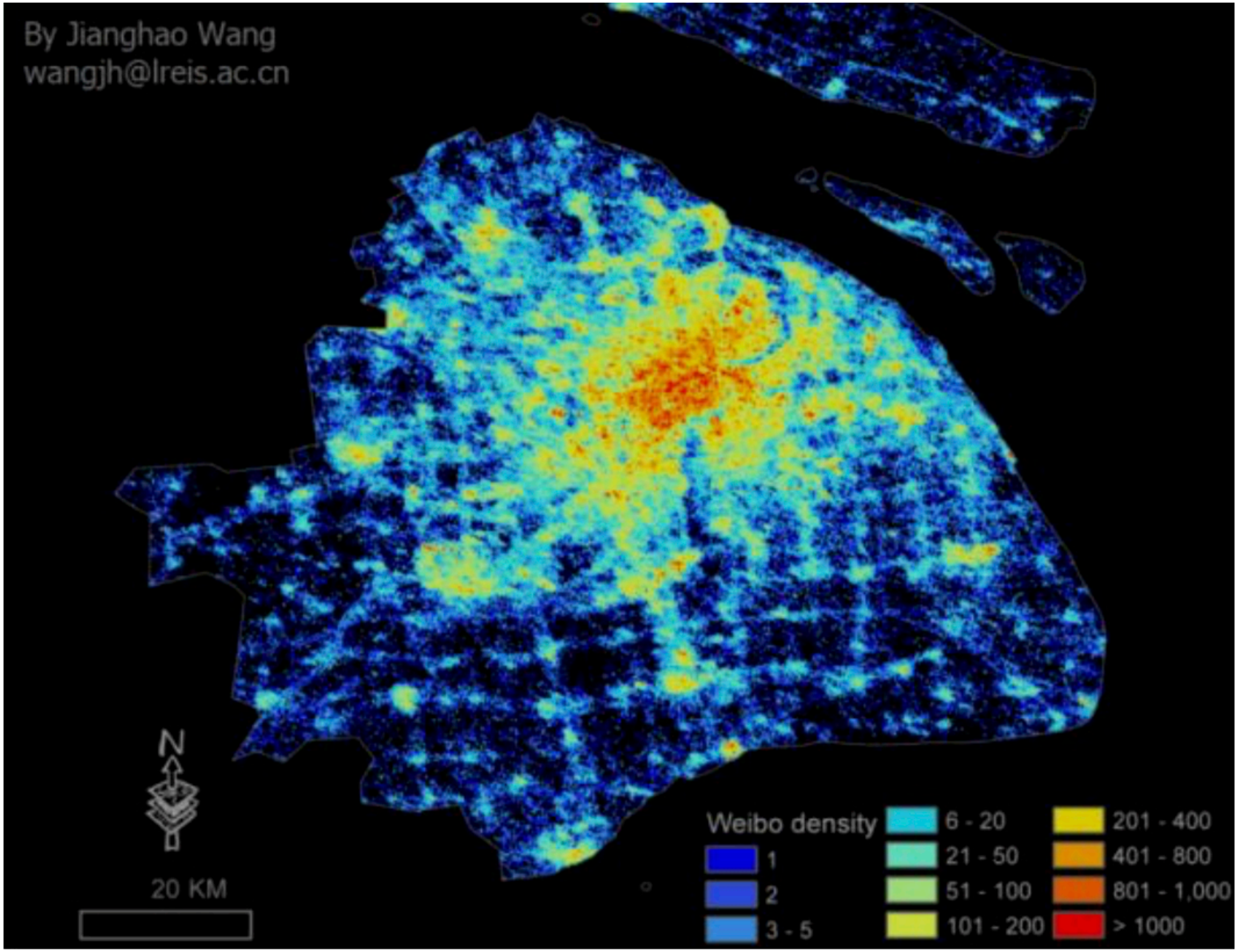
Social interaction density from www.facebook.com

Urban Vibrancy in World: Weibo

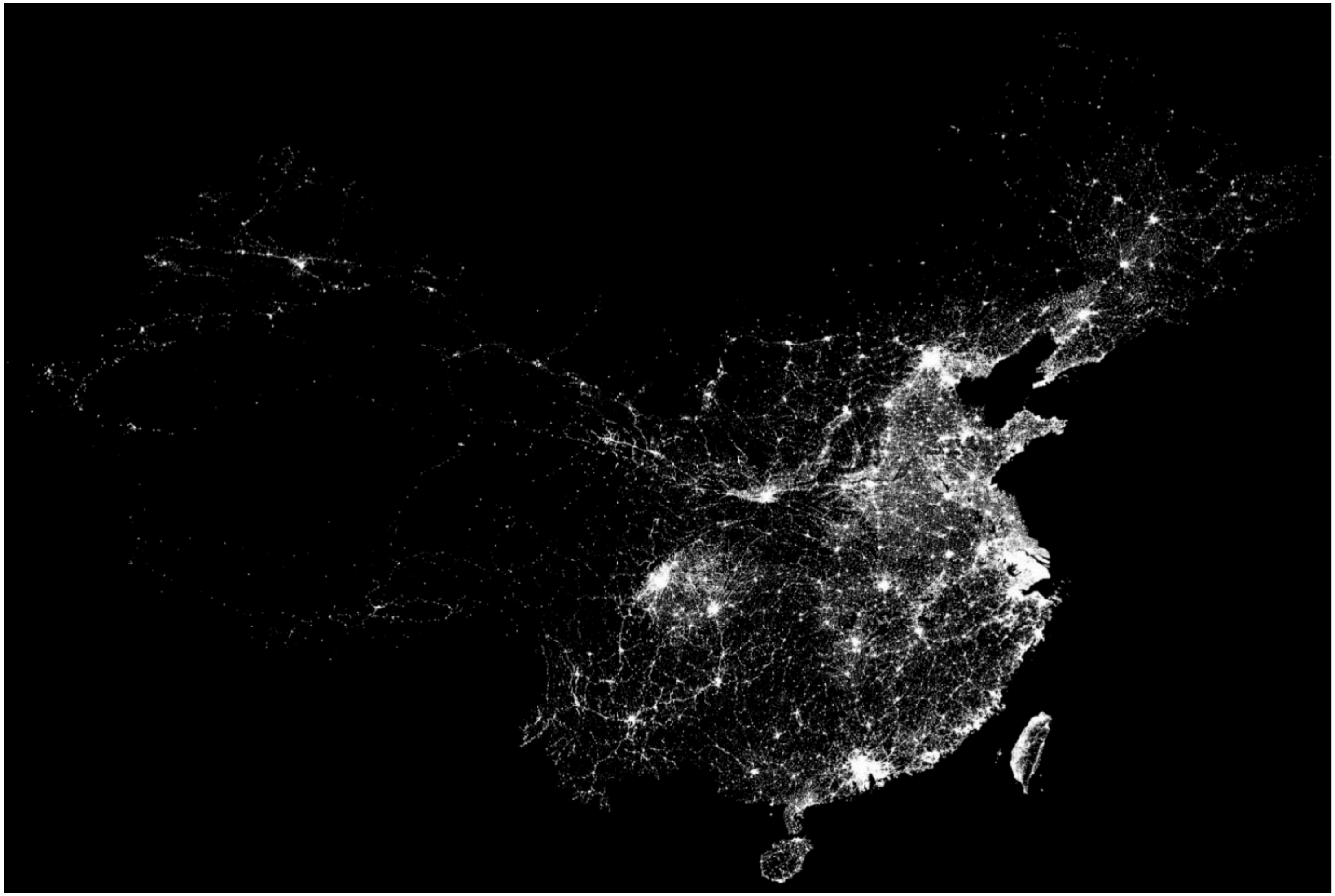


Weibo usage density in the world from www.weibo.com

Weibo Hotspots in Shanghai

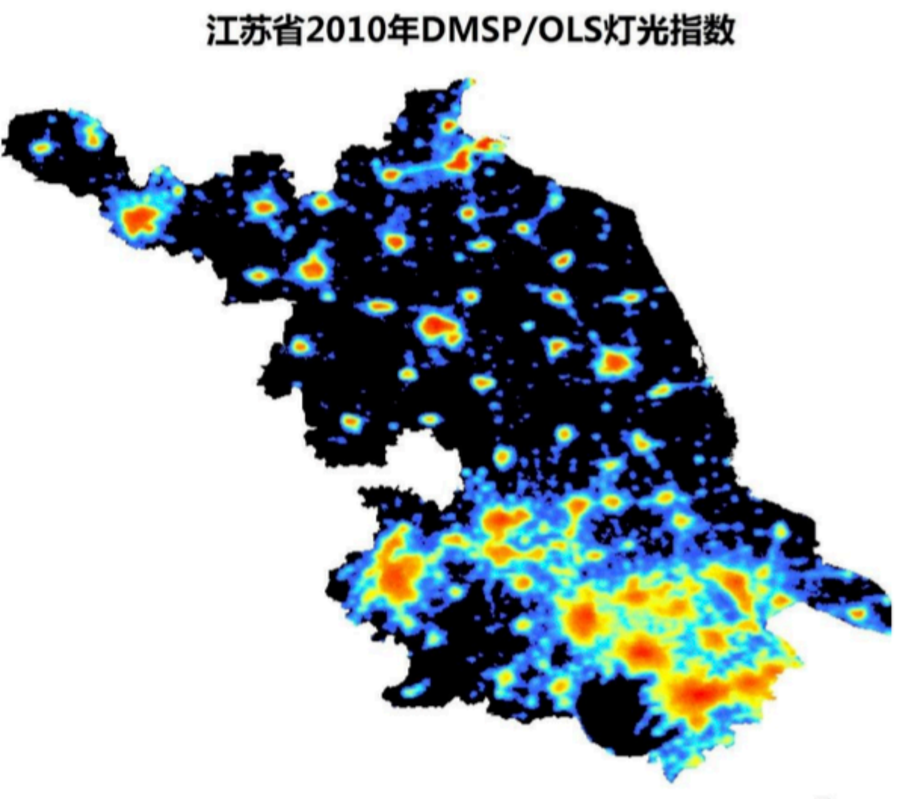


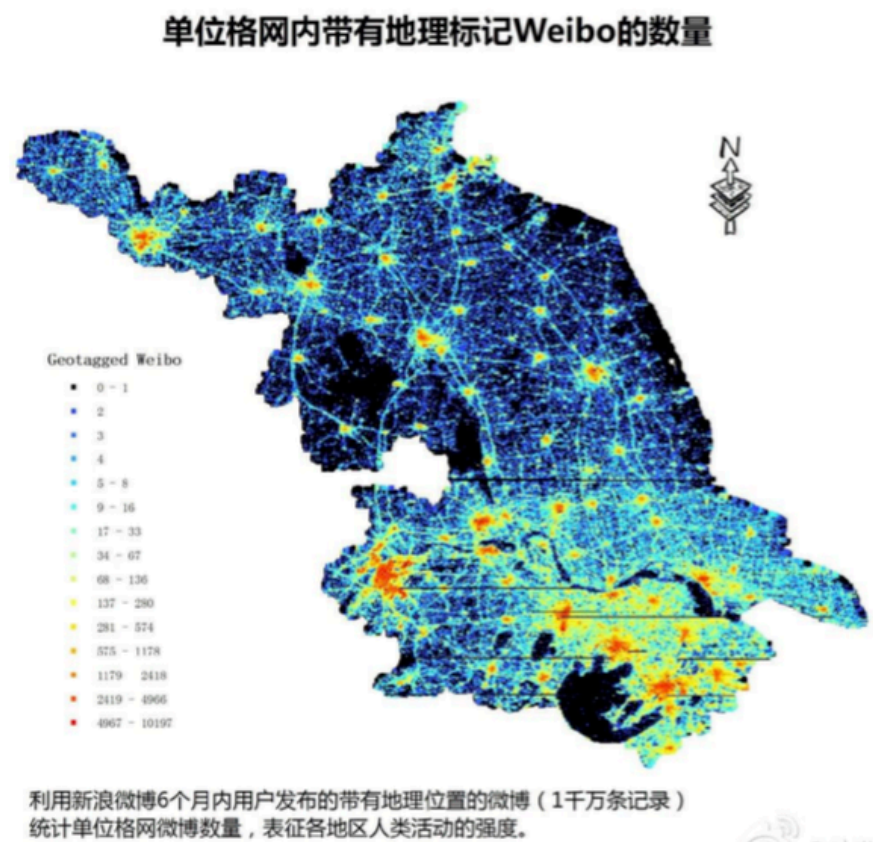
Urban Vibrancy in China



POI checkin data from www.dianping.com

City Light and Mobile Checkins





High similarity between city light distribution and Weibo Checkin distribution

Indicators of Vibrant Communities

- **Spatial Characters**
  - Walkable
  - Dense
  - Compact
  - Diverse
  - Accessible
  - Connected
  - Mixed-use
- **Socio-economic Characters**
  - **Willingness To Pay (WTP)**
  - Intensive social interactions
  - Attract talented workers and cutting-edge firms

Our Research Thrusts

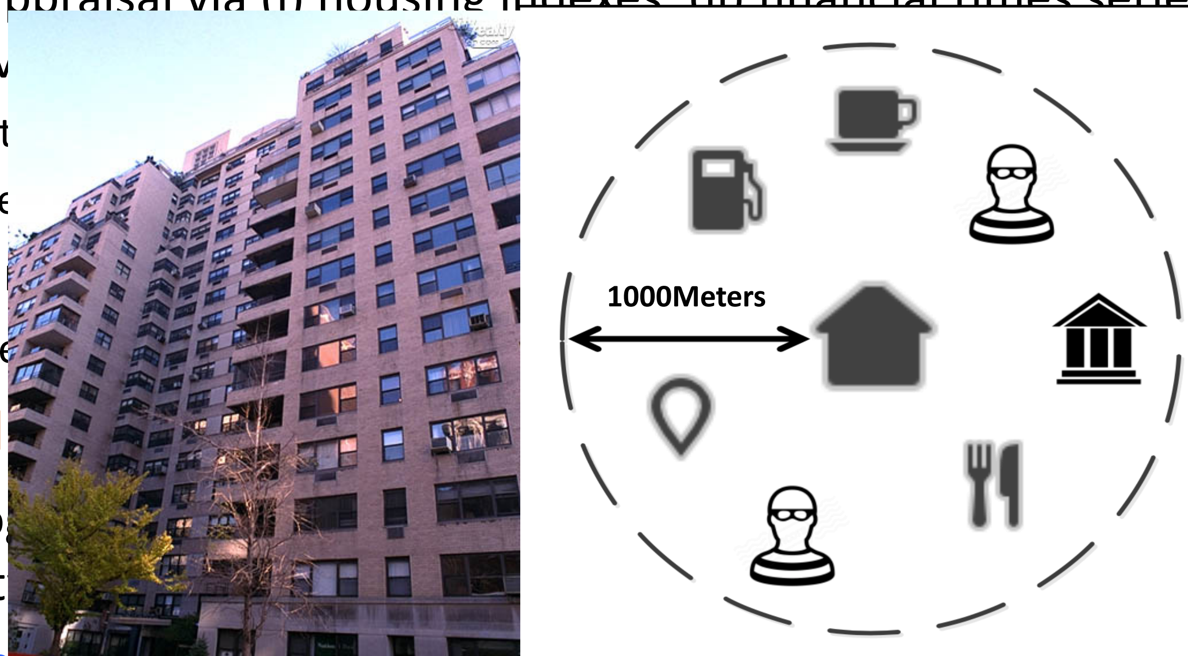
- **Measurement**
  - An alternative index for unmeasurable urban vibrancy: willingness to pay
- **Patterns**
  - Spatial structure character from urban geography
  - Social interaction characters from human mobility
- **Mechanism**
  - How to develop effective ranking systems for identifying high-rated communities with high willingness to pay?
  - What are the underlying drivers for vibrant and sustainable communities?

