Lab for Media Search

National University of Singapore

Social Media Computing

Lecture 6: Case Study – Multi-Source Profile Learning

Lecturer: Aleksandr Farseev

E-mail: farseev@u.nus.edu

Slides: http://farseev.com/ainlfruct.html

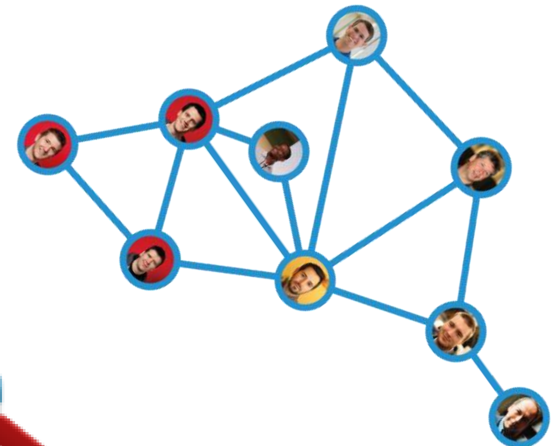
References

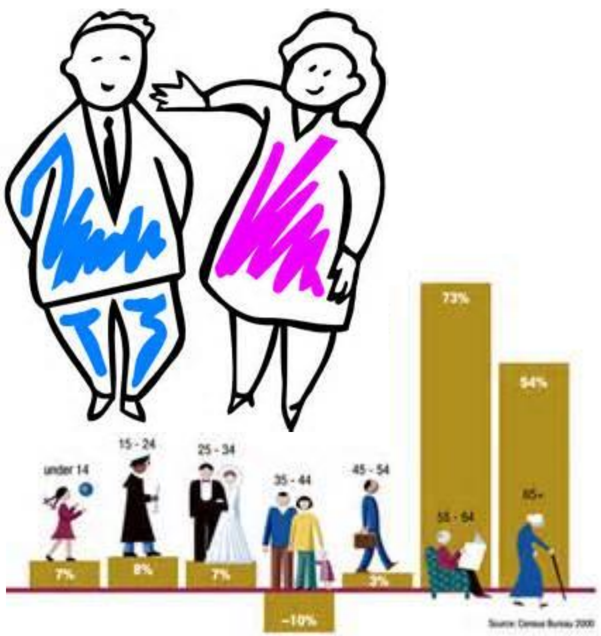
- A. Farseev, N. Liqiang, M. Akbari, and T.-S. Chua. **Harvesting multiple sources for user profile learning: a Big data study.** *ACM International Conference on Multimedia Retrieval (ICMR)*. China. June 23-26, 2015.

- A. Farseev, D. Kotkov, A. Semenov, J. Veijalainen, and T.-S. Chua. **Cross-Social Network Collaborative Recommendation.** *ACM International Conference on Web Science (WebSci)* 2015.

- А. Фарсеев, Н. Жуков, И. Государев, и Ю. Заричняк. **Разработка Кросплатформенной Рекомендательной Системы на Основе Извлечения Данных из Социальных Сетей** *Компьютерные Инструменты в Образовании*. June 2014.

What is user profile?





What is human mobility?

- Mobility - contemporary paradigm, which explores various types of people movement.

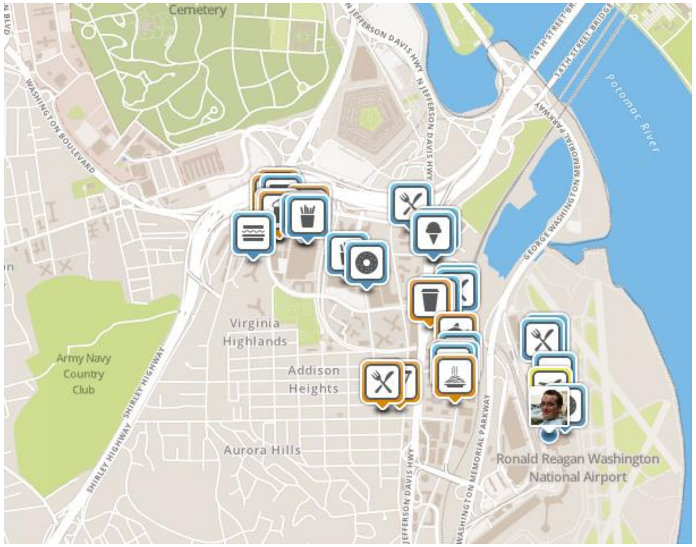
What is human mobility?

- Mobility - contemporary paradigm, which explores various types of people movement.

- The movement of people
- The quality or state of being mobile
- (Physiology) the ability to move physically
- (Sociology) movement within or between social classes and occupations
- (Chess) the ability of a chess piece to move around the board

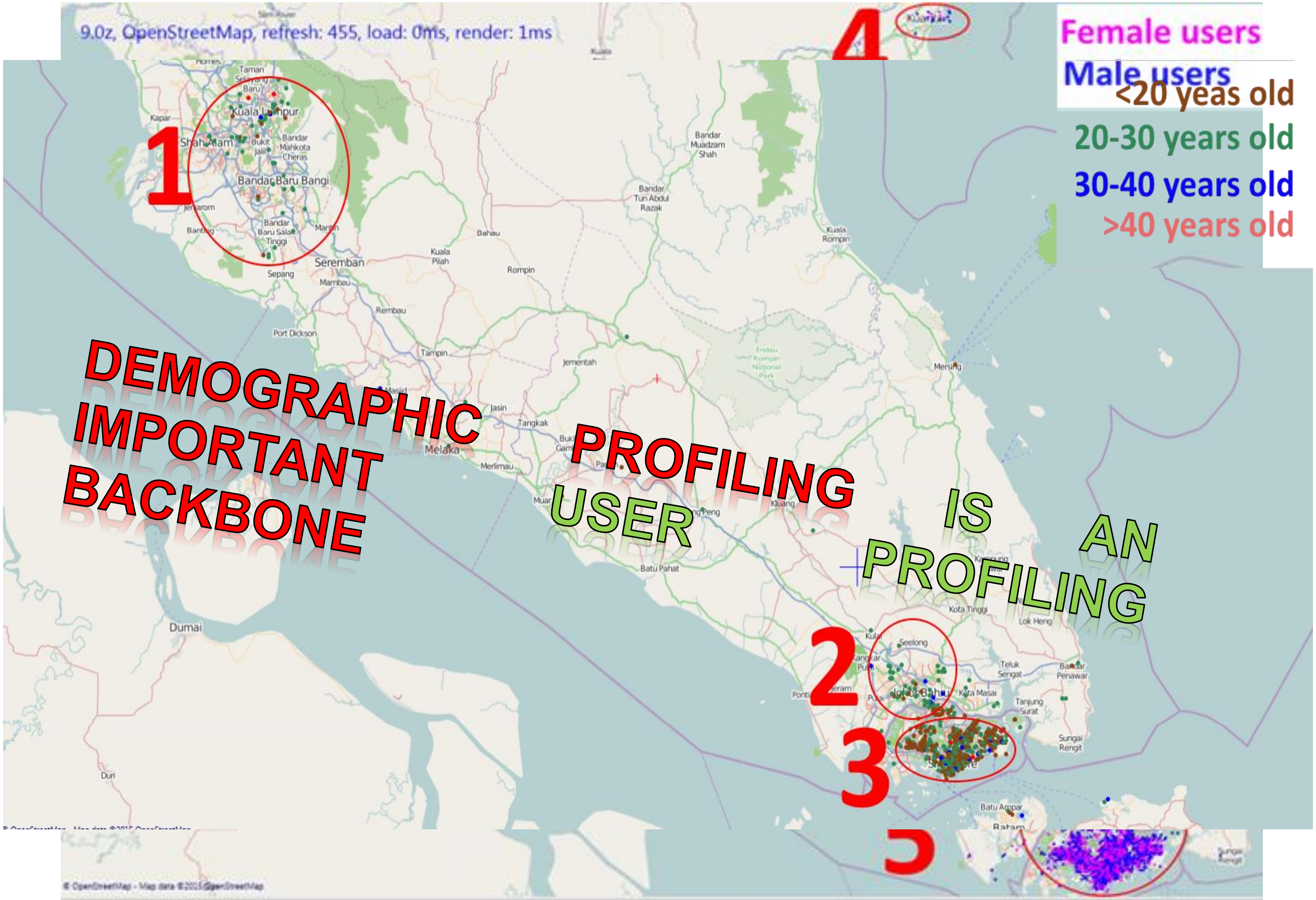
Why human mobility?

- **Urban planning: understand the city and optimize services**
- **Mobile applications and recommendations: study the user and offer services**



Mobility can describe people

How we want to know more?



AGE, GENDER

40, MALE

Marketing
Trade are analysis
Demography and interest - based marketing

Tent to stay at home, visit local pubs and shopping mall daily.

Wellness
Health group prediction
Lifestyle recommendation

Medium overweight, potential hypertonia and diabetes.