Outline

- Background and Motivation
- **Preliminary Analysis**
- Modeling Geographic Dependencies
- Exploring Mixed Land Use
- Conclusion and Future Work

Real Estate Ranking

- Prior literature
  - Market price appraisal via (i) housing indexes, (ii) financial times series analysis, (iii) ev… …uct)
    - ***DON'T*** evaluat…
    - ***DON'T*** provide…
  - Learning To Ra…
    - ***DON'T*** conside…
- Revolution in I…
  - Big and hetero… human mobilit…
  - Tobler's first la… *but ne…*
- Real Est…
  - Rank real estate (i.e., residential complexes in big cities) with urban geography, human mobility, and Point of Interests (POIs) data



a residential complex != a single-family house
a residential complex = an apartment building + a neighboring circle area which provides diverse urban functions

Research Challenges

- **Application Challenge**
  - Prior literature
    - ***MOSTLY*** consider prices, coarse-grained location info (e.g., zip code, school area), apartment info (e.g., construction year)
    - ***DON'T*** consider fine-grained urban geography with GPS locations and dynamic human mobility data
  - Location! Location! Location!
    - We are the first to bring in fine-grained urban geography and dynamic mobility data
- **Modeling Challenge**
  - Once we bring in urban geography and human mobility, these data make the modeling difficult
    - **How to combine ranking with geographic dependencies?**
    - **How to combine ranking with mobility patterns?**

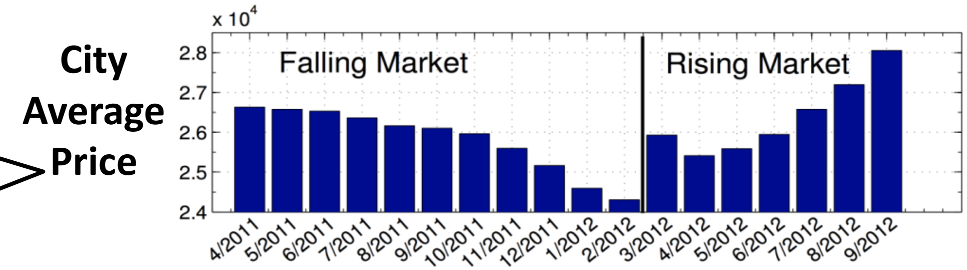
Quantifying Community Ratings

- Investment return rate over a holding period

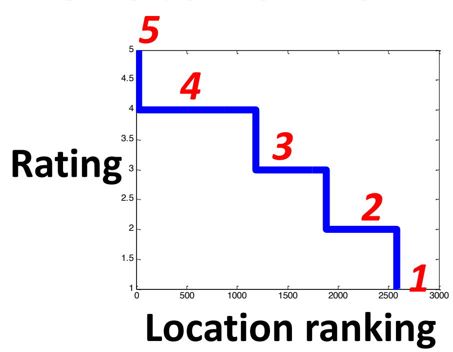
$$r = \frac{P_f - P_i}{P_i} \quad (P_f\text{: final sale price, } P_i\text{: initial sale price})$$

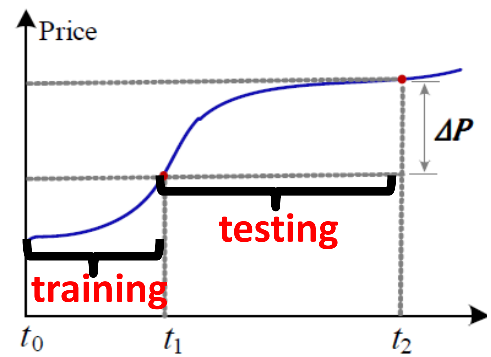
- Rising market and falling market periods

We compute **monthly-based average per-square-feet** prices from transactions for each residential complex



- Segmenting return rates into location ratings for training



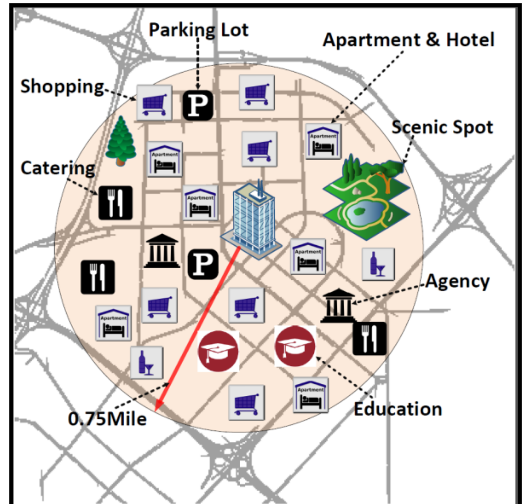


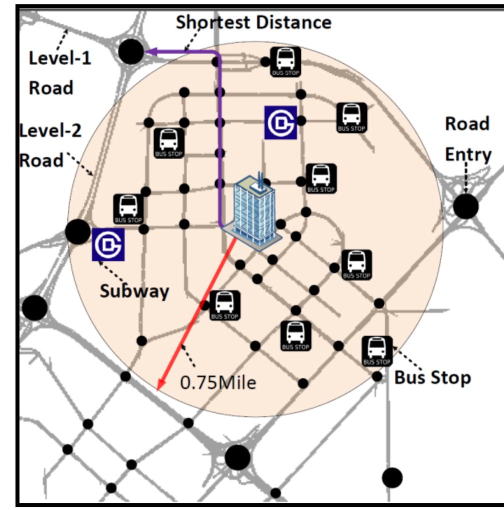
  - Identify rising/falling market periods
  - Calculate the investment returns of each residential complex
  - Segment and grade locations into ratings (5>4>3>2>1) in rising/falling markets

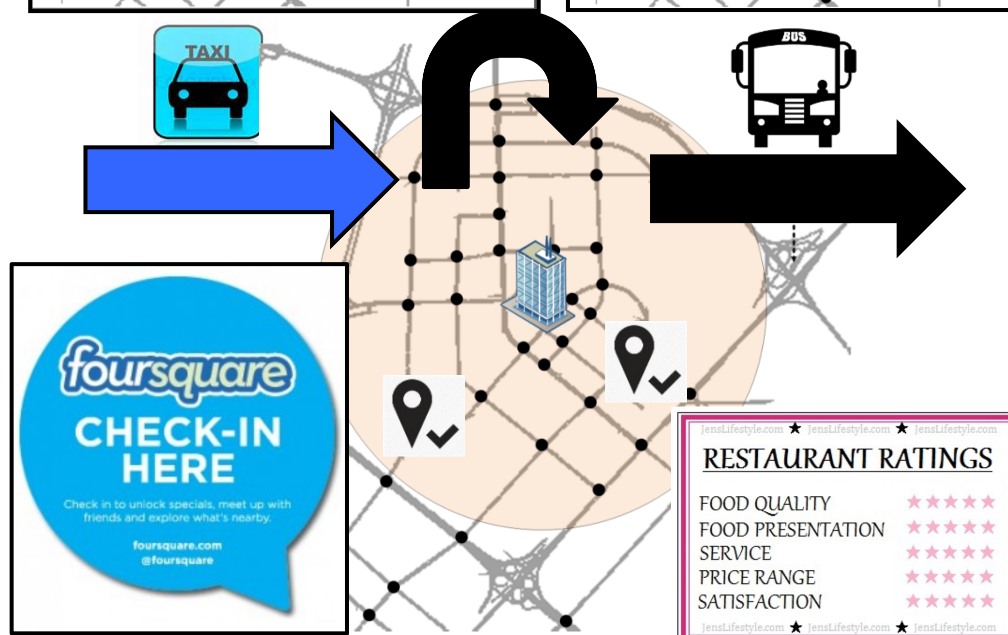
Feature Extraction

A neighborhood is defined as a cell area with radius of 1KM

- Features of urban geography
  - Number of bus stops, subway stations, road networks, POIs
  - Walking distance to bus stops, subways, road networks, POIs
- Features of human mobility
  - Arriving volume, leaving volume, transition volume, driving velocity, trajectory distance of taxies and buses
- Features of customer reviews
  - Overall, service, and environment ratings
  - Number of checkin events
  - Topical profile of checkin comments





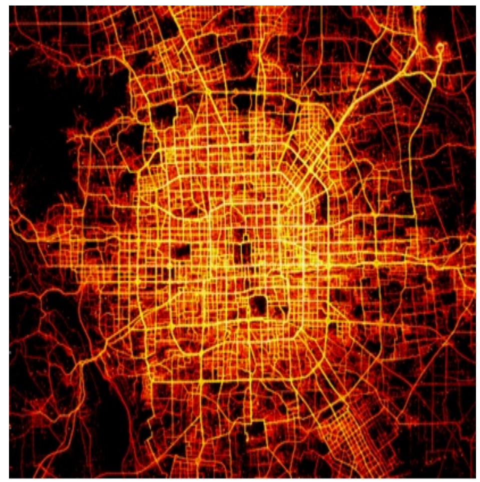


Our Modeling Objective

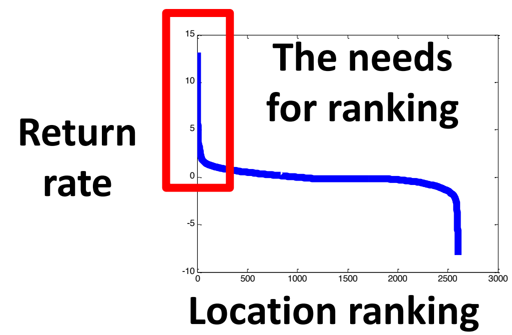
Urban Geography Human Mobility Location Rankings

f(

,



) =

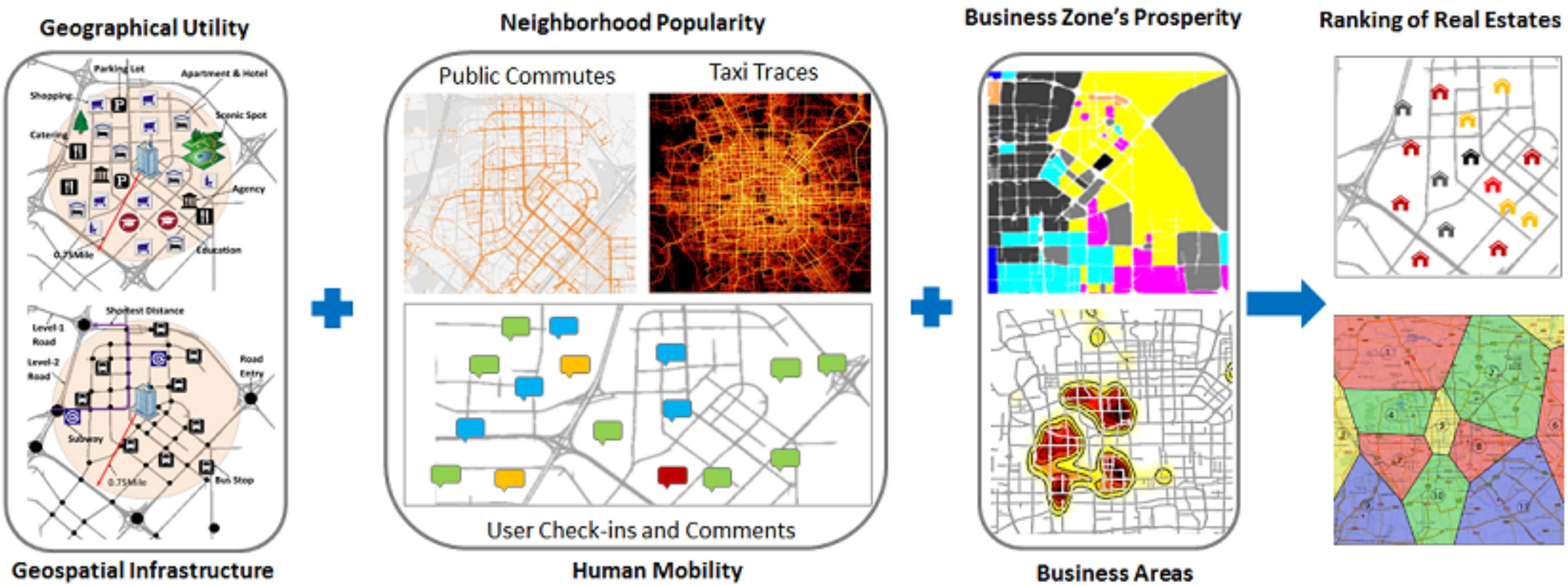


Identify locational insights for developing vibrant and sustainable communities

Outline

- Background and Motivation
- Preliminary Analysis
- **Modeling Geographic Dependencies**
- Exploring Mixed Land Use
- Conclusion and Future Work

Predictors of Location Ratings



Three predictors

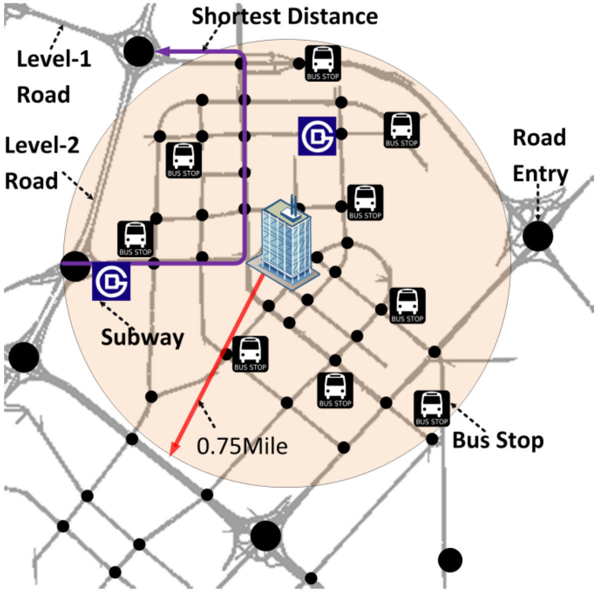
- Geographic utility (land uses)
- Neighborhood popularity (human mobility)
- Influence of business area (business potential)

(1) Daniel Baldwin Hess, Tangerine Maria Almeida. 2007.
(2) Robert Cervero, Chang Deok Kang. 2011.
(3) Montanari, Armando, Barbara Staniscia. 2012.
(4) Hur, Misun, Hazel Morrow-Jones. 2008.
(5) Hj. Mar Iman al Murshid, Abdul Hamid. 2008.

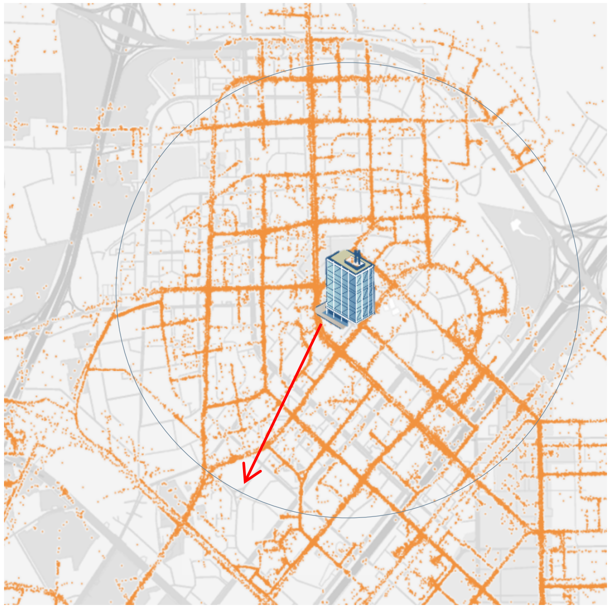
Geographic Dependencies (1)

A neighborhood is defined as a cell area with radius of 1KM

Point of Interest



Transportation Accessibility

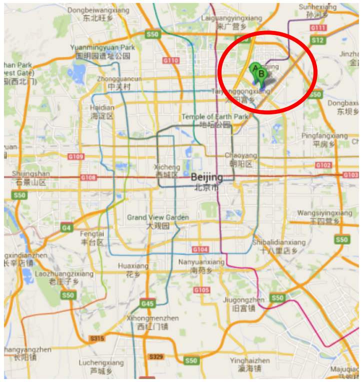


Taxi Mobility

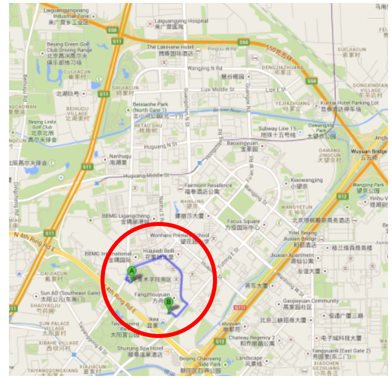
Individual dependency

- The rating of a residential complex is determined by the geographic characteristics of its own neighborhood

Geographic Dependencies (2)



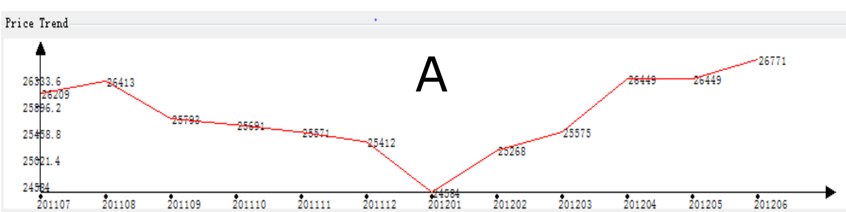
Zoom In

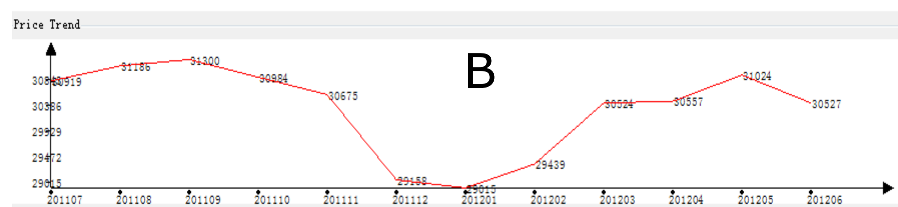


Zoom In

Comparison of locational characteristics

Attribute	A	B
Distance to Level2 road network	156 meters	143 meters
Distance to subway station	1385 meters	1585 meters
#Restaurants	3	4
#Transportation facilities	8	8

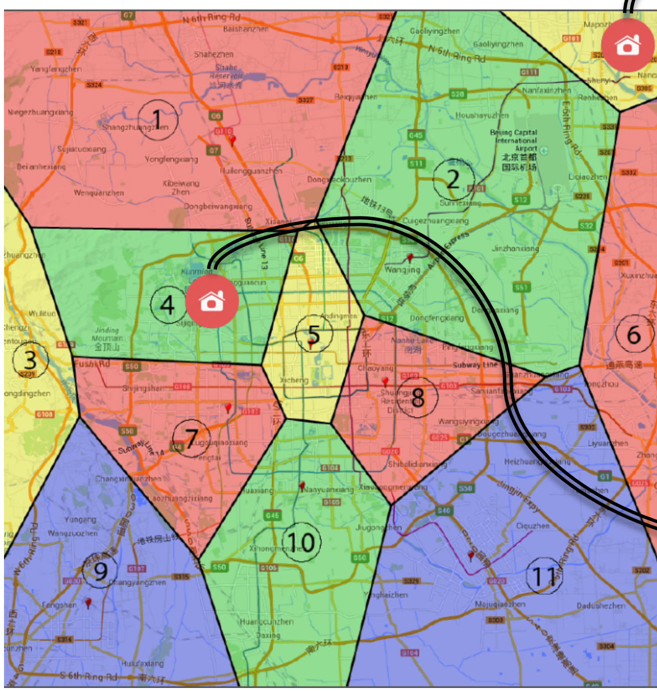




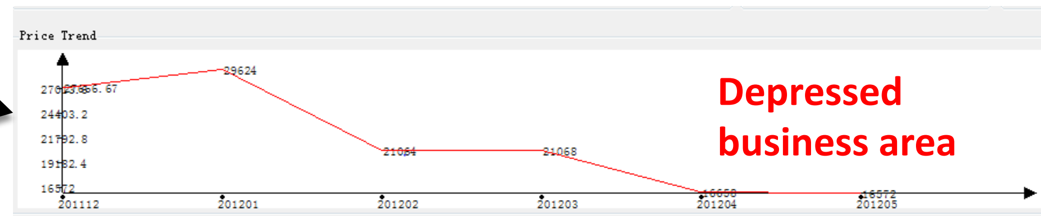
Peer dependency

- Inside a business area, the location rating can be reflected by its nearby residential complexes

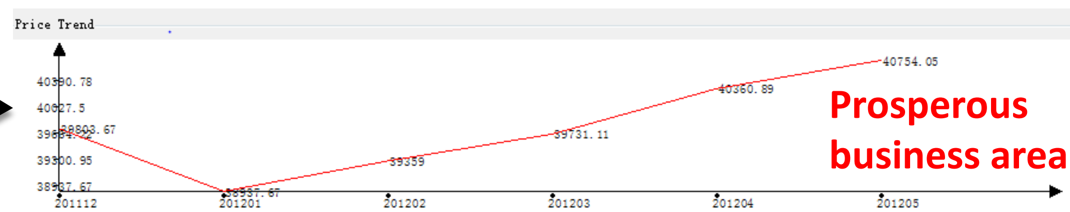
Geographic Dependencies (3)



Rising Market Period
(02/2012-05/2012)



Average regional prices



Average regional prices

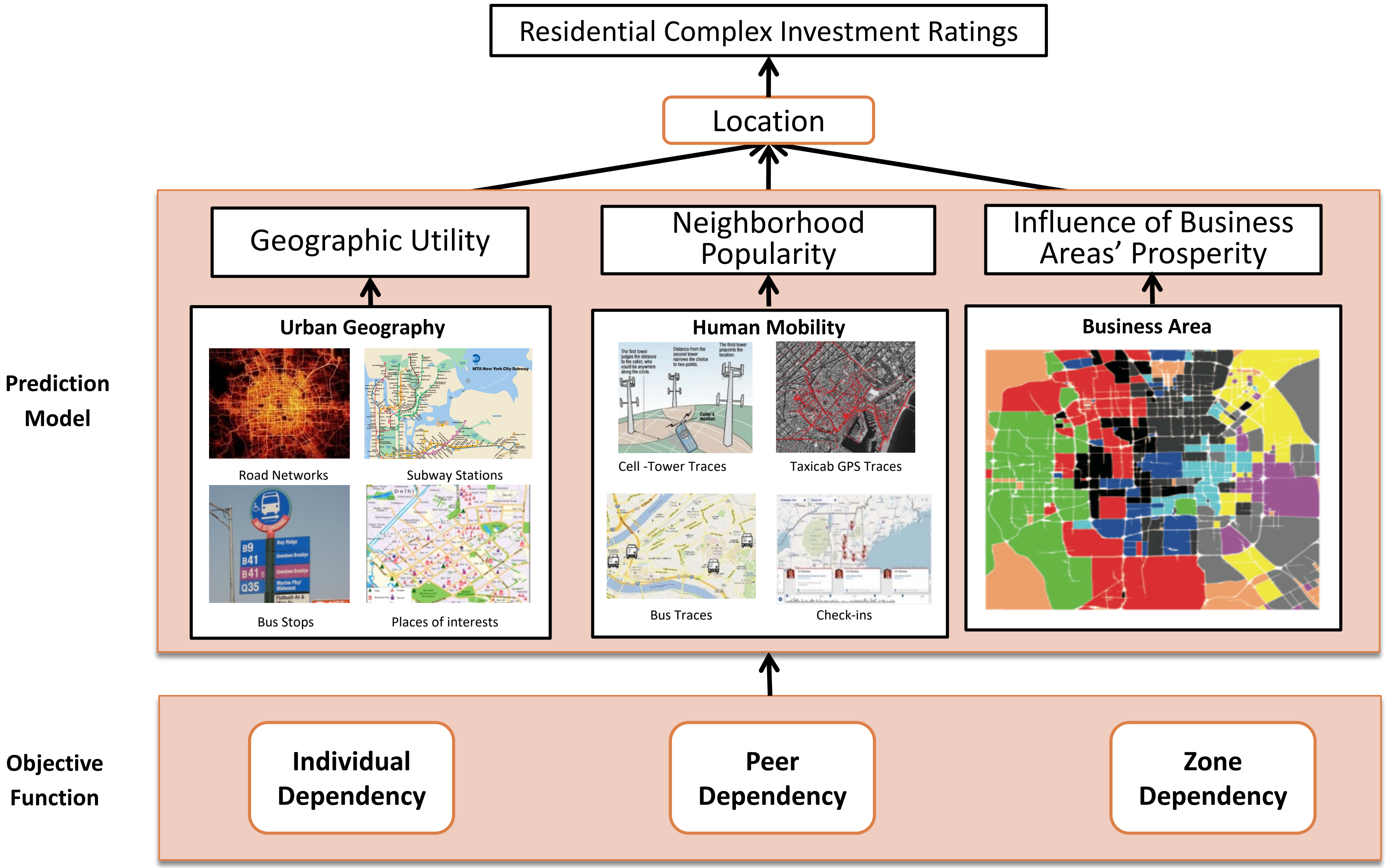
Zone dependency

- The rating of a residential complex can also be influenced by the prosperity of its associated business area

Problem Definition

- **Given**
  - Residential complexes with locations and ratings
  - Urban geography (e.g., POIs, road networks, etc.)
  - Human mobility (e.g., taxi GPS traces)
- **Objective**
  - Rank and classify residential complexes based on their ratings
- **Core tasks**
  - Extract and combine **geographic utility, neighborhood popularity,** and **influence of business areas** to predict ratings
  - Jointly model **individual, peer,** and **zone** dependencies as an objective function to learn a ranking system

Overview of ClusRanking



Modeling Location Rating (1)

- Geographic Utility
  - Feature extraction by spatial indexing (Rtree, Grid index)
  - Linearly combine geographic features of each residential complex to geographic utility

Data Source	Feature Design
Transportation	Number of bus stop
	Walking distance to bus stop
	Number of subway station
	Walking distance to subway station
	Number of road network entries
	Walking distance to road network entries
POIs	Number of POIs of different POI categories

Log norm for count data

$$P[c_i] = \frac{\#_i}{\sum_{i=1}^{|C|} \#_i} log \frac{|H|}{|\{h|c_i \in h\}|}$$

TF-IDF norm for doc-word data

Neighborhood Profiling (a neighborhood is defined as a cell area with radius of 1KM)

Modeling Location Rating (2)

- **Influence of business areas (A generative view)**
  - There are ***K*** business areas in a city
  - Each business area is a cluster of residential complexes

l: the latlng of a complex i

$$l_i \sim \mathcal{N}(\mu_r, \Sigma_r)$$

μ: the center (latlng) of a business area r
Σ: the covariance of lat and lng

  - The more prosperous, the easier we identify a high-rated residential complex from a business area

r: the business area assignment of a complex

$$r \sim \text{Multinomial}(\eta)$$

η: the prosperities of K business areas

  - K business areas are K spatial hidden states; their business prosperities can inversely show influence on residential complexes in terms of geo-distance

ρ: the influence of business area prosperities

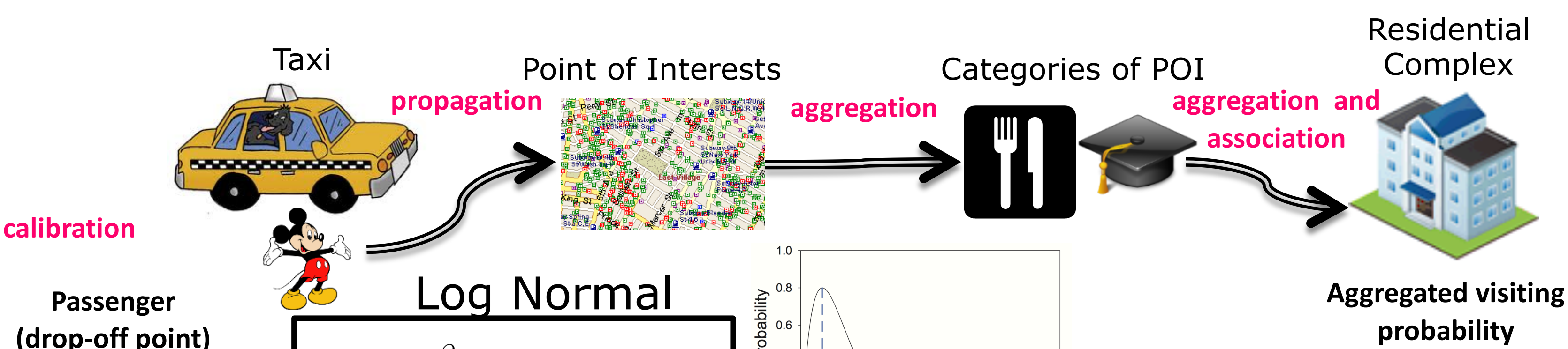
$$\rho_i = \sum_{k=1}^{K} \left(\frac{d_0}{d_0 + d(i, r_k)}\right)^e \frac{\eta_k}{\sum_{k=1}^{K} \eta_k}$$

d_0/[d_0+d(i, r)]: the influence is inversely proportional to distance

Gaussian Mixture Model + Learning To Rank

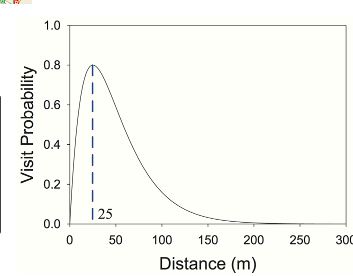
Modeling Location Rating (3)

- **Neighborhood Popularity (A propagation view)**
  - Propagate visit probability to POIs per drop-off point
  - Aggregate visit probability per POI
  - Aggregate visit probability per POI category
  - Compute popularity score
- **Spatial propagation and aggregation from taxi to residential complex**



Log Normal

$$P(x) = \frac{\beta_1}{\beta_2} \cdot x \cdot exp(1 - \frac{x}{\beta_2})$$

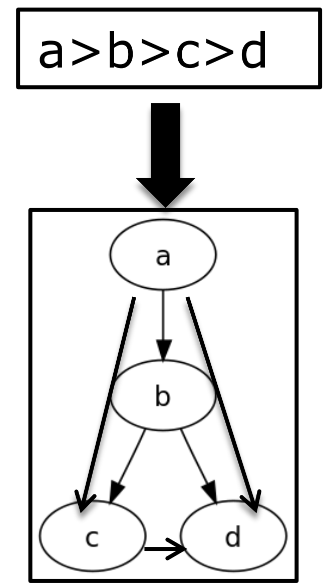


Modeling Three Dependencies

- **Individual Dependency (Lik_{id}: point-wise analysis)**
  - Model the accuracy of predicting observed data, e.g., investment ratings, locations, and business area assignments
  - Maximize the likelihood ≈ minimize square loss