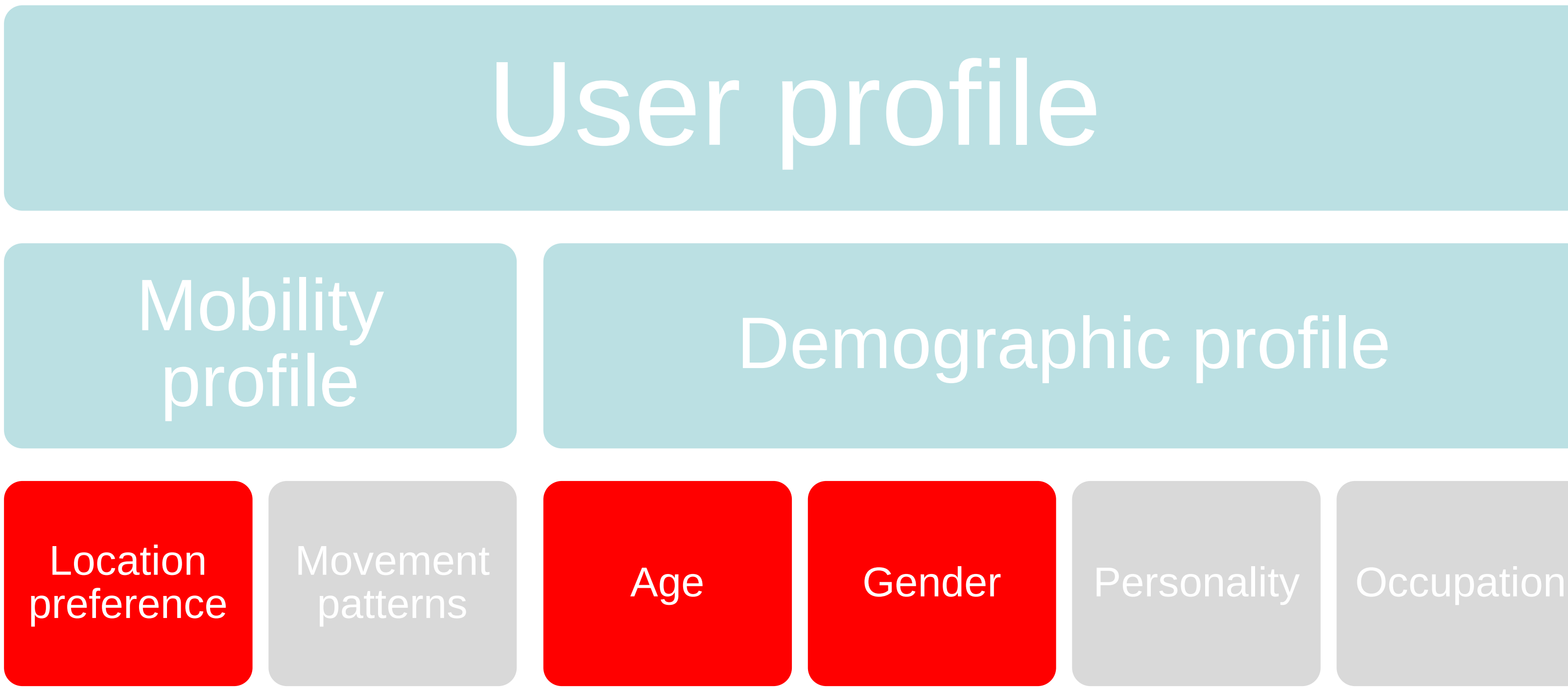
Advertisement
Demography and interest - based personalized advertisement

Advertise new Beer brand and new car models.

Assistance
Activity recommendation,
Venue recommendation,
Etc.

Morning excursive with medium intensity.

User profile: Mobility + Demography



Multiple sources describe user from multiple views

More than 50% of online-active adults use more than one social network in their daily life*

*According Paw Research Internet Project's Social Media Update 2013 (www.pewinternet.org/fact-sheets/social-networking-fact-sheet/)

Multiple sources describe user from multiple views

Research Problems

- Multi-source user profiling:
  - Geographical user mobility profiling
  - User demographic profiling
  - Data incompleteness
  - Multi–source multi–modal data integration