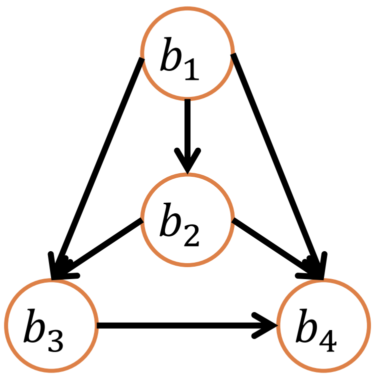
$$Lik_{id} = \prod_{i}^{I} P(\{y_i, l_i, r_i\}|\Psi, \Omega) = \prod_{i}^{I} N(y_i|f_i)N(l_i|u, \sigma)Multi(r_i|\eta)$$

- **Graph Representation of Rankings**
  - A ranked list of estates is viewed as a directed graph
  - A node ≈ a residential location
  - A directed edge *a* → *b* ≈ *a* ranks higher than *b*
  - Our model generates edges with certain probability
  - Maximizing the likelihood ≈ minimizing the ranking loss of graph-based ranking structure



Modeling Three Dependencies

- **Peer Dependency (Lik_{pd}: pairwise analysis on residential complex level)**
  - Consider a ranked list of residential complexes: $b_1 > b_2 > b_3 > b_4$

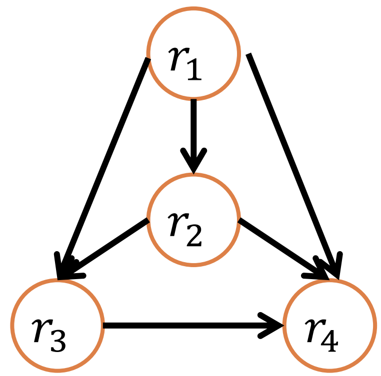


  - Maximize the ranking consistency of residential complex pairs
  - ≈ Maximize the likelihood of edges of complex-level ranking graph

$$Lik_{pd} = \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(i \rightarrow h|\Psi, \Omega)^{I(r_i = r_h)} = \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} \left[\frac{1}{1 + \exp(-(f_i - f_h))}\right]^{I(r_i = r_h)}$$

Modeling Three Dependencies

- **Zone Dependency (Lik_{zd}: pairwise analysis on area level)**
  - Map the graph of residential complex rankings $b_1 > b_2 > b_3 > b_4$ to the graph of business area rankings: $r_1 > r_2 > r_3 > r_4$



  - Maximize the ranking consistency of corresponding business area pairs
  - ≈ Maximize the likelihood of edges of area-level ranking graph

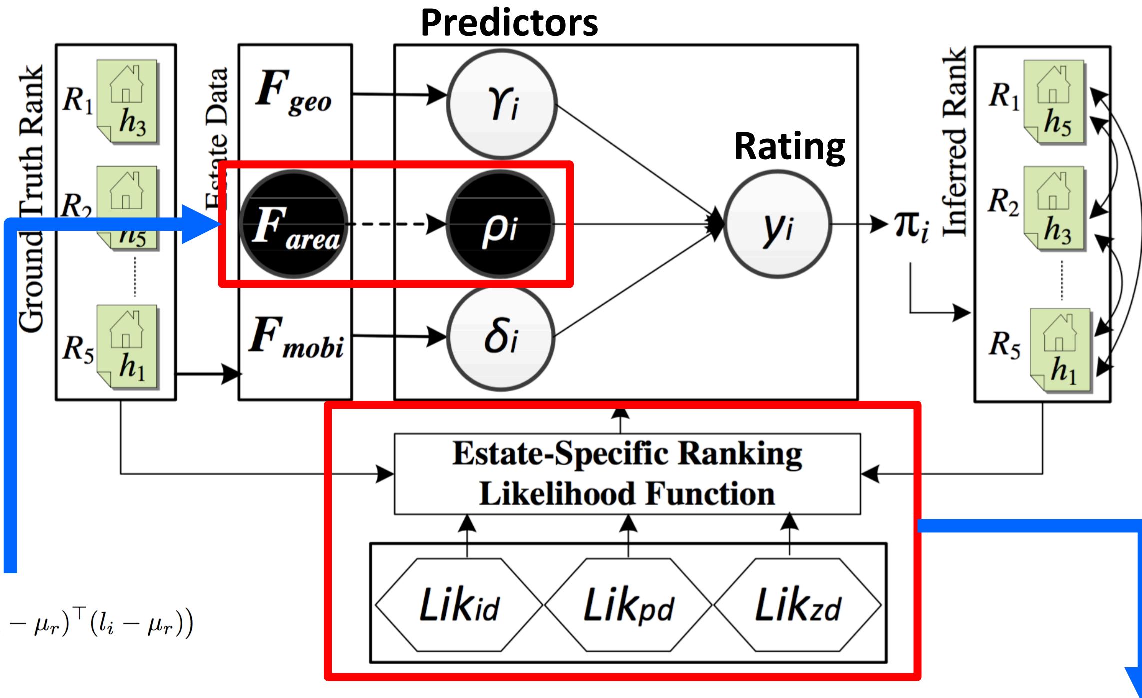
$$Lik_{zd} = \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(r_i \rightarrow r_h | \Psi, \Omega)^{I(r_i \neq r_h)} = \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} \left[\frac{1}{1 + \exp(-(\eta_i - \eta_h))}\right]^{I(r_i \neq r_h)}$$

- **By Bayesian inference, the posterior is**

$$P(\mathcal{D}|\Psi, \Omega) = Lik_{id} \times Lik_{pd} \times Lik_{zd}$$

Solving the Co-Training

The co-training of Geo-Clustering and Multi-View Learning-To-Rank via EM mixed with a sampling (MCEM)



$$\eta_r^{(t+1)} = \frac{C_r^{(t+1)} + \gamma_r}{C^{(t+1)} + \sum_r \gamma_r}$$

$$\mu_r = \frac{1}{\#(i,r)} \sum_{i=1}^{I} \mathbb{I}(r_i = r) l_i,$$

$$\Sigma_r = \frac{1}{\#(i,r) - 1} \sum_{i=1}^{I} \mathbb{I}(r_i = r) \left((l_i - \mu_r)^\top (l_i - \mu_r) \right)$$

E-step: update the latent business area assignment by maximizing the posterior of r via sampling

M-step: maximize the three dependencies by gradient decent

$$r \sim P\left(l_i | r, \Psi^{(t)}\right) P\left(\{Y, \Pi\} | r, \Psi^{(t)}\right) P\left(r | \eta^{(t)}\right)$$

r is updated by the location emission probability, the ranking consistency, and the prosperities of multiple areas

$$\mathcal{L}(q, W | R^{(t+1)}, \mathcal{D}) = \sum_{i=1}^{I} \left[-\frac{1}{2} ln\sigma^2 - \frac{(y_i - f_i)^2}{2\delta^2} \right] + \sum_{i=1}^{I-1} \sum_{h=i+1}^{I} ln \frac{1}{1 + exp(-(f_i - f_h))} \mathbb{I}(r_i = r_h)$$
$$+ \sum_{m=1}^{M} \left[-\frac{1}{2} ln\sigma_q^2 - \frac{(q_m - \mu_q)^2}{2\sigma_q^2} \right] + \sum_{m=1}^{M} \sum_{n=1}^{N} \left[-\frac{1}{2} ln\sigma_w^2 - \frac{(w_{mn} - \mu_w)^2}{2\sigma_w^2} \right]$$

Experimental Data

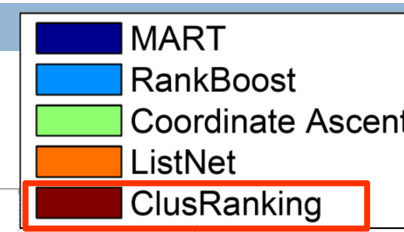
Beijing real-world Data

- Beijing real estate data
  - 2851 estates with transaction records from 04/2011 to 09/2012
  - Falling market(04/2011 to 02/2012) and Rising market (02/2012 to 09/2012)
- Beijing transportation facility data including bus stop, subway, road networks
- Beijing POI data
- Beijing taxi GPS traces

Data Sources	Properties	Statistics
Real estates	Number of real estates	2,851
	Size of bounding box (km)	40*40
	Time period of transactions	04/2011 - 09/2012
Bus stop(2011)	Number of bus stop	9,810
Subway(2011)	Number of subway station	215
Road networks (2011)	Number of road segments	162,246
	Total length(km)	20,022
	Percentage of major roads	7.5%
POIs	Number 0f POIs	300,811
	Number of categories	13
Taxi Trajectories	Number of taxis	13,597
	Effective days	92
	Time period	Apr. - Aug. 2012
	Number of trips	8,202,012
	Number of GPS points	111,602
	Total distance(km)	61,269,029

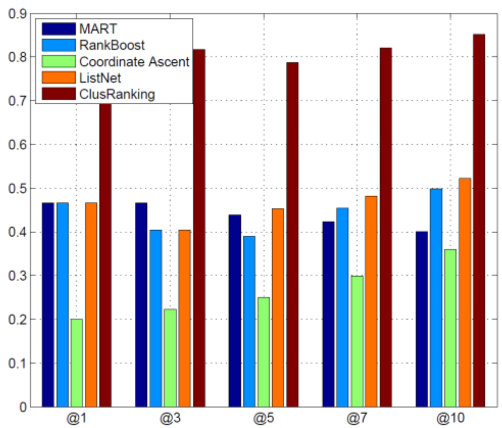
Table 4: Statistics of the experimental data.

Top-K Recommendation

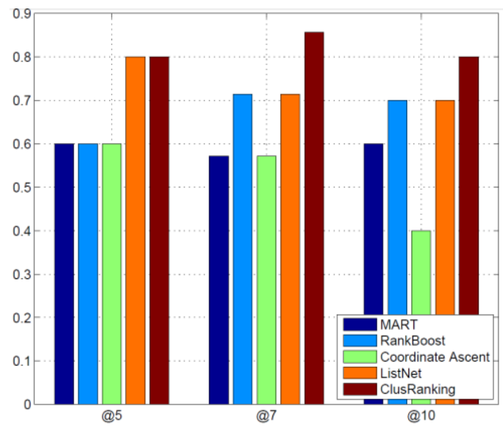


Peer and zone dependences can boost top-k recommendation

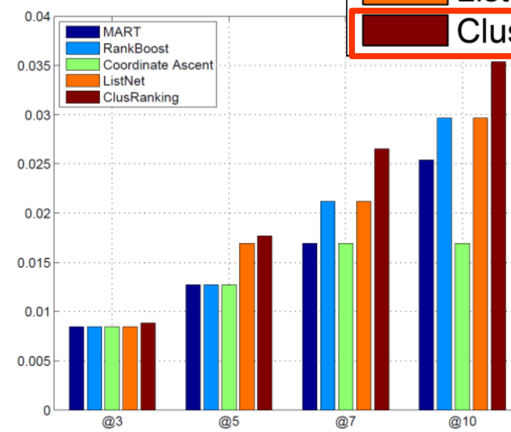
Comparison in rising markets



NDCG@K

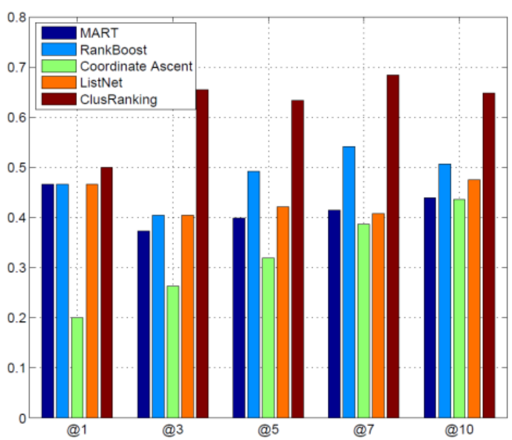


Precision@K

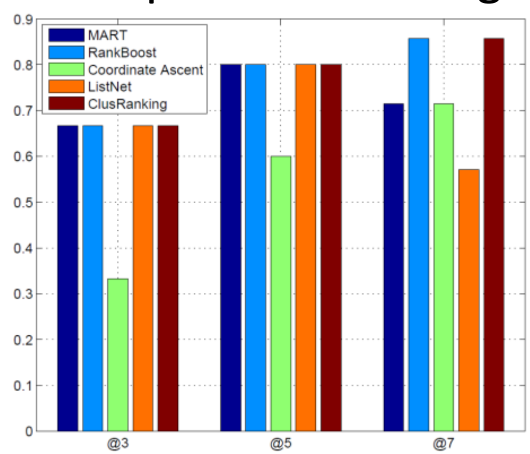


Recall@K

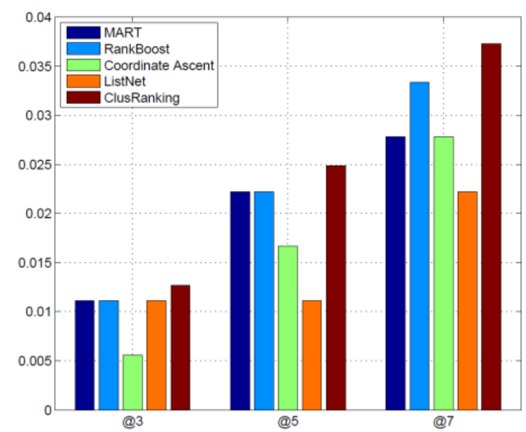
Comparison in falling markets



NDCG@K



Precision@K



Recall@K

Importance of Three Predictors

Step1: extract the values of the three predictors from the learned model
Step2: feed the three predictors along with the benchmark location ratings into a random forest model
Step3: extract the Gini importance of the three predictors

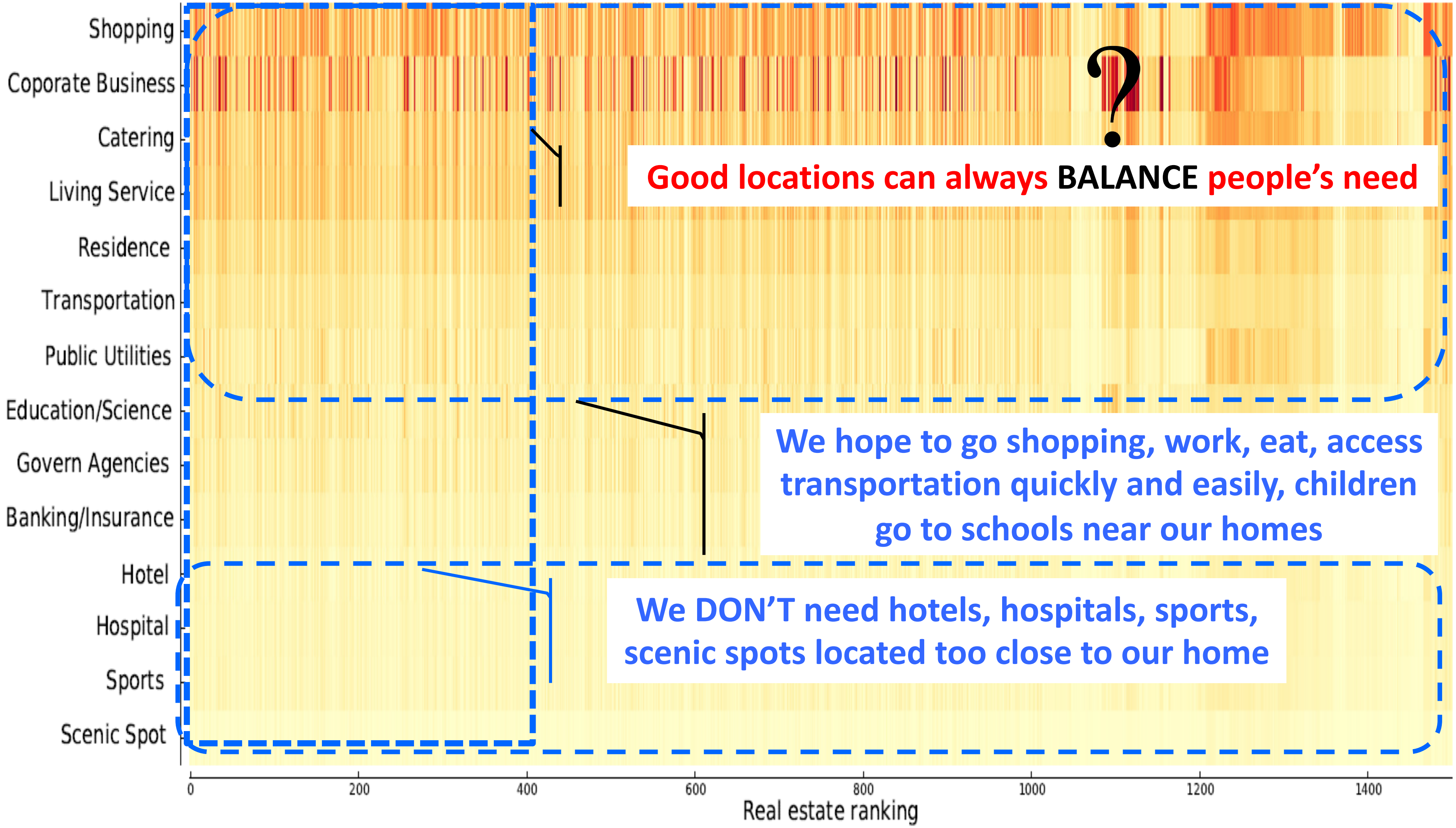
The Gini importance of the three factors.

Market	Geo-Utility	Business Areas Influence	Popularity
Rising	40.92804	40.37436	31.07325
Falling	34.79067	34.03652	28.14835

1, **Geographic Utility** (land uses) **>= Influence of Business Areas** (business prosperity) **>> Neighborhood Popularity** (human mobility)
2, Influence of business areas is implicit, latent, but significant

Understanding Human Needs

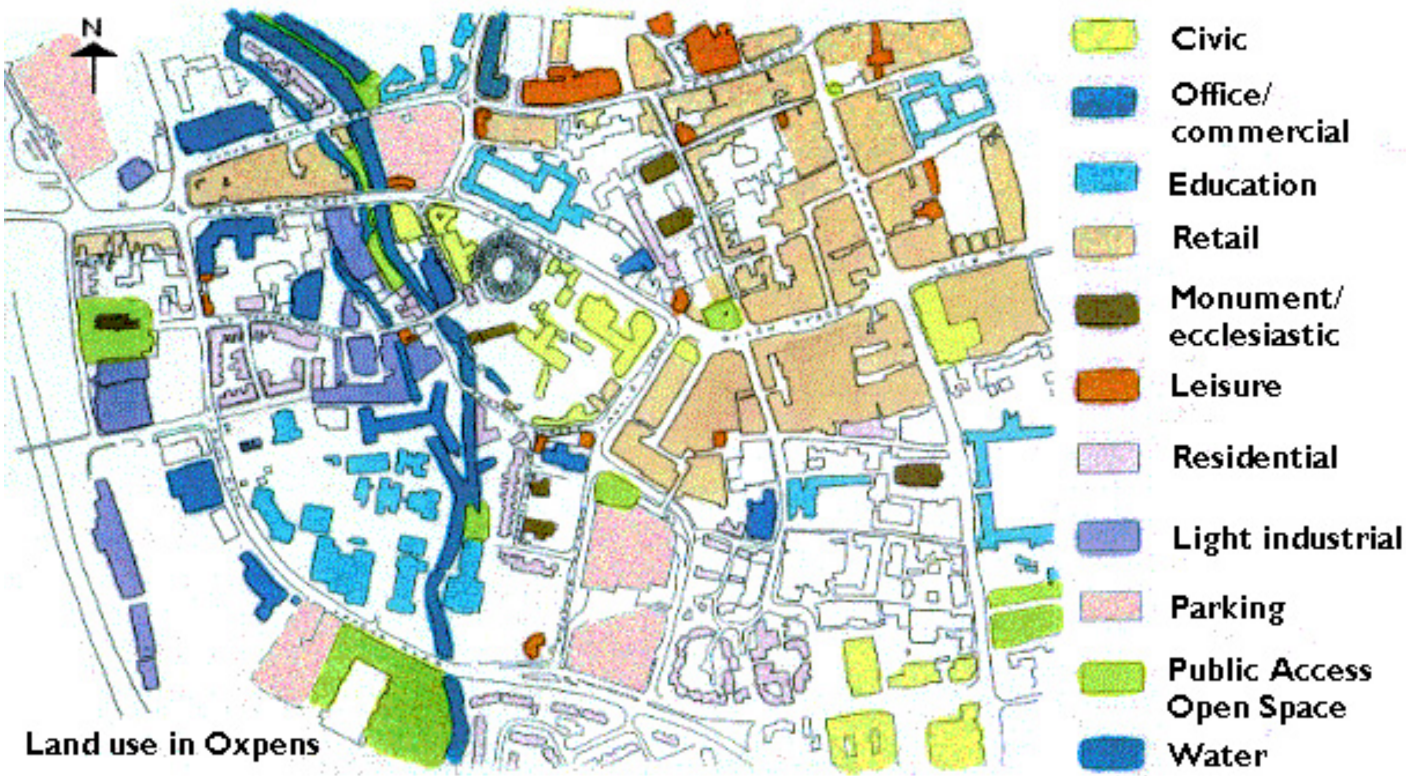
POI density spectrum of different categories over house rankings



Outline

- Background and Motivation
- Preliminary Analysis
- Modeling Geographic Dependencies
- **Exploring Mixed Land Use**
- Conclusion and Future Work

Definition of Mixed Land Use

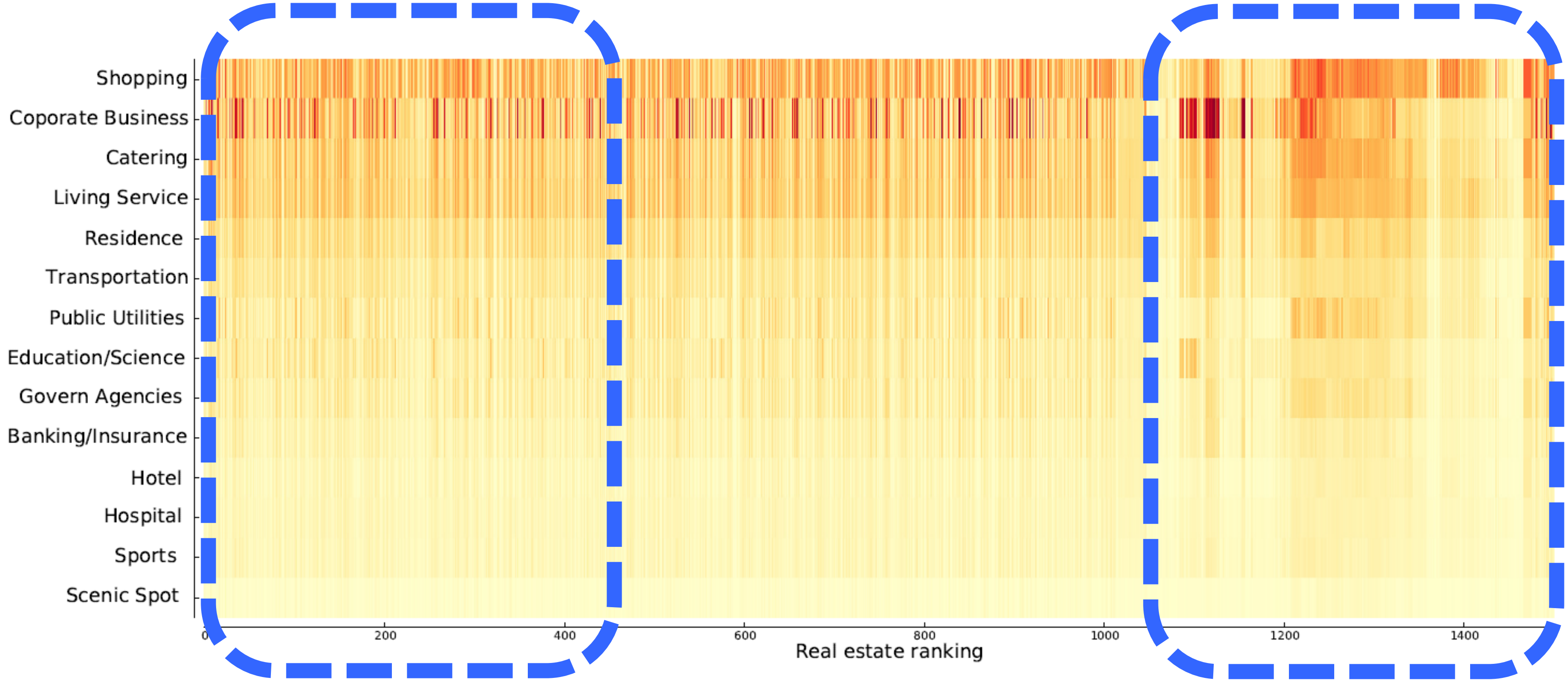


- Defined as a mixture of residential uses and compatible non-residential uses (*e.g., commercial, education, and office uses*) within a certain area
- Implying proximity of households to each other, but also to different types of community functions

Importance of Mixed Land Use

- Contribute economic benefits
  - Commercial areas in close proximity to residential areas can increase property values
- Support viable public transit
- Enhance the perceived security
  - By helping increase activity and hence the presence of people on the street
- Lead to co-location of socio-economic functions
- Yield livable, sustainable, and viable neighborhoods

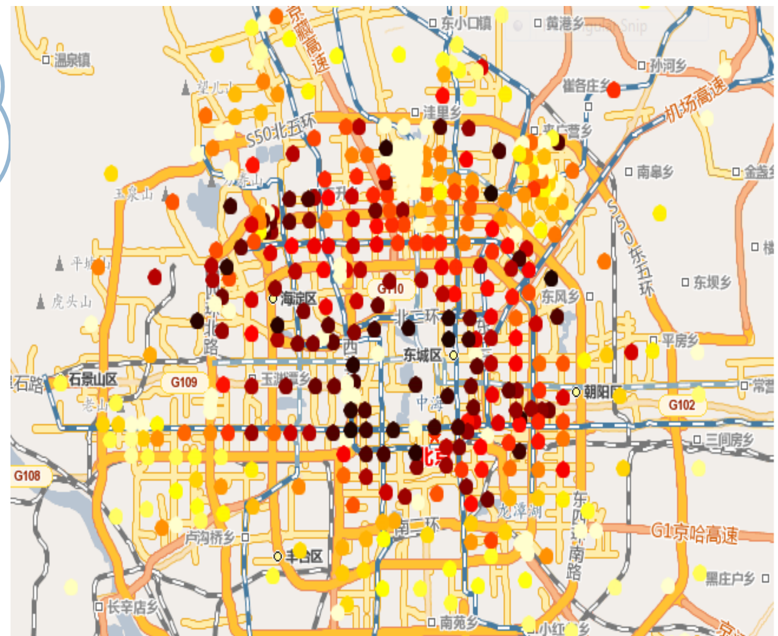
Mixed Land Use Increases Value



- In big cities, people value a balanced mix of land uses more than other key indicators of real estate value.
- People are willing to pay almost 25% more for a residential complex in an area with appropriate mixed land use.

Ranking via Mixed Land Use

Why not explore the impact of mixed land using to rank real estate?



Real Estate Investment Rating ⟷ **Location** ⟷ **Mixed Land Use**

What and How to Mix

The composition of community functions

Office

Retail

Education

Transportation

Light industry

Government agencies

Residential

Entertainment

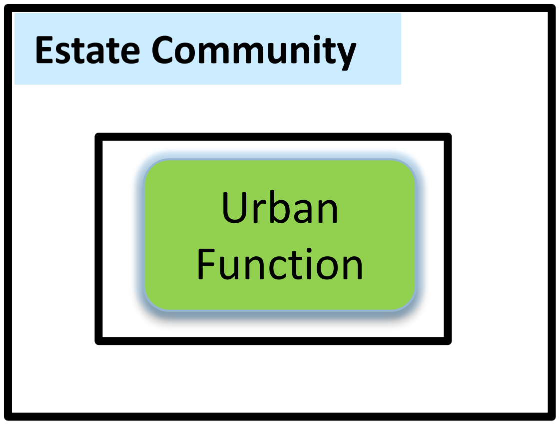
Finance and banking

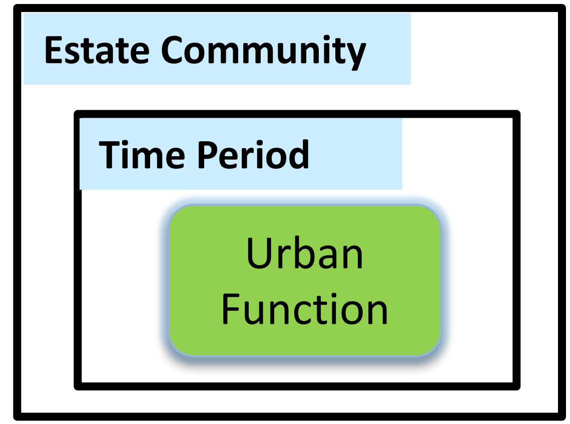
Heavy industry

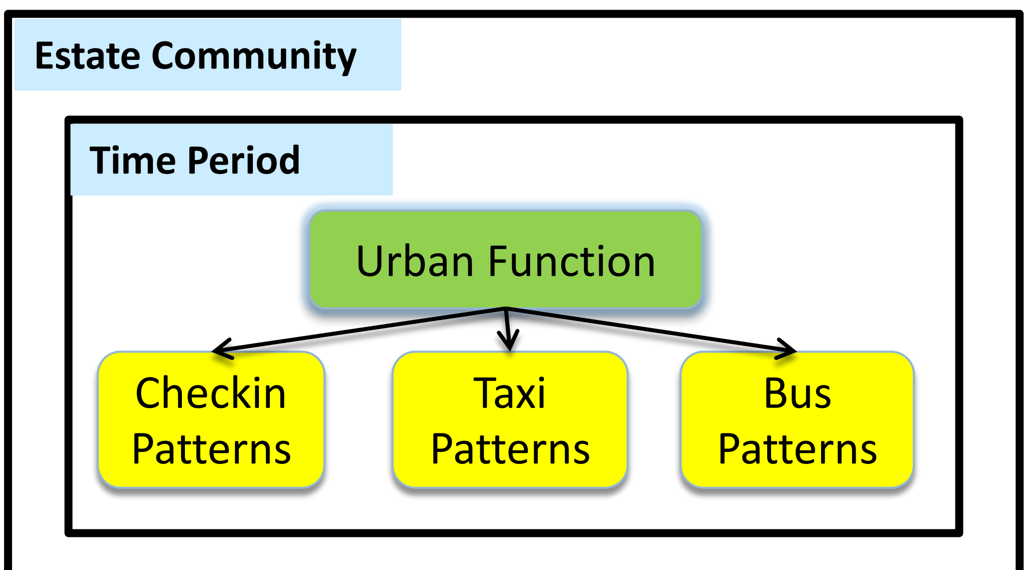
- **What to be mixed?**
  - Identify compatible urban functions that help increase real estate value
- **How to mix?**
  - Learn the optimal portfolio of these compatible functions in a community