Multi-source dataset:
NUS-MSS*

*http://lms.comp.nus.edu.sg/research/NUS-MULTISOURCE.htm

NUS-MSS: Data sources

FOURSQUARE

BIGGEST LBSN

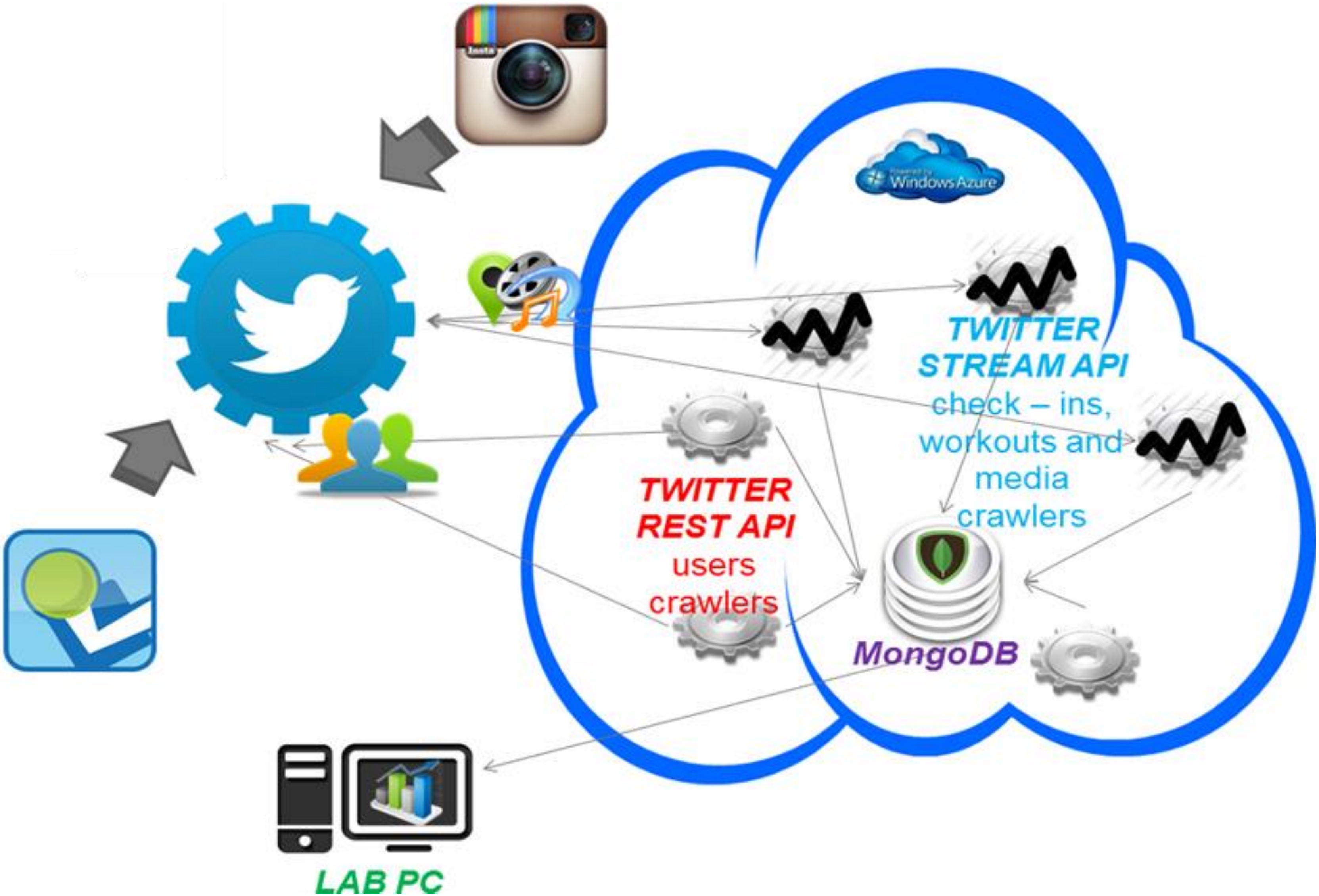
Instagram

BIGGEST PHOTO SHARING SERVICE

twitter

BIGGEST ENGLISH-SPEAKING MICROBLOG

NUS-MSS: Data collection



NUS-MSS: Dataset Description

Singapore

11,732,489 TWEETS

366,268 CHECK-INS

263,530 IMAGES

FROM 7,023 USERS

NUS-MSS: Dataset Description

London

2,973,162 TWEETS

127,276 CHECK-INS

65,088 IMAGES

FROM 5,503 USERS

NUS-MSS: Dataset Description

New York

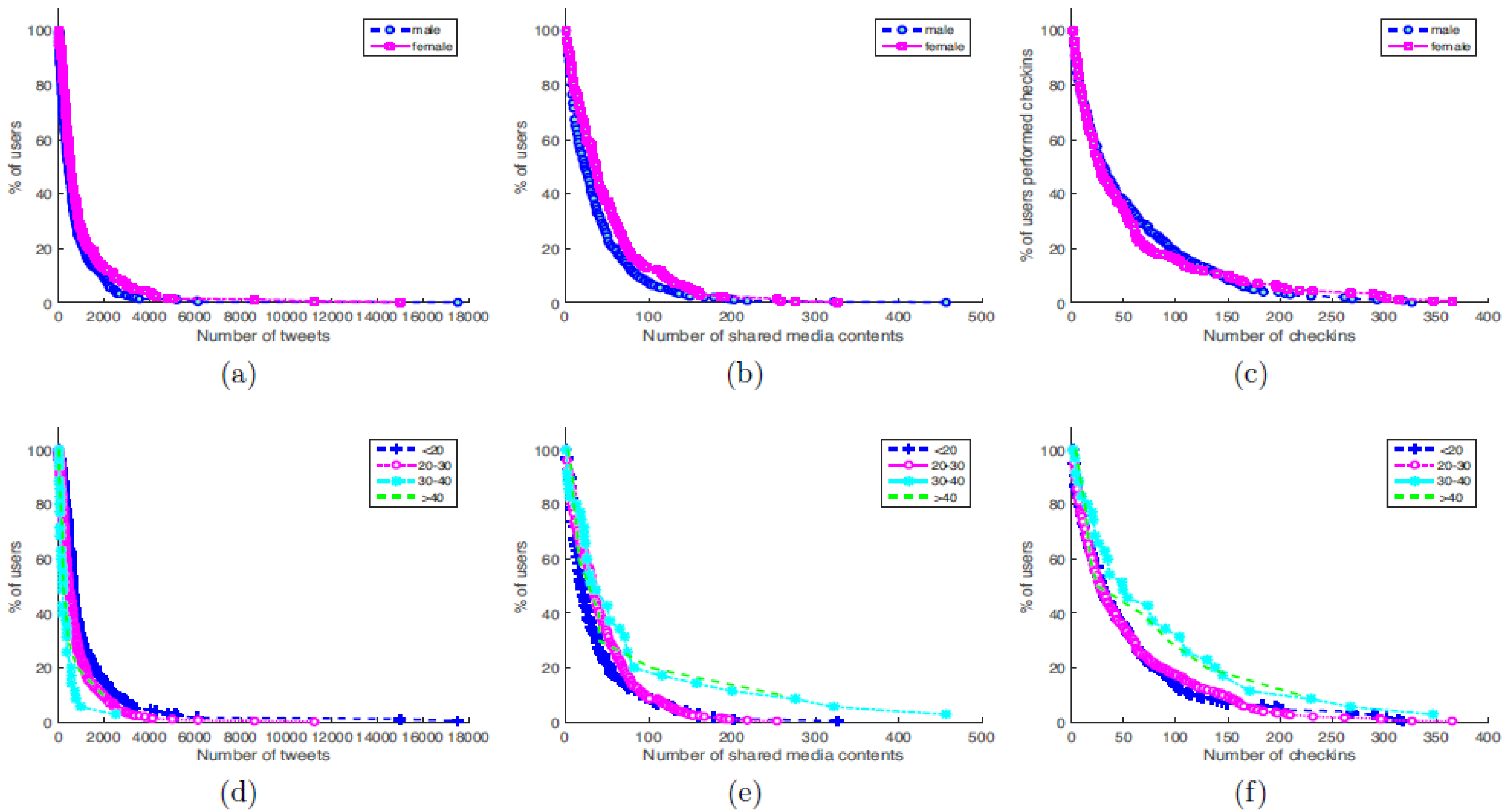
5,263,630 TWEETS

304,493 CHECK-INS

230,752 IMAGES

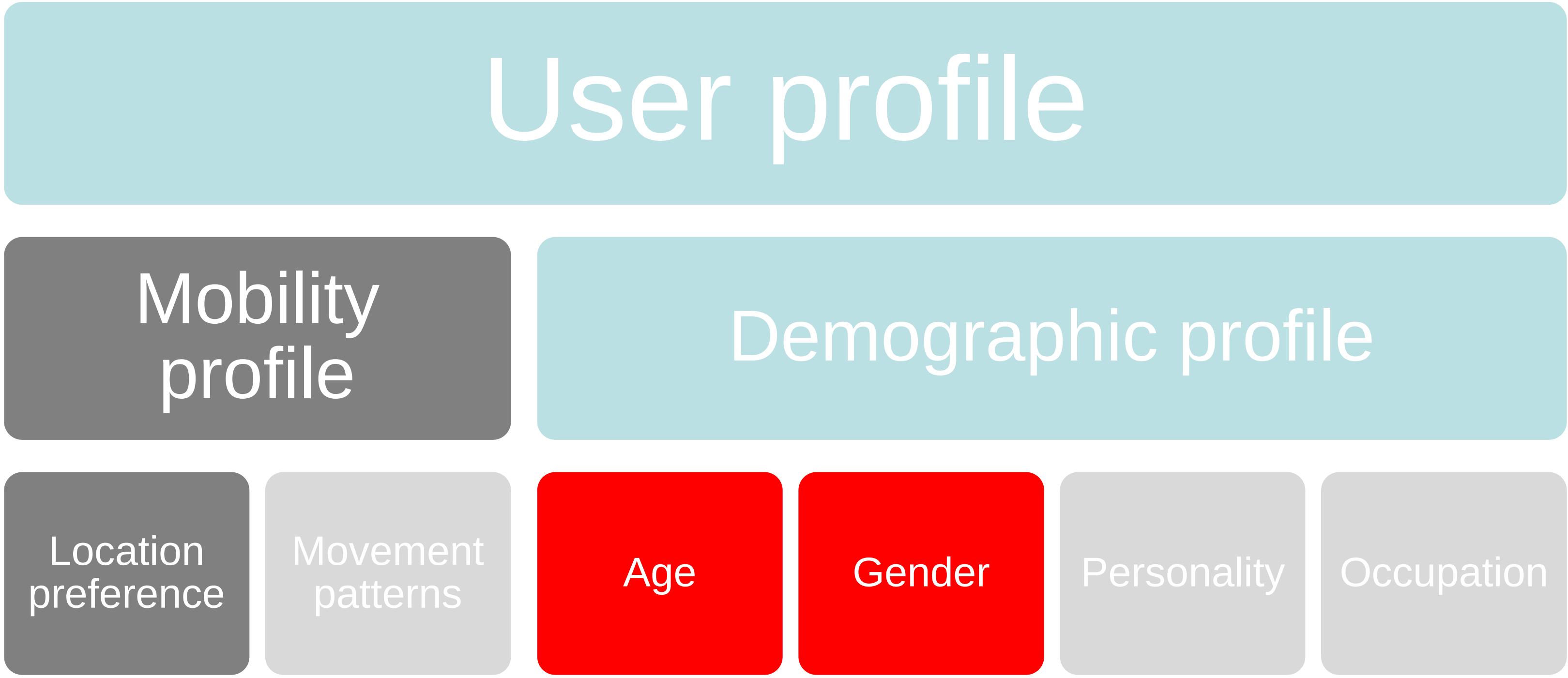
FROM 7,957 USERS

NUS-MSS: Dataset Statistics in Singapore



Demographic profiling

User profile: Mobility + Demography



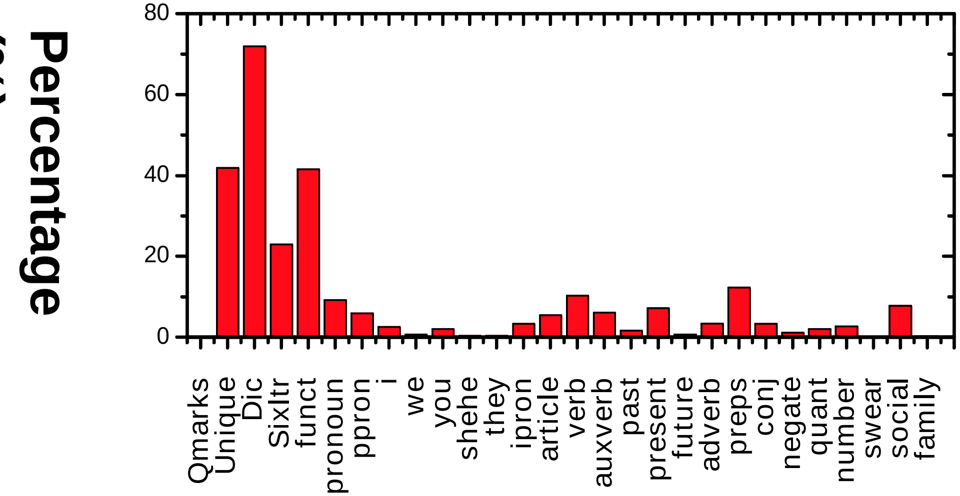
Data representation

- Linguistic features
  - **LIWC**
  - User Topics
- Heuristic features
  - Writing behavior

A text analysis software.

Dictionary

Word category



An efficient and effective method for studying the various emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples.

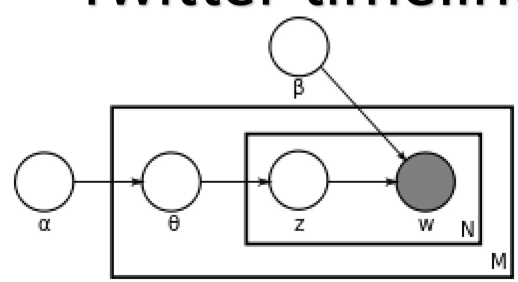
Can be highly related to one's demography.

Data representation

- Linguistic features
  - LIWC
  - **User Topics**
- Behavioral features
  - Writing behavior

Users of similar gender and age may talk about similar topics e.g. female users – about shopping, male – about cars; youth – about school while elderly – about health.

LDA word distribution over 50 topics for collected Twitter timeline.



Data representation

- Linguistic features
  - LIWC
  - User Topics
- Heuristic features
  - **Writing behavior**

As we mention from our research – user's **writing behavioral patterns** are **highly correlated** with e.g. age (individuals from **10 – 20 years** old are making **two times less grammatical errors** than **20 -30 years** old individuals)

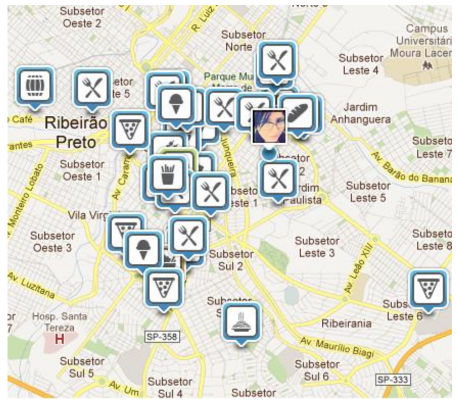
Feature name	Description
Number of hash tags	Number of hash tags mentioned in message
Number of slang words	Number of slang words one use in his tweets. We calculate number of slang words / tweet and compute average slang usage
Number of URLs	Number of URL's one usually use in his/her tweets
Number of user mentions	Number of user mentions – may represent one's social activity
Number of repeated chars	Number of repeated characters in one tweets (e.g. noooooooo, wahhhhhhh)
Number of emotion words	Number of words that are marked with not – neutral emotion score in Sentiment WordNet
Number of emoticons	Number of common emoticons from Wikipedia article
Average sentiment level	Module of average sentiment level of tweet obtained from Sentiment WordNet
Average sentiment score	Average sentiment level of tweet obtained from Sentiment WordNet
Number of misspellings	Number of misspellings fixed by Microsoft Word spell checker
Number Of Mistakes	Number of words that contains mistake but cannot be fixed by Microsoft Word spell checker
Number of rejected tweets	Number of tweets where 70% of words either not in English or cannot be fixed by Microsoft Word spell checker
Number of terms average	Average number of terms per / tweet
Number of Foursquare check-ins	Number of Foursquare check-ins performed by user
Number of Instagram medias	Number of Instagram medias posted by user
Number of Foursquare tips	Number of Foursquare Tips that user post in a venue
Average time between check-ins min	Average time between two sequential check-ins - represents Foursquare user activity frequency

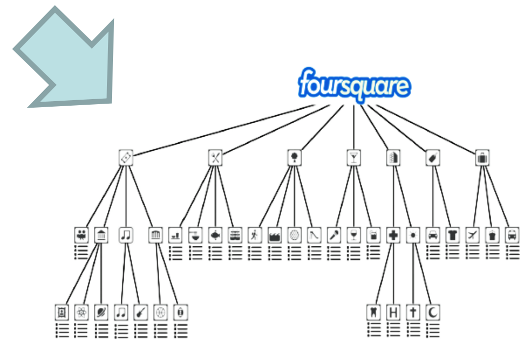
Data representation

- Location features
  - **Location semantics**
  - Location topics

Venue semantics such as venue categories can be related to **users demography**. E.g. individuals who tent to visit **night clubs** are usually belong to **10 – 20** or **20 – 30** years old age groups.