Research Insight (1)

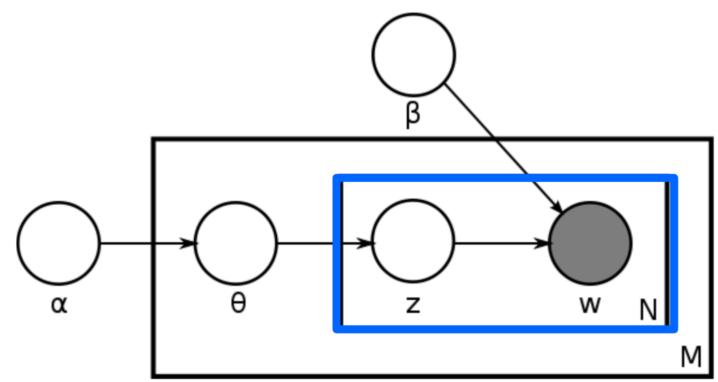
The correlations among estate communities, urban functions, temporal effect, and mobility patterns



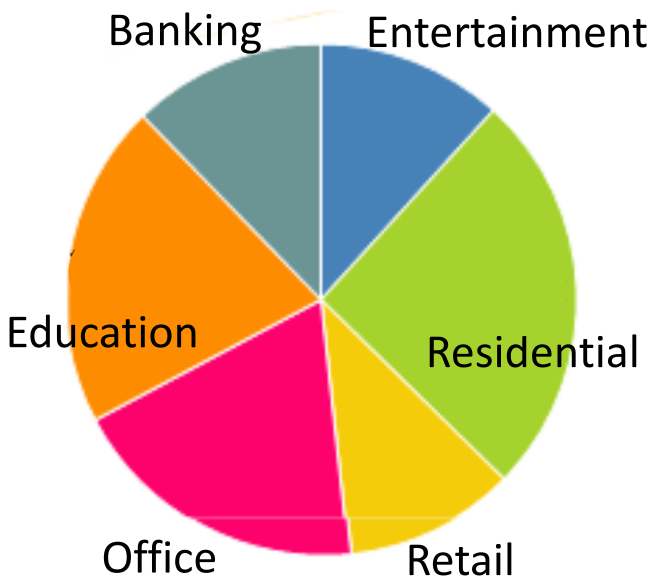




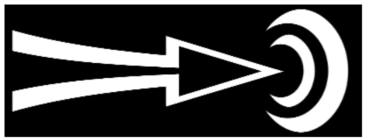
Community-Mobility Patterns	Doc-Words



How Diversity Impacts Value



Affect



Return of Investment

- Question: What is the impact of the portfolio of community functions on real estate values?
- Idea: Model the correlation between functional portfolios and real estate rankings via functional diversity

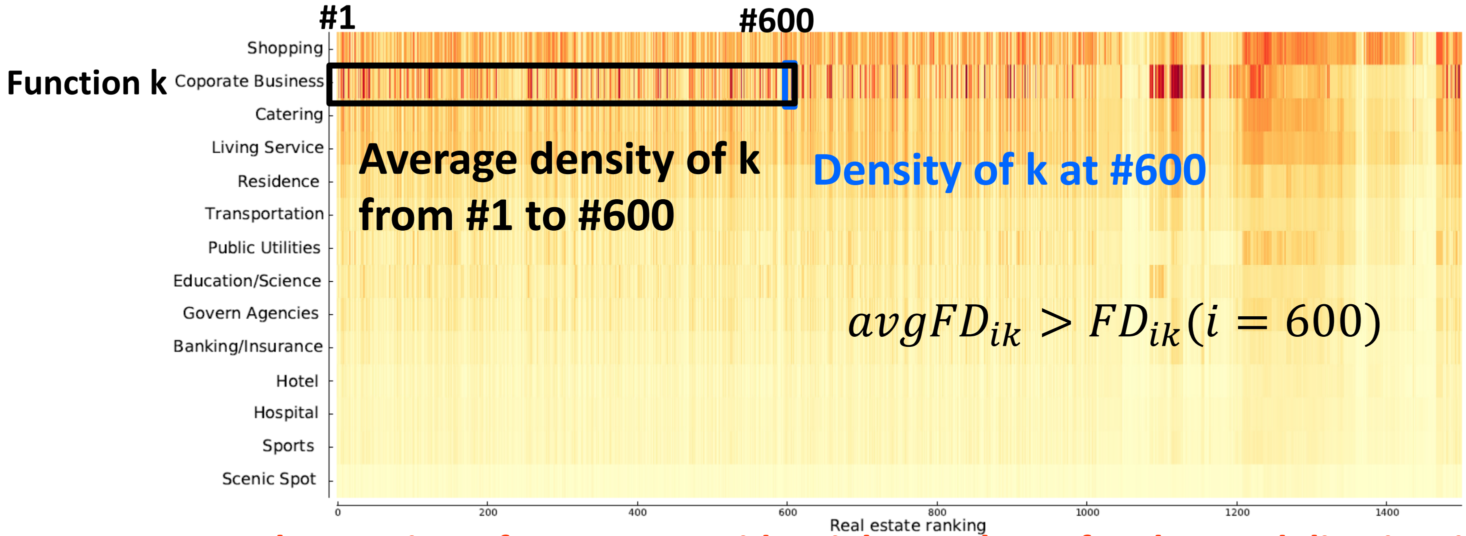
Research Insight (2)

- **Functional diversity: a generalized weighted sum function (two-step method)**

$$\sum_{k=1}^{K} P(k) f(k|\Xi, \Phi, \Lambda)$$

where $P(k)$ is the weight of the k-th urban function, $f(k|\Xi, \Phi, \Lambda)$ is the relevance score (information gain) of the entire complex ranked list given the function k

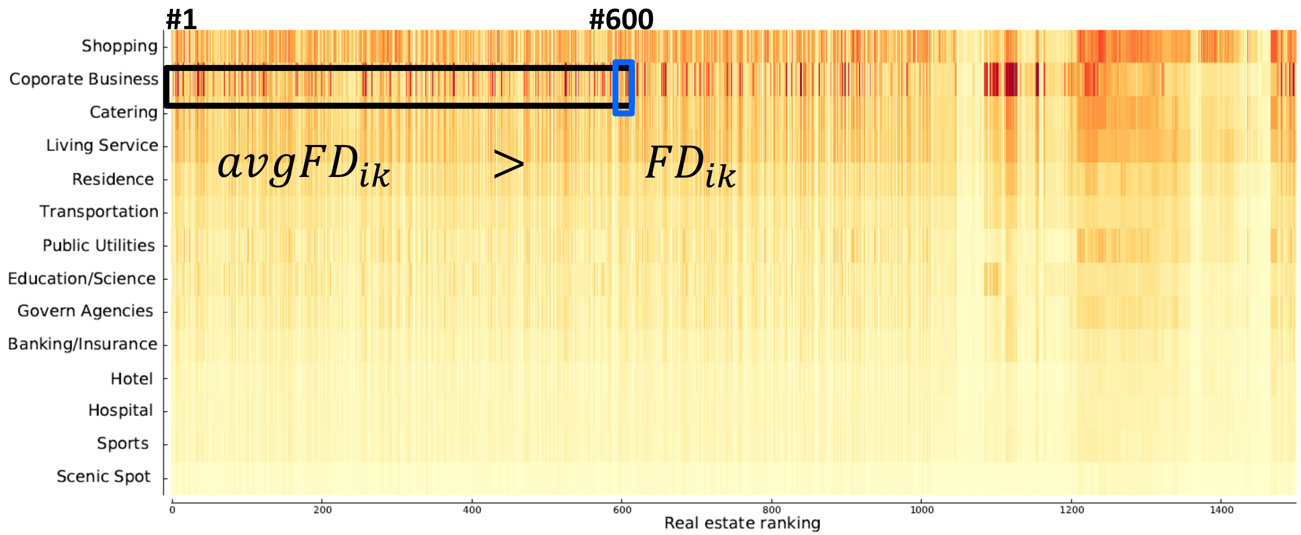
- **What is relevance?**
  - Modeling intuition: *if a urban function k can significantly increase value, a high-rated residential complex is likely to contain more urban function k*



An observation of top #600 residential complexes for the modeling intuition

Research Insight (2) Cont.

Functional portfolio as the listwise density distribution of K functions along the ranked list



- **How to quantify the relevance score given a function k?**
  - Borrow the idea of normalized Discounted Accumulated Gain

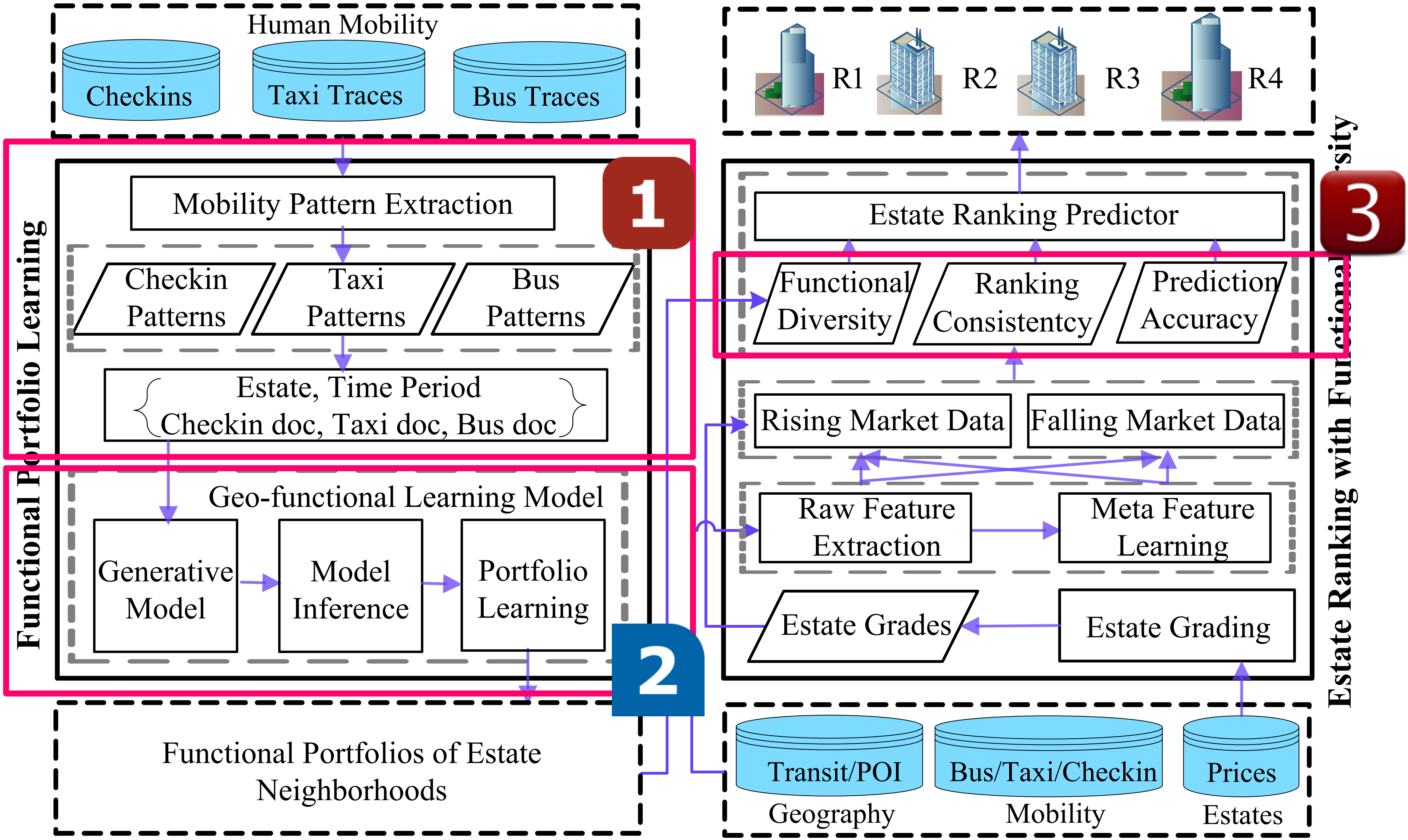
$$f(k|\Xi,\Phi,\Lambda) = sigmoid(\sum_{i=1}^{I} rating_i \times (avgFD_{ik} - FD_{ik}))$$

- Aggregate the weighted sum of K relevance scores to incorporate diversity
- Joint model of functional diversity and ranking consistency

Problem Definition

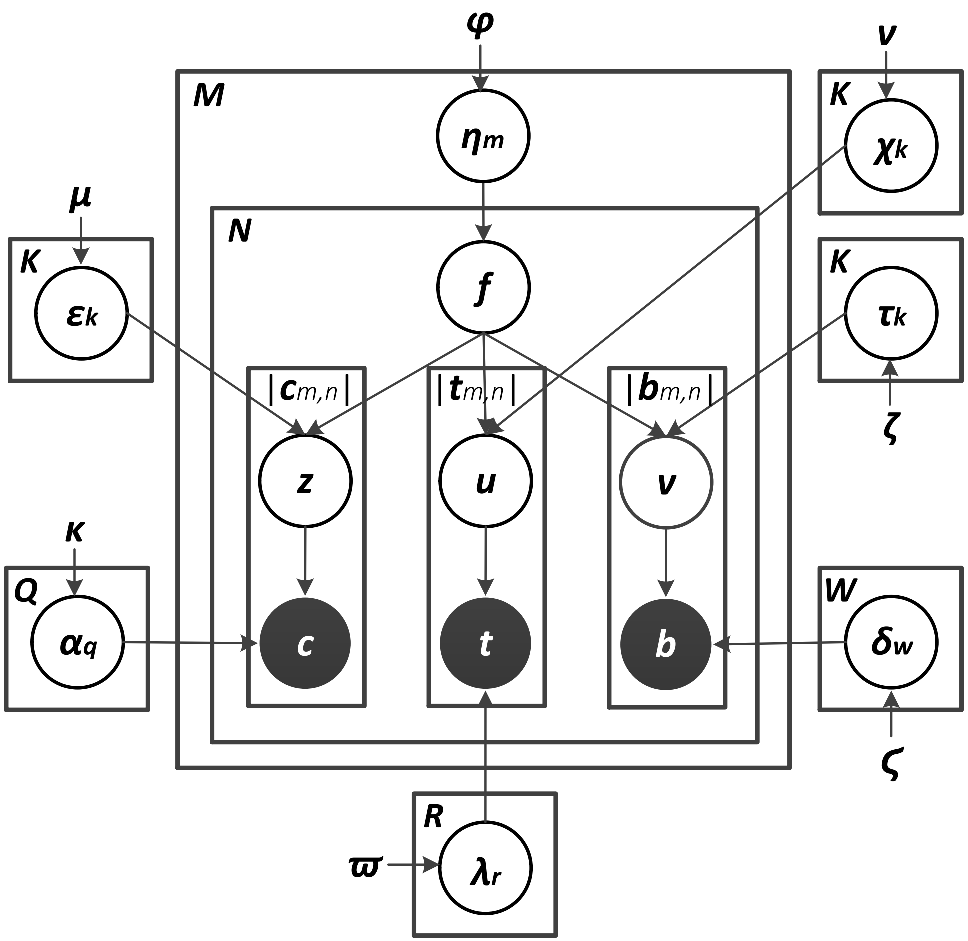
- **Given**
  - Estates with locations and historical prices
  - Urban geography (e.g., POIs, road networks)
  - Human mobility (e.g., taxi, bus, checkin)
  - Customer reviews of business venues
- **Objective**
  - Rank and classify residential complexes based on their ratings
- **Core tasks**
  - Identify **compatible community functions** and **their corresponding portfolios** for each residential complex
  - Incorporate **functional diversity** into **objective function** to enhance real estate ranking

Overview of FuncDivRank



Geographic Function Learning

Learning the portfolio of community functionalities (M estates for N time periods on K urban functions with C/T/B mobility)



- An estate community **m** is a mixture of urban functions (**η**)
- The urban function **f** of a community changes over time period **n**
- In a period, a community shows checkin (**C**), taxi (**T**), and bus (**B**) clusters of mobility patterns reflecting an urban function **f**
- A cluster of mobility pattern = a document
- A mobility pattern = a word
- Model doc-word with topic modeling

Solving GeoFuncLearning

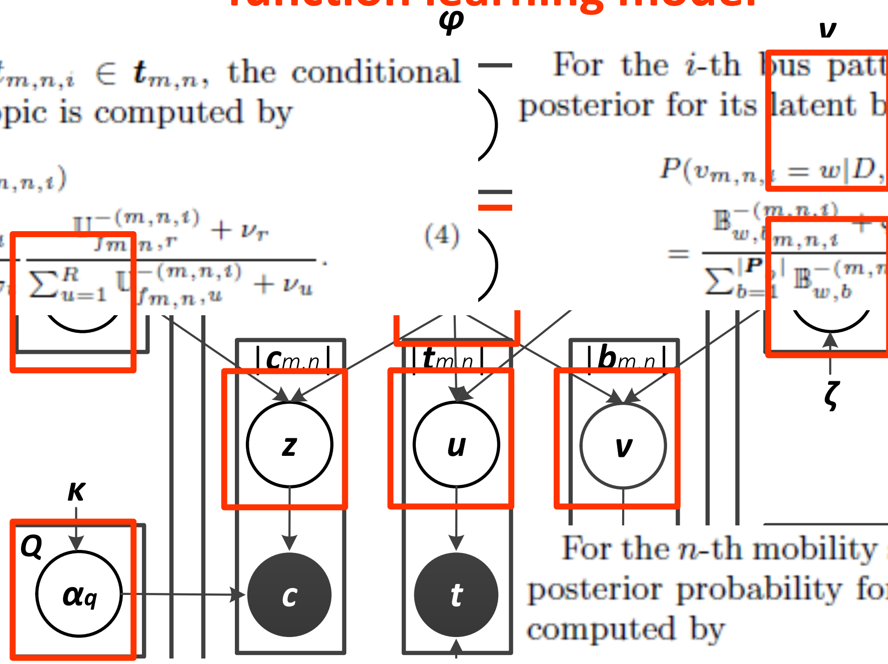
Collapsed Gibbs Sampling to learn the generative process of geographic function learning model

For the i-th taxi pattern $t_{m,n,i} \in \boldsymbol{t}_{m,n}$, the conditional posterior for its latent taxi topic is computed by

$$P(u_{m,n,i} = r | D, \Upsilon - u_{m,n,i}) = \frac{\mathbb{T}^{-(m,n,i)}_{r,t_{m,n,i}} + \varpi_{t_{m,n,i}}}{\sum_{t=1}^{|\boldsymbol{P}_t|} \mathbb{T}^{-(m,n,i)}_{r,t} + \varpi_t} \cdot \frac{\mathbb{U}^{-(m,n,i)}_{f_{m,n},r} + \nu_r}{\sum_{u=1}^{R} \mathbb{U}^{-(m,n,i)}_{f_{m,n},u} + \nu_u}. \quad (4)$$

For the i-th bus pattern $b_{m,n,i} \in \boldsymbol{b}_{m,n}$, the conditional posterior for its latent bus topic is computed by

$$P(v_{m,n,i} = w | D, \Upsilon - v_{m,n,i}) = \frac{\mathbb{B}^{-(m,n,i)}_{w,b_{m,n,i}} + \varsigma_{b_{m,n,i}}}{\sum_{b=1}^{|\boldsymbol{P}_b|} \mathbb{B}^{-(m,n,i)}_{w,b} + \varsigma_b} \cdot \frac{\mathbb{V}^{-(m,n,i)}_{f_{m,n},w} + \zeta_w}{\sum_{v=1}^{W} \mathbb{V}^{-(m,n,i)}_{f_{m,n},v} + \zeta_v}. \quad (5)$$



For the n-th mobility segment in estate m, the conditional posterior probability for its latent function assignment f is computed by

$$P(f_{m,n} = k | D, \Upsilon - f_{m,n}) = \frac{\mathbb{F}^{-(m,n)}_{m,k} + \rho_k}{\sum_{f=1}^{K} \mathbb{F}^{-(m,n)}_{m,f} + \rho_f} \times \frac{\prod_{z=1}^{Q} \Gamma(\mathbb{Z}_{k,z} + \mu_z)\Gamma(\sum_{z=1}^{Q} \mathbb{Z}^{-(m,n)}_{k,z} + \mu_z)}{\prod_{z=1}^{Q} \Gamma(\mathbb{Z}^{-(m,n)}_{k,z} + \mu_z)\Gamma(\sum_{z=1}^{Q} \mathbb{Z}_{k,z} + \mu_z)} \times \frac{\prod_{u=1}^{R} \Gamma(\mathbb{U}_{k,u} + \nu_u)\Gamma(\sum_{u=1}^{R} \mathbb{U}^{-(m,n)}_{k,u} + \nu_u)}{\prod_{u=1}^{R} \Gamma(\mathbb{U}^{-(m,n)}_{k,u} + \nu_u)\Gamma(\sum_{u=1}^{R} \mathbb{U}_{k,u} + \nu_u)} \times \frac{\prod_{v=1}^{W} \Gamma(\mathbb{V}_{k,v} + \zeta_v)\Gamma(\sum_{v=1}^{W} \mathbb{V}^{-(m,n)}_{k,v} + \zeta_v)}{\prod_{v=1}^{W} \Gamma(\mathbb{V}^{-(m,n)}_{k,v} + \zeta_v)\Gamma(\sum_{v=1}^{W} \mathbb{V}_{k,v} + \zeta_v)}. \quad (2)$$

For the i-th checkin pattern $c_{m,n,i} \in \boldsymbol{c}_{m,n}$, the conditional posterior for its latent checkin topic is computed by

$$P(z_{m,n,i} = q | D, \Upsilon - z_{m,n,i}) = \frac{\mathbb{C}^{-(m,n,i)}_{q,c_{m,n,i}} + \kappa_{c_m}}{\sum_{c=1}^{|\boldsymbol{P}_c|} \mathbb{C}^{-(m,n,i)}_{q,c} [illegible]}$$

After all the latent assignments a[illegible] update rules of the model parameter[illegible]

$$\epsilon_{f,z} = \frac{\mathbb{Z}_{f,z} + \mu_z}{\sum_{q=1}^{Q} \mathbb{Z}_{f,q} + \mu_q}, \chi_{f,u} = \frac{\mathbb{U}_{f,u} + [illegible]}{\sum_{r=1}^{R} \mathbb{U}_{f,} [illegible]}$$

$$\alpha_{z,c} = \frac{\mathbb{C}_{z,c} + \kappa_c}{\sum_{p=1}^{|\boldsymbol{P}_c|} \mathbb{C}_{z,p} + \kappa_p}, \lambda_{u,t} = \frac{\mathbb{T}_{u,t} [illegible]}{\sum_{p=1}^{|\boldsymbol{P}_t|} \mathbb{T} [illegible]}$$

update the latent a[illegible] functions, checkin/t[illegible] each mobility patter[illegible]

Predicting Investment Rating

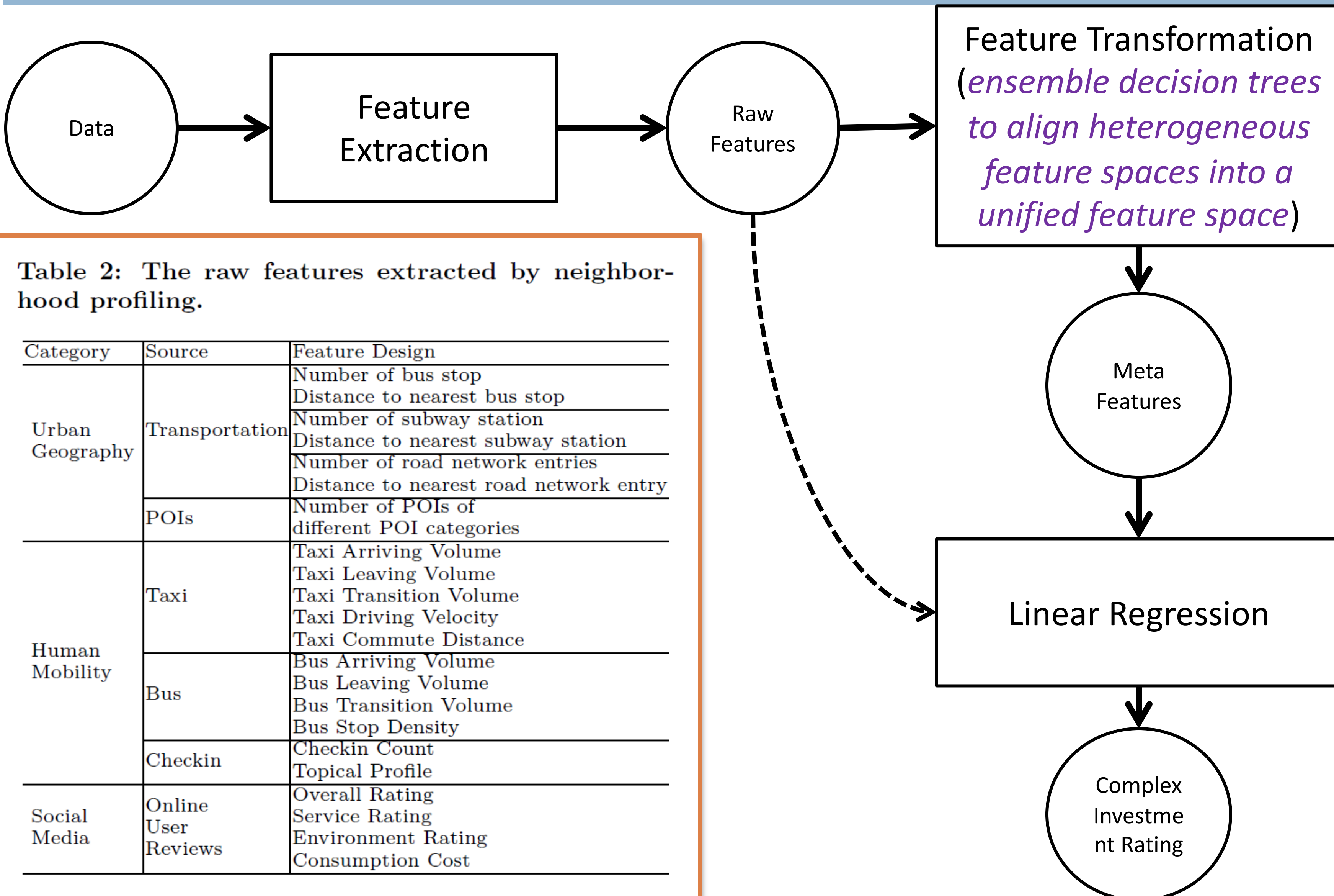


Table 2: The raw features extracted by neighborhood profiling.

Category	Source	Feature Design
Urban Geography	Transportation	Number of bus stop Distance to nearest bus stop
		Number of subway station Distance to nearest subway station
		Number of road network entries Distance to nearest road network entry
	POIs	Number of POIs of different POI categories
Human Mobility	Taxi	Taxi Arriving Volume Taxi Leaving Volume Taxi Transition Volume Taxi Driving Velocity Taxi Commute Distance
	Bus	Bus Arriving Volume Bus Leaving Volume Bus Transition Volume Bus Stop Density
	Checkin	Checkin Count Topical Profile
Social Media	Online User Reviews	Overall Rating Service Rating Environment Rating Consumption Cost

Ranking with Function Diversity

- **Prediction Accuracy (pointwise: investment ratings are categorical with distinctness)**
  - Describe prediction accuracy of real estate investment values

$$P\left(Y|\Phi,\Lambda\right)=\prod_{m=1}^{M}\mathcal{N}(y_m|g_m,\sigma)=\prod_{m=1}^{M}\frac{1}{\sigma}\exp\left(-\frac{(y_m-g_m)^2}{2\sigma^2}\right)$$

- **Ranking Consistency (pairwise: investment ratings are ordinal with order)**
  - Describe pairwise accuracy of real estate rankings

$$P\left(\Pi|\Phi,\Lambda\right)=\prod_{m=1}^{M-1}\prod_{h=m+1}^{M}P(m\rightarrow h|\Phi,\Lambda)$$

- **Functional Diversity (listwise: high-rated locations maximally cover K functions)**
  - Describe functional coverage of real estate rankings

$$P\left(\Xi|\Phi,\Lambda\right)=\sum_{f=1}^{K}P(f)P(\Xi|f,\Phi,\Lambda)=\sum_{f=1}^{K}\frac{\theta_f}{1+exp(-(\sum_{m=1}^{M}g_m\frac{\sum_{h=1}^{m}\eta_{h,f}}{m}-\sum_{m=1}^{M}g_m\eta_{m,f}))}$$

- **By Bayesian inference, the posterior probability is**

$$P\left(\Delta|\Phi,\Lambda\right)=P\left(\{Y,\Pi,\Xi\}|\Phi,\Lambda\right)$$
$$=\underbrace{P\left(Y|\Phi,\Lambda\right)}_{\text{Prediction Accuracy}}\times\underbrace{P\left(\Pi|\Phi,\Lambda\right)}_{\text{Ranking Consistency}}\times\underbrace{P\left(\Xi|\Phi,\Lambda\right)}_{\text{Functional Diversity}}$$

Experimental Data

- Beijing real-world Data
  - Beijing estate data
    - 2851 estates with transaction records from 04/2011 to 09/2012
    - Falling market(04/2011 to 02/2012) and Rising market (02/2012 to 09/2012)
  - Beijing road networks
  - Beijing bus and subway systems
  - Beijing taxi GPS traces
  - Beijing bus GPS traces
  - Beijing check-ins
  - Beijing business review

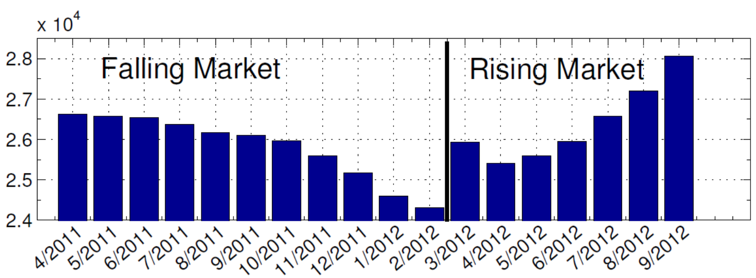
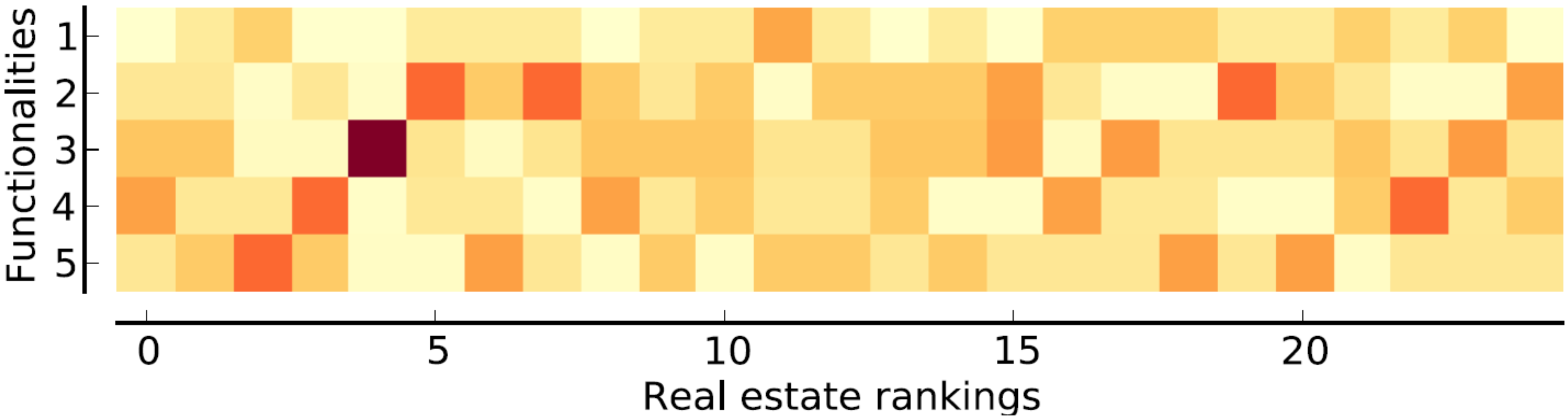


Table 3: Statistics of the experiment data.