We map all Foursquare check – ins to Foursquare categories from category hierarchy.





For case when user performed check-ins in two restaurants and airport but did not perform check-ins in other venues:

	$Category_1$	...	$Category_{restaurant}$	...	$Category_{airport}$	...	$Category_n$
U_1	0	0	**2**	0	**1**	0	0
...	*	*	*	*	*	*	*
U_n	*	*	*	*	*	*	*

Data representation

- Image features
  - **Image concept**

Extracted image concepts may represents **user interests** and be related to one's demography. For example **female** user may take pictures of **flowers, food**, while **male** – of **cars or buildings**.

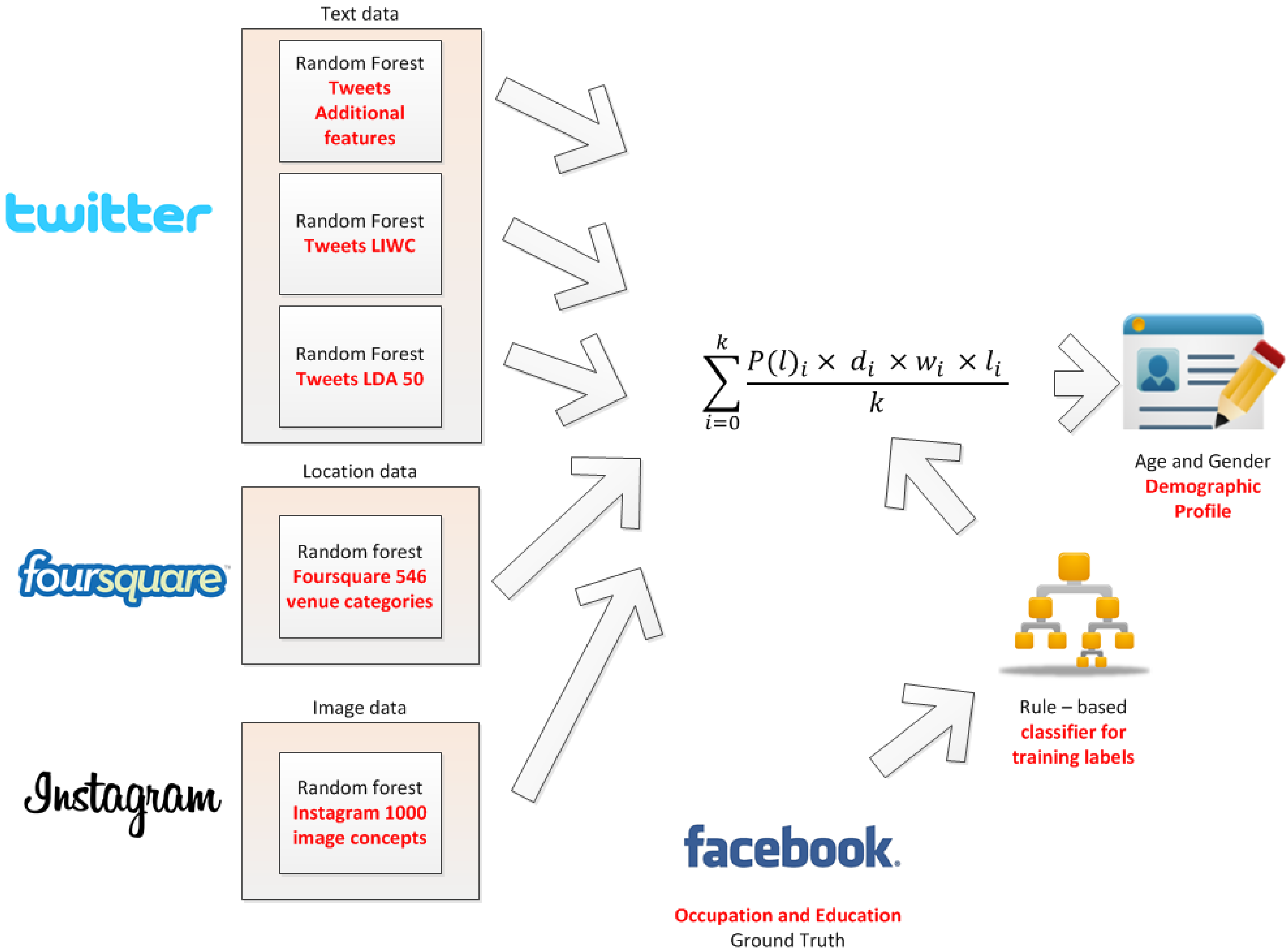
SIFT, LBP, CH FEATURES*

IMAGENET

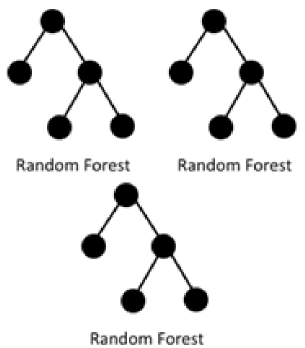
IMAGE CONCEPTS VECTOR

*The concept learning Tool was provided by Lab of Media Search LMS.
It was evaluated based on ILSVRC2012 competition dataset and performed with average accuracy @10 - 0.637

Ensemble learning

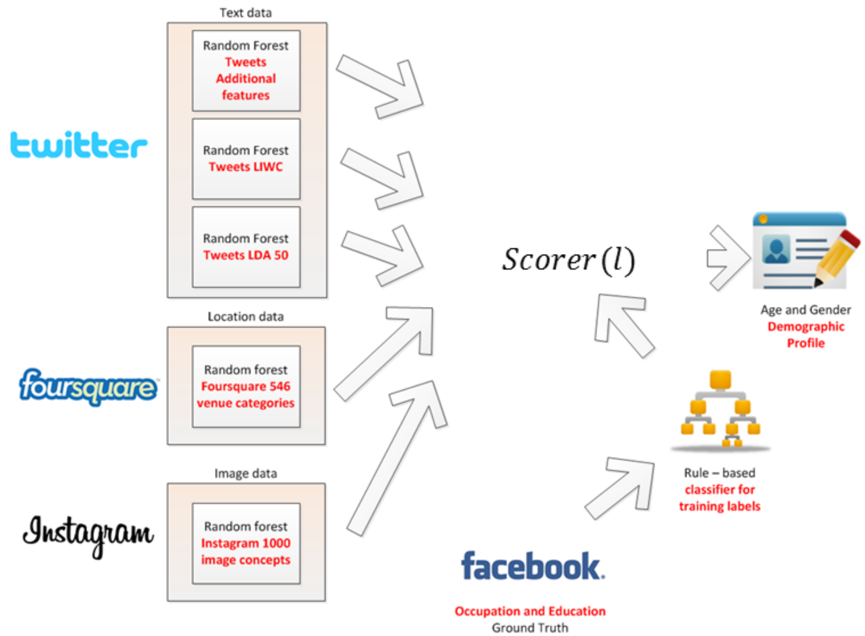


Ensemble learning



$$d_i \times w_i \times l_i$$

AGE, GENDER
CONFIDENCE
SCORES



$$Score(l) = \sum_{i=0}^{k} \frac{P(l)_i \times\ d_i\ \times w_i\ \times l_i}{k}$$

$P(l)_i$ - model prediction confidence
d_i - normalized data records number
w_i - model trust weight
l_i - model “strength” – learned by “Hill Climbing” optimization with step 0.05

Ensemble learning details

- According to our evaluation, the bias of estimated ages does not exceed ±2.28 years. It is thus reasonable to use the estimated age for age group prediction task.
- We have adopted SMOTE* oversampling to obtain balanced age-group labeling
- By performing 10-fold cross validation, we determine the optimal number of constructed random trees for each classifier with iteration step equal to 5 as 45, 25, 35, 40, 105 random trees for Random Forest Classifiers learned based on location, LIWC, heuristic, LDA 50, and image concept features respectively.
- We jointly learn the l_i model "strength" coefficient by performing "Hill Climbing" optimization* * with step 0.05. The randomized "Hill Climbing" approach is able to obtain local optimum for non-convex problems and, thus, can produce resolvable ensemble weighting.

*N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002.

**An iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution by incrementally changing a single element of the solution. If the change produces a better solution, an incremental change is made to the new solution, repeating until no further improvements can be found.