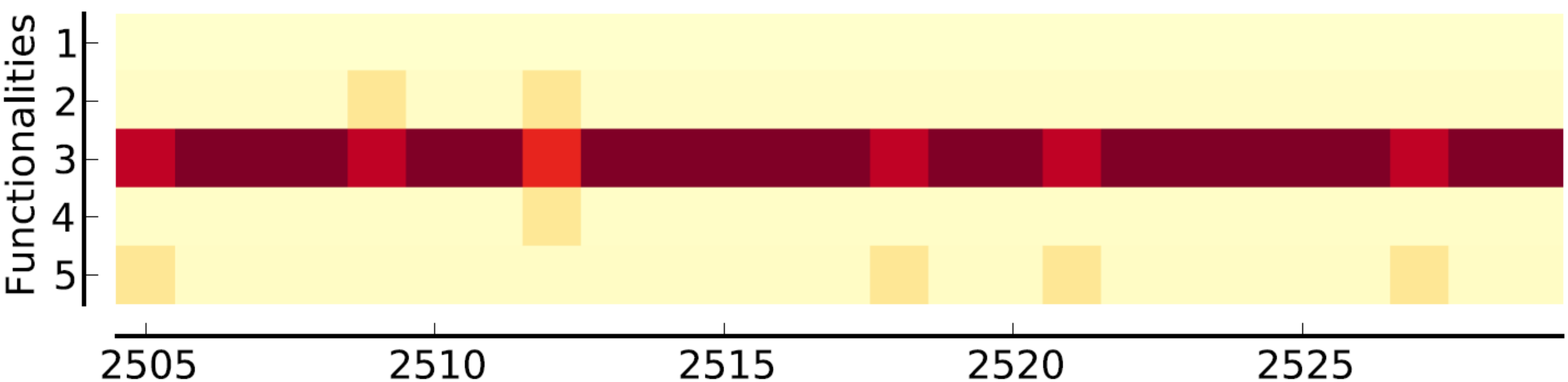
Data Sources	Properties	Statistics
Bus stop(2011)	Number of bus stop	9,810
Subway(2011)	Number of subway station	215
Road networks (2011)	Number of road segments Total length(km) Percentage of major roads	162,246 20,022 7.5%
POIs	Number 0f POIs Number of categories	300,811 13
Taxi Traces	Number of taxis Effective days Time period Number of trips Number of GPS points Total distance(km)	13,597 92 Apr. - Aug. 2012 8,202,012 111,602 61,269,029
Smart Card Transactions	Number of bus stops Time Period Number of car holders Number of trips	9,810 Aug 2012 to May 2013. 300,250 1,730,000
Check-Ins	Number of check-in POIs Number of check-in events Number of POI categories Time Period	5,874 2,762,128 9 01/2012-12/2012
Business Review	Number of reviews Number of users	470846 159820
Real Estates	Number of estates Size of bounding box (km) Time period of transactions	2,851 40*40 04/2011 - 09/2012

Study of GeoFuncLearning

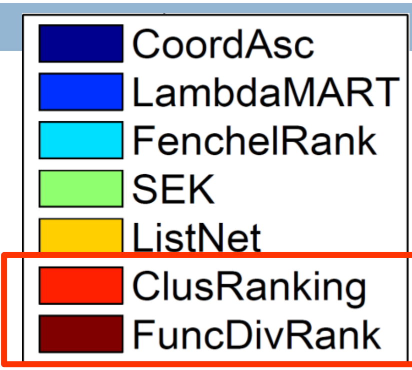
The portfolios of urban functions for **high-rated** complexes



The portfolios of urban functions for **low-rated** complexes

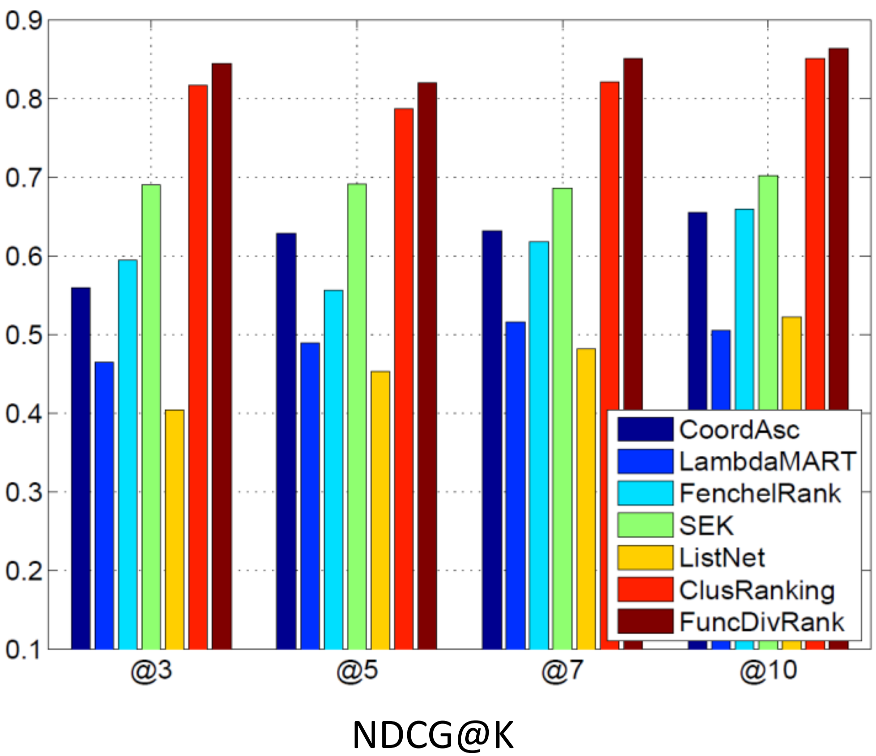


Top-K Recommendation

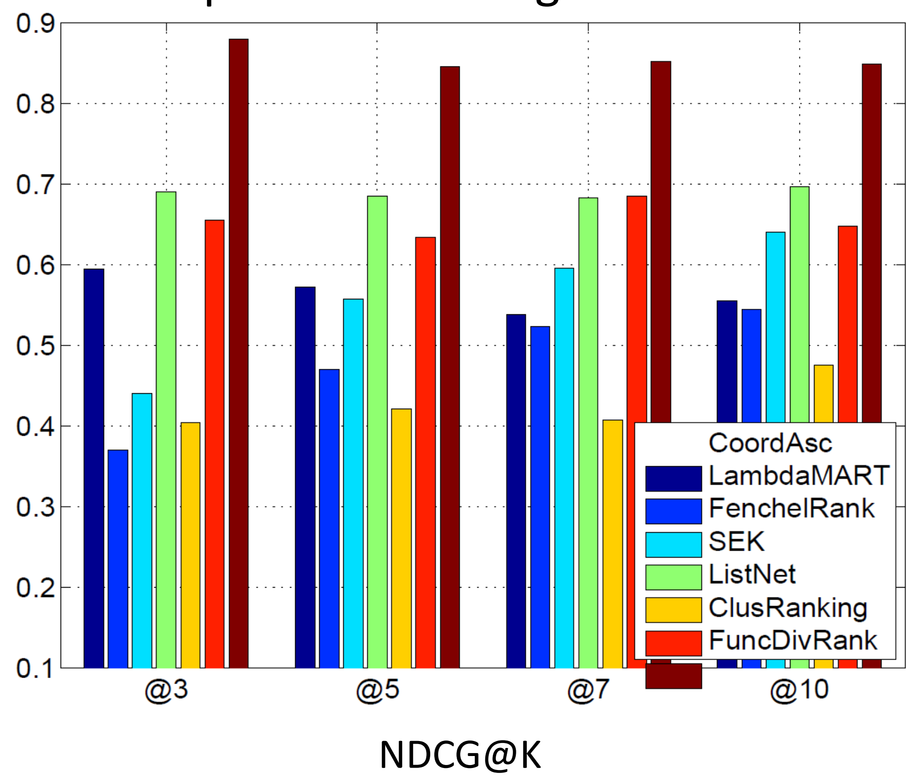


Capturing functional diversity can help spot low-rated residential locations and continue to enhance ranking performances

Comparison in rising markets



Comparison in falling markets



Outline

- Background and Motivation
- Preliminary Analysis
- Modeling Geographic Dependencies
- Exploring Mixed Land Use
- **Conclusion and Future Work**

Conclusion (1)

- **Real Estate Ranking**
  - To *rank and classify* high-rated residential complexes with locational interpretations and explanations
  - The *first* to bring in fine-grained urban geography and dynamic human mobility
- **Multi-view learning-to-rank (insights)**
  - Ranking with geographic dependencies
    - Model geographic individual, peer, and zone dependencies
  - Ranking with mobility patterns
    - Explore the impact of mixed land use via the diversity of community functions with heterogeneous human mobility
- **Effective methods to turn big data into decision making support (performances)**

Conclusion (2)

- **Generalization Potential and Benefits (capabilities)**
  - Geographic dependencies can be generalized for
    - Market segmentation
    - Other geo-items (e.g., restaurants, retail stores, etc.)
    - Other cities of similar mixed use developments
    - Social network (individual and group)
  - Geo-function learning method can be generalized for
    - Modeling various mobility data
    - Profiling urban function portfolio
    - Business site selection
    - Discovering urban lifestyle
    - Toward personalized real estate recommendation by considering personalized preference on the portfolios of urban functions
  - Diversity modeling over a ranking list
    - Weighted sum-up function + normalized discounted accumulated gain

Future Work (1)

Urban and Mobile Analytics

(1) Urban Region Level

- ❖ Plan transportation, facilities, functions for sustainable communities

(2) Mobile User Level

- ❖ Profile mobile users for personalized customer targeting

(3) Network Systems and Device Level

- ❖ Enhance effectiveness and security for smart customer care

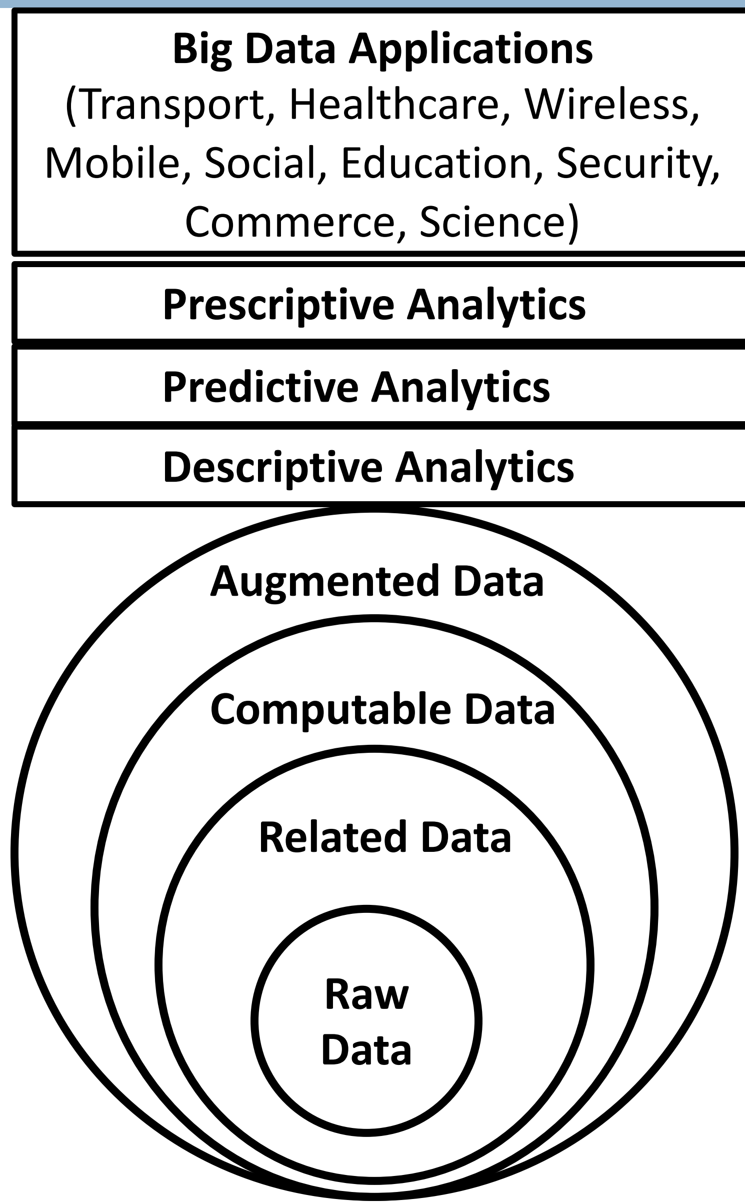
Learning with Cross-Domain Data

(1) How do analytical approaches alleviate information heterogeneity and asymmetry in data space?

(2) What role do modeling regulations play in exploring the correlations among heterogeneous information?

Future Work (2)

- Big picture



Acknowledgements

- Rutgers University
  - Advisor: Professor Hui Xiong
  - Dissertation Committee Member: Professor Jian Yang
  - Dissertation Committee Member: Professor Spiros Papadimitriou
  - All group members
- University of Minnesota Twin Cities
  - Dissertation Committee Member: Professor Rui Kuang
- IBM Thomas J. Watson Research Center
  - Dr. Charu Aggarwal, Dr. Deepak S Turaga, and all group members
- Microsoft Research Asia
  - Dr. Xing Xie, Dr. Yu Zheng, and all group members
- Nanjing University
  - Professor Zhi-Hua Zhou
- Tsinghua University
  - Professor Guoqing Chen
- Huawei 2012 Labs
  - Dr. Jin Yang, Dr. Nandu Gopalakrishnan, Dr. Xin Yan, and all group members

Questions?

WE ARE JUST ON THE WAY

THANK YOU.

Homepage: https://sites.google.com/site/yanjiefoo/