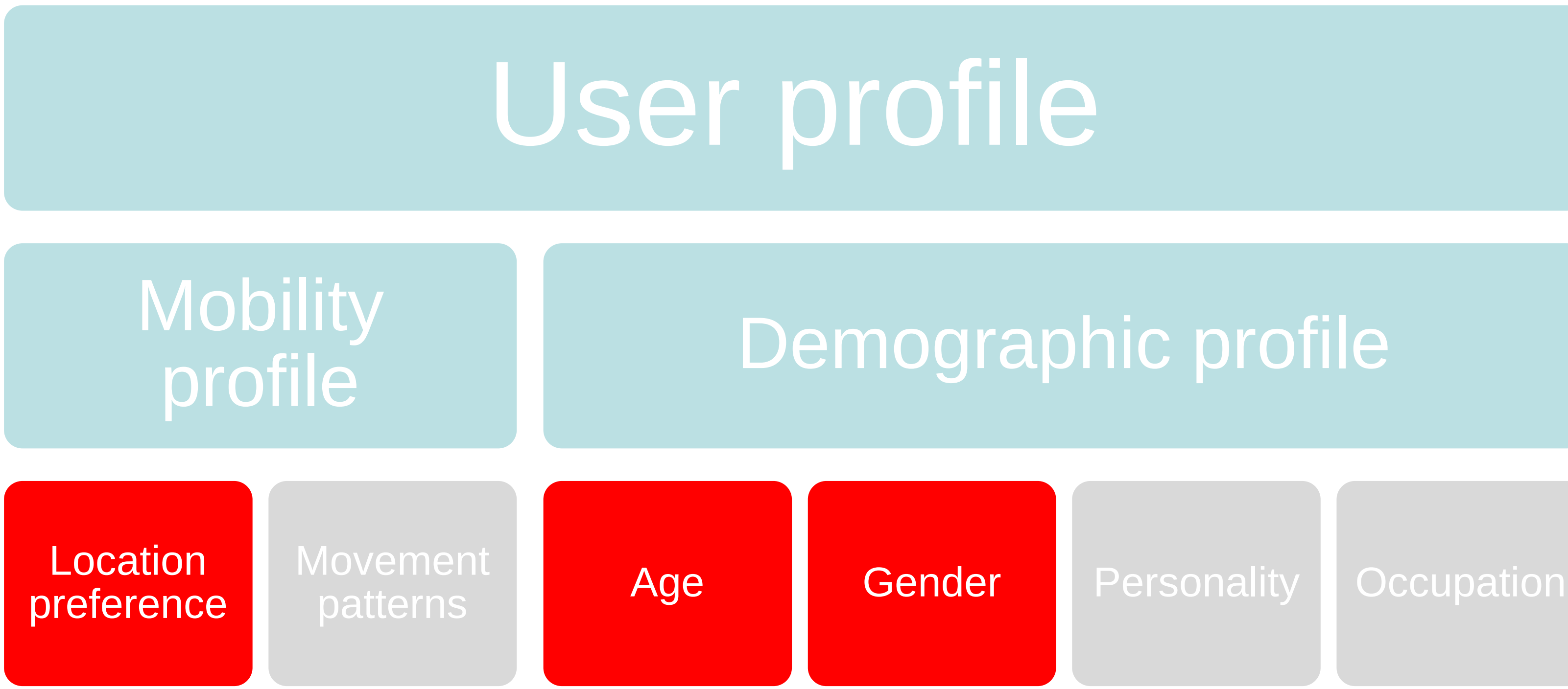
Experimental results (Singapore)

Method	Gender	Age
State-of-the-arts techniques		
SVM Location Cat. (Foursquare)	0.581	0.251
SVM LWIC Text(Twitter)	0.590	0.254
SVM Heuristic Text(Twitter)	0.589	0.290
SVM LDA 50 Text(Twitter)	0.595	0.260
SVM Image Concepts(Instagram)	0.581	0.254
NB Location Cat. (Foursquare)	0.575	0.185
NB LWIC Text(Twitter)	0.640	0.392
NB Heuristic Text(Twitter)	0.599	**0.394**
NB LDA 50 Text(Twitter)	**0.653**	0.343
NB Image Concepts(Instagram)	0.631	0.233
Single-Source		
RF Location Cat. (Foursquare)	0.649	0.306
RF LWIC Text(Twitter)	0.716	0.407
RF Heuristic Text(Twitter)	0.685	**0.463**
RF LDA 50 Text(Twitter)	**0.788**	0.357
RF Image Concepts(Instagram)	0.784	0.366
Multi-Source combinations		
RF LDA + LIWC(Late Fusion)	0.784	0.426
RF LDA + Heuristic(Late Fusion)	0.815	0.480
RF Heuristic + LIWC (Late Fusion)	0.730	0.421
RF All Text (Late Fusion)	0.815	0.425
RF Media + Location (Late Fusion)	0.802	0.352
RF Text + Media (Late Fusion)	**0.824**	**0.483**
RF Text + Location (Late Fusion)	0.743	0.401
All sources together		
RF Early fusion for all features	0.707	0.370
RF Multi-source (Late Fusion)	**0.878**	**0.509**

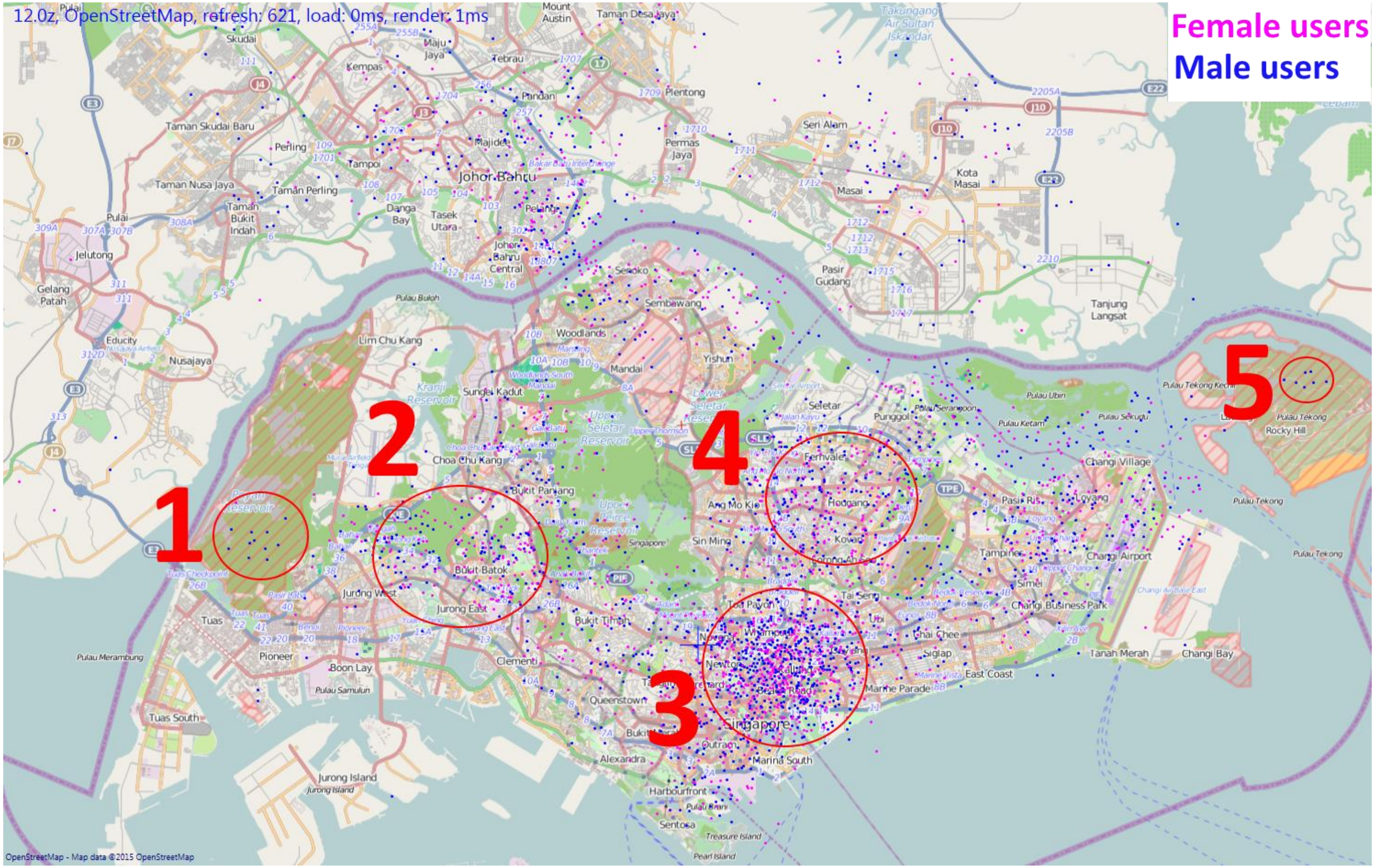
AGE GROUPS:
< 20 YEARS OLD,
20 – 30 YEARS OLD,
30 – 40 YEARS OLD,
> 40 YEARS OLD

Demographic mobility

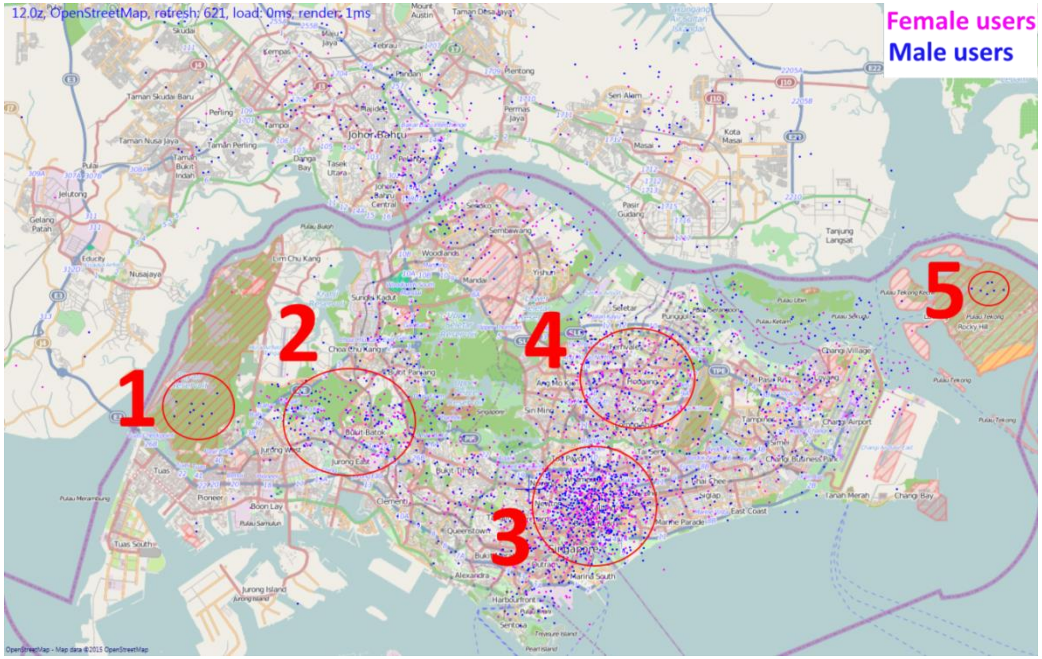
User profile: Mobility + Demography



Geographical user mobility: users movement (city level)

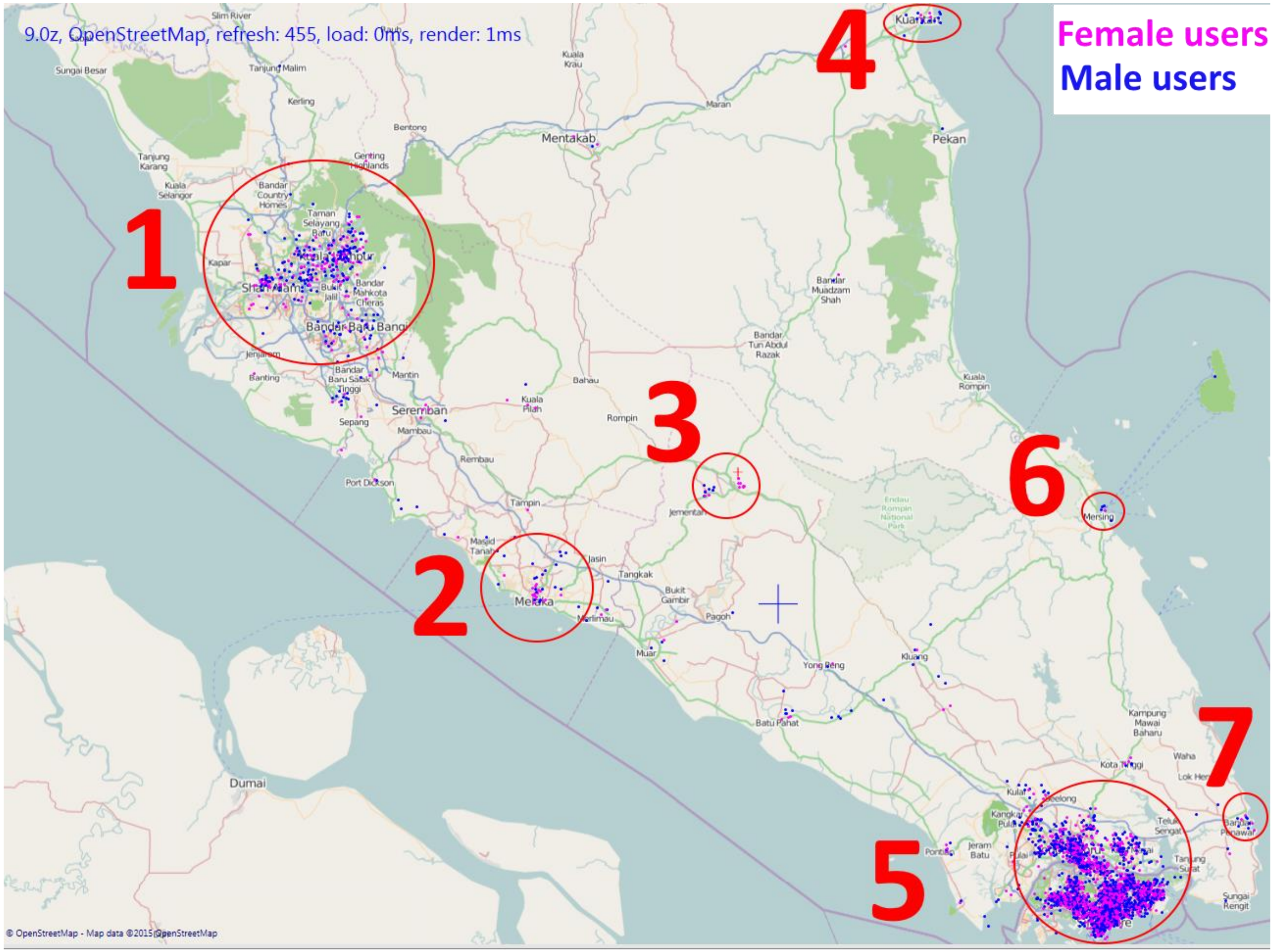


Geographical user mobility: users movement (city level)

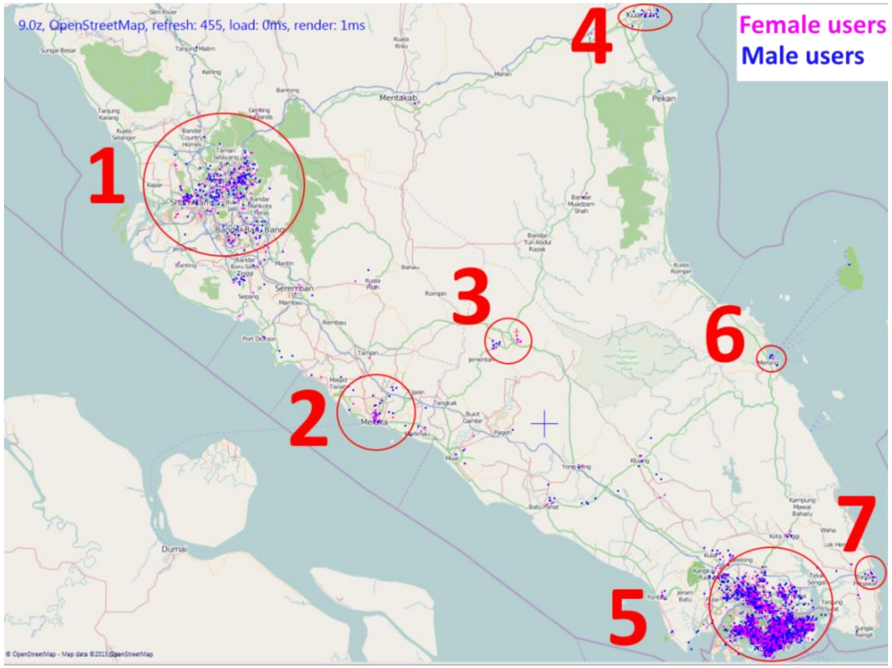


- Singapore population is concentrated in several regions, which represent peoples' housing (Regions 2 and 3) and working (Region 3) areas.
- There are some regions where male (Blue markers) user check-in density is much higher than female (Pink markers).

Geographical user mobility: users movement (region level)



Geographical user mobility: users movement (region level)



- Both female and male users often perform trips to nearby cities for shopping and leisure purposes (Regions 1, 2, 4, 5).
- Regions 2 and 3 are popular among female users, since 2 is "Malacca resorts", while 3 – National park. Both regions are famous by it's family time spending facilities.

Geographical user mobility: users movement (city level)

Geographical user mobility: users movement (city level)

- Teenagers and children (Brown markers) mostly perform check-ins in housing city areas and around schools (Regions 1,2,3,5).
- Students (Green markers) and working professionals (Blue and Red markers) are concentrated in city center (Region 4).

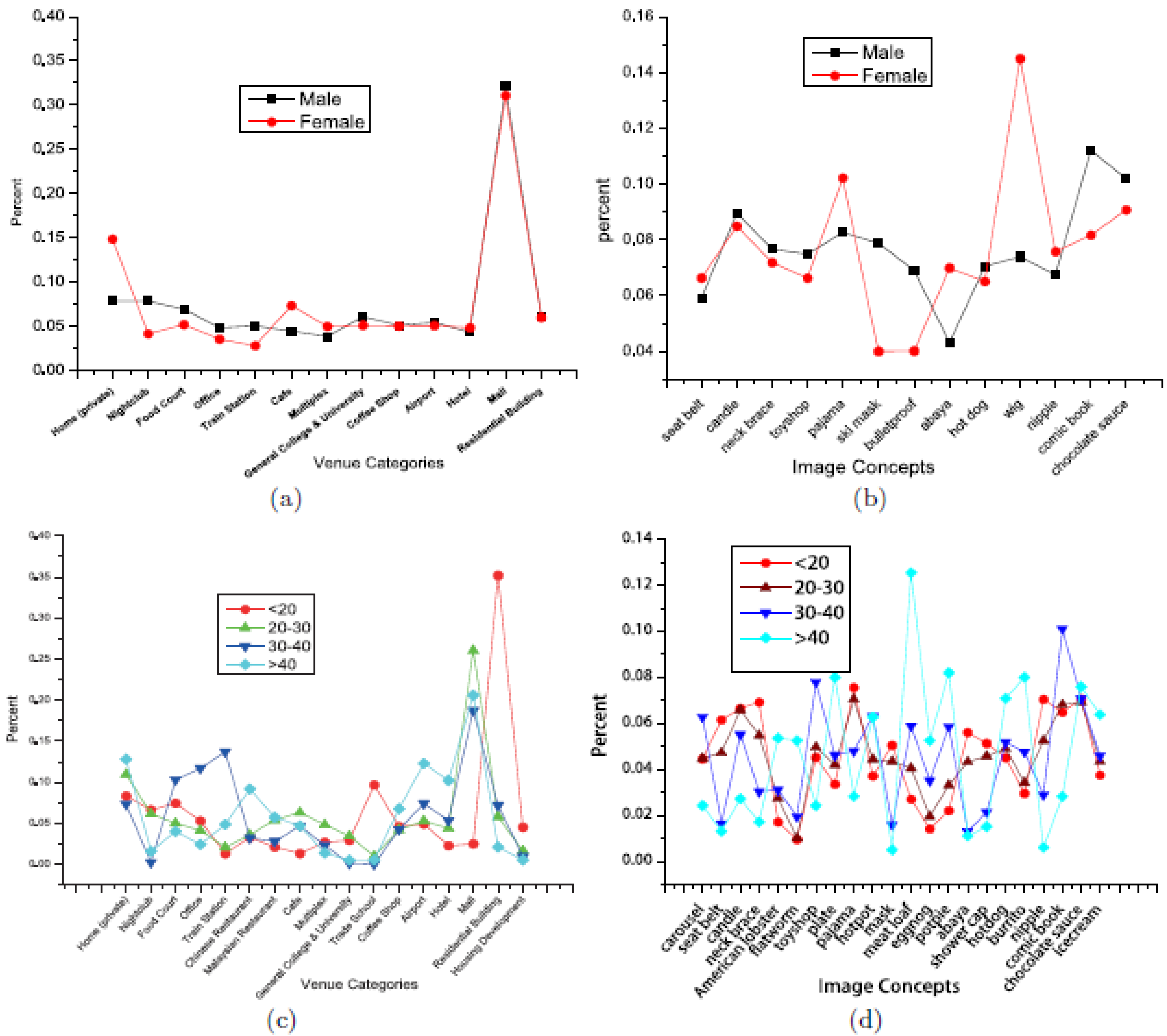
38

Geographical user mobility: users movement (region level)

Geographical user mobility: users movement (region level)

- Young users (brown circles) are rarely travel to nearby cities due to their age (Region 3)
- Adults (green circles) often make such trips (Regions 1 and 2). These users may be students or young professionals who visit their families during weekends.

Dataset Statistics: Content



(a) (b) (c) (d)

Geographical user mobility: venue semantics profiling

- We extract location topics based on venue categories to model user mobility semantics

LDA word distribution over 6 topics for collected Foursquare check-ins.

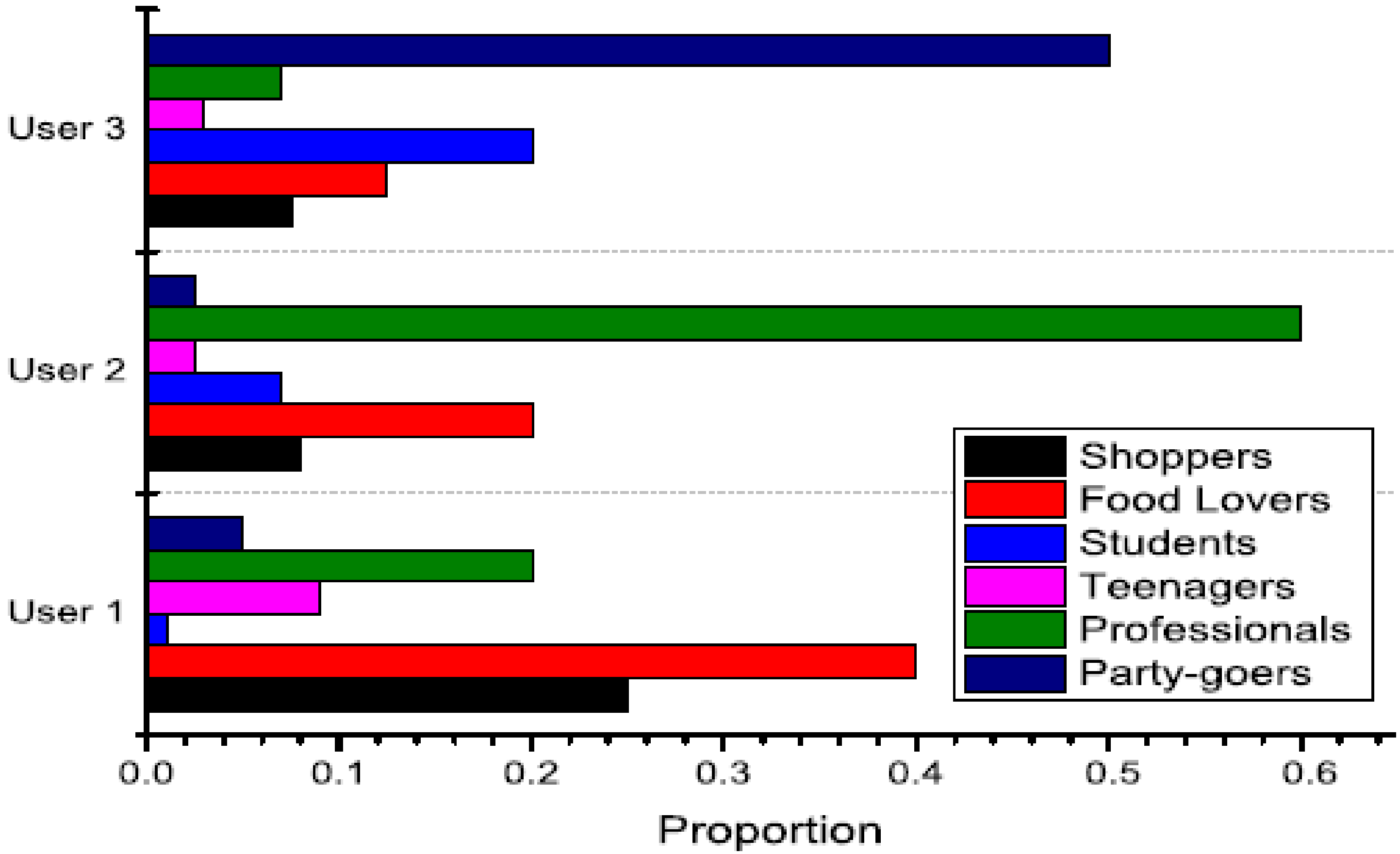
Location topics may serve as an user interest clusters for distinguishing user demography attributes such as age or gender.

Every venue category

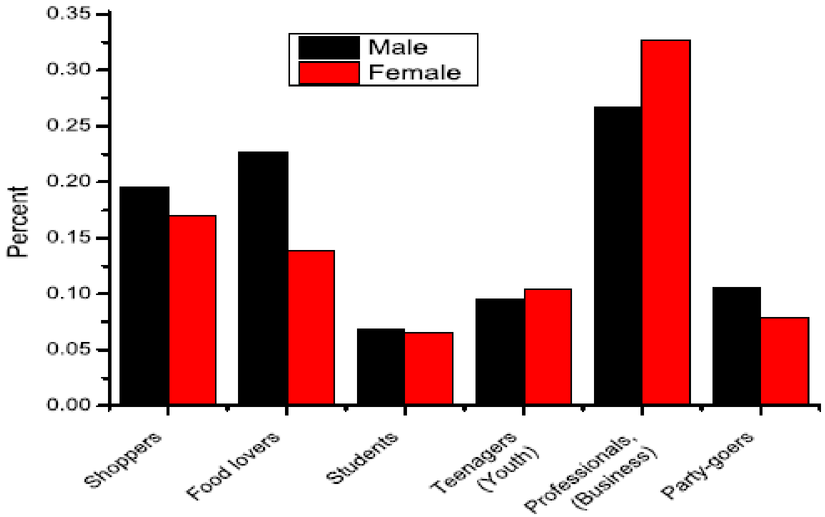
Table 2: Category distribution among LDA topics

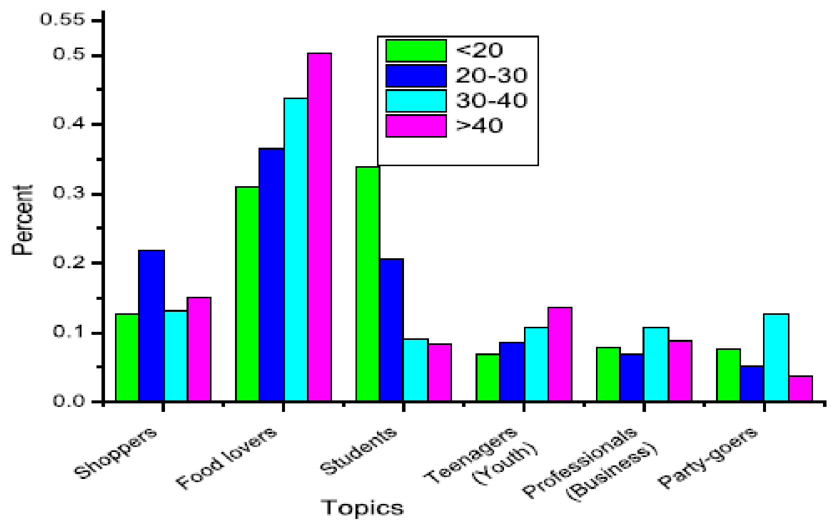
ID	Categories	LDA Topics
T1	Malay Res-t, Mall, University, Indian Res-t, Aisian Res-t	Food Lovers
T2	Cafe, Airport, Hotel, Coffee Shop, Chinese Res-t	Travelers (Business)
T3	Nightclub, Mall, Food Court, Trade School, Res-t, Coffee Shop	Party Goers
T4	Home, Office, Build., Neighbor-d, Gov. Build., Factory	Family Guys (Youth)
T5	University (Collage), Gym, Airport, Hotel, Fitness Club	Students
T6	Train St., Apartment, Mall, High School, Bus St.	Teenagers (Youth)

Geographical user mobility: venue semantics profiling



Geographical user mobility: venue semantics profiling

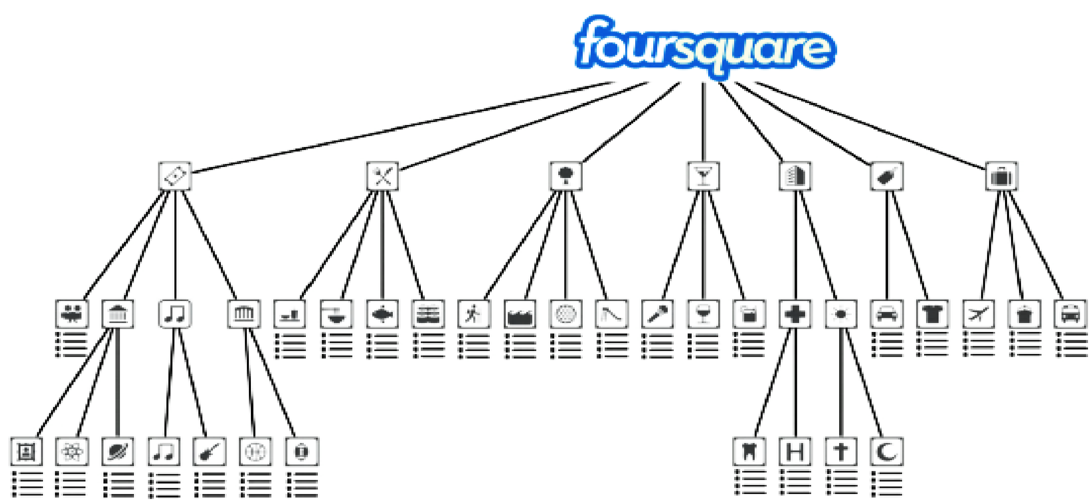




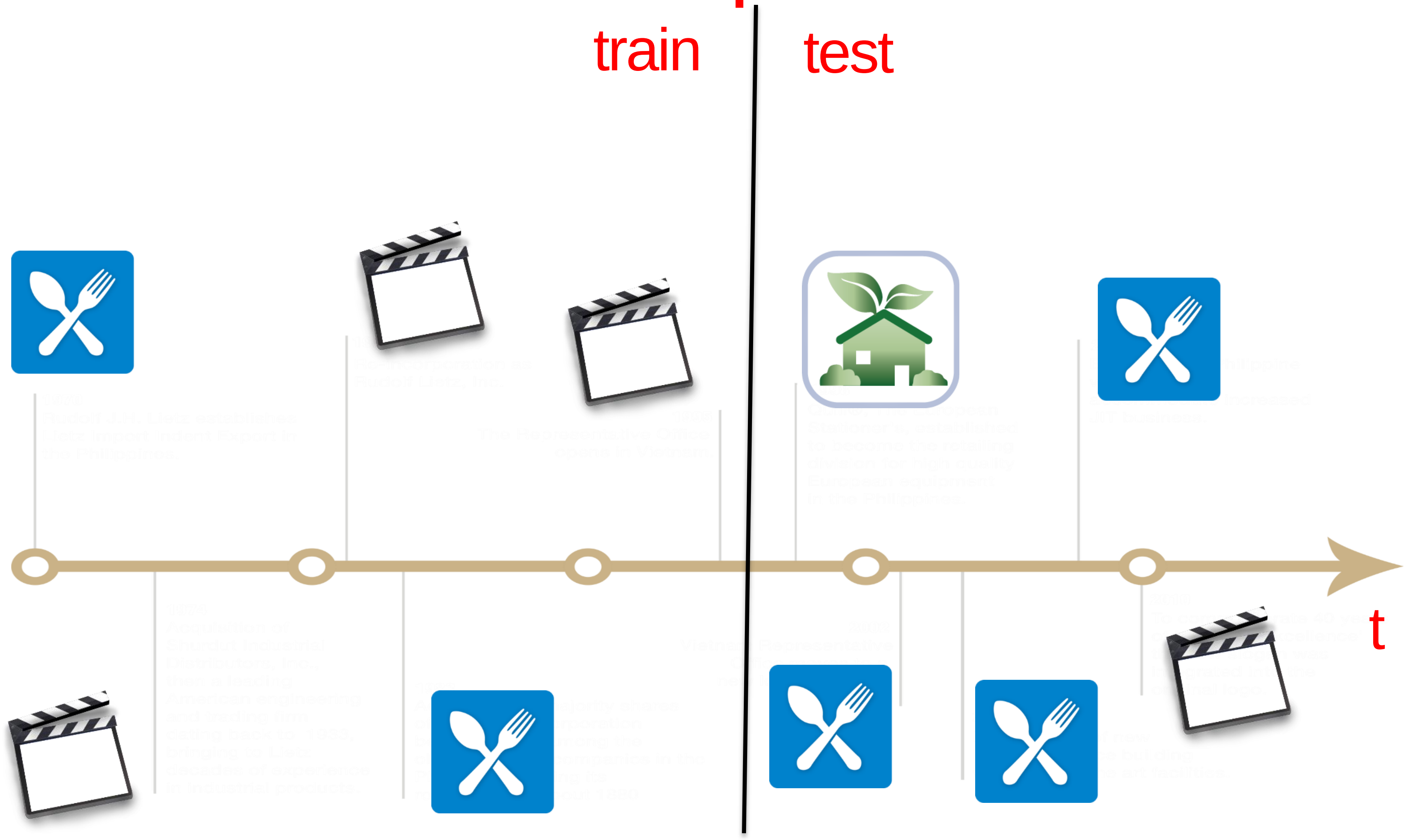
- Male users more often do shopping than male, while female users often show-up in job-related venues.
- > 30 years old users often show-up in dining-related places, while < 20 – often visit education-related venues.

Venue Category Recommendation

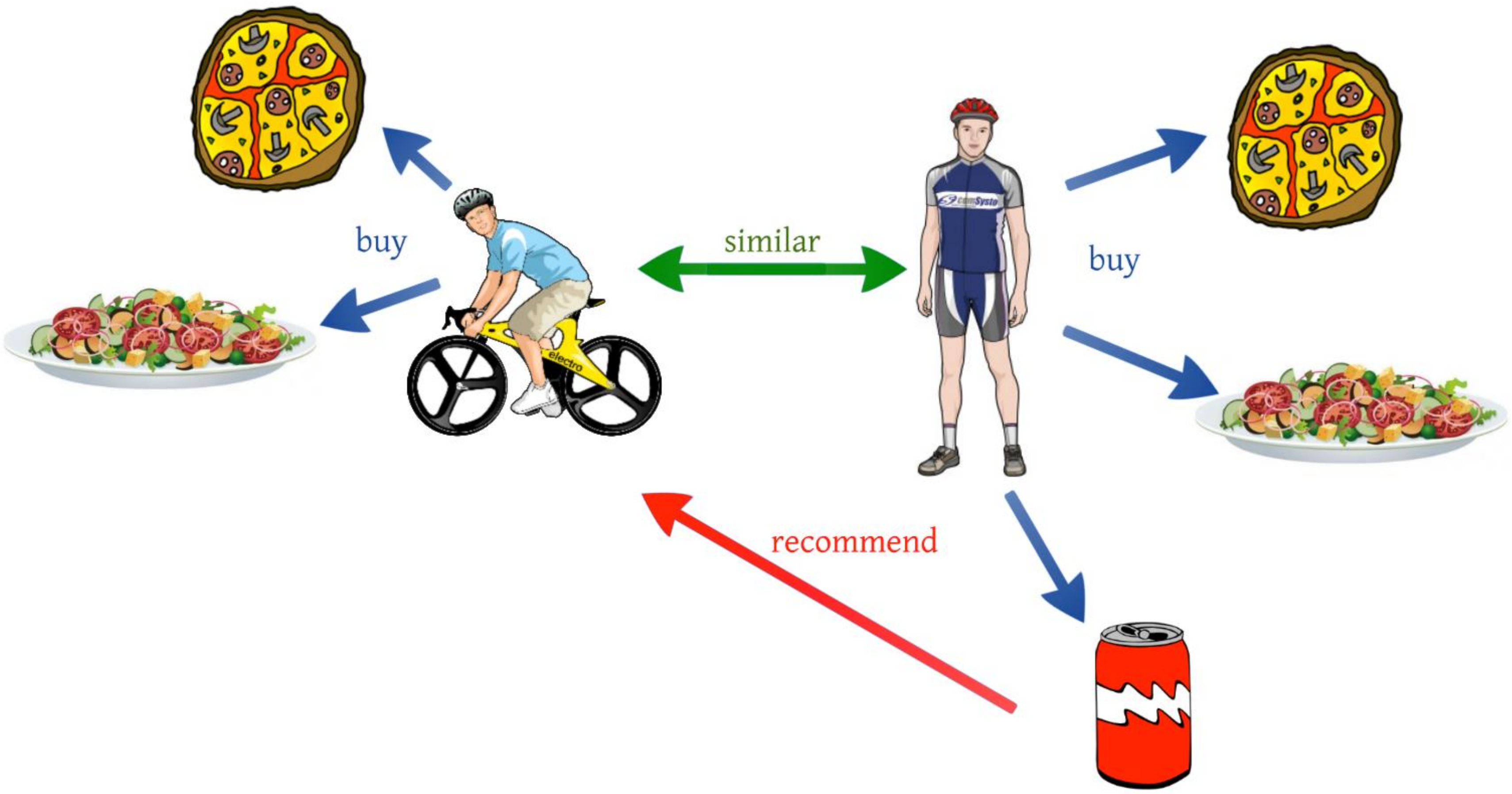
Which category(s) of 4sq venues to go next?



Evaluation – split time on train and test periods



We use Collaborative Filtering (CF)



Multi-Source re-ranking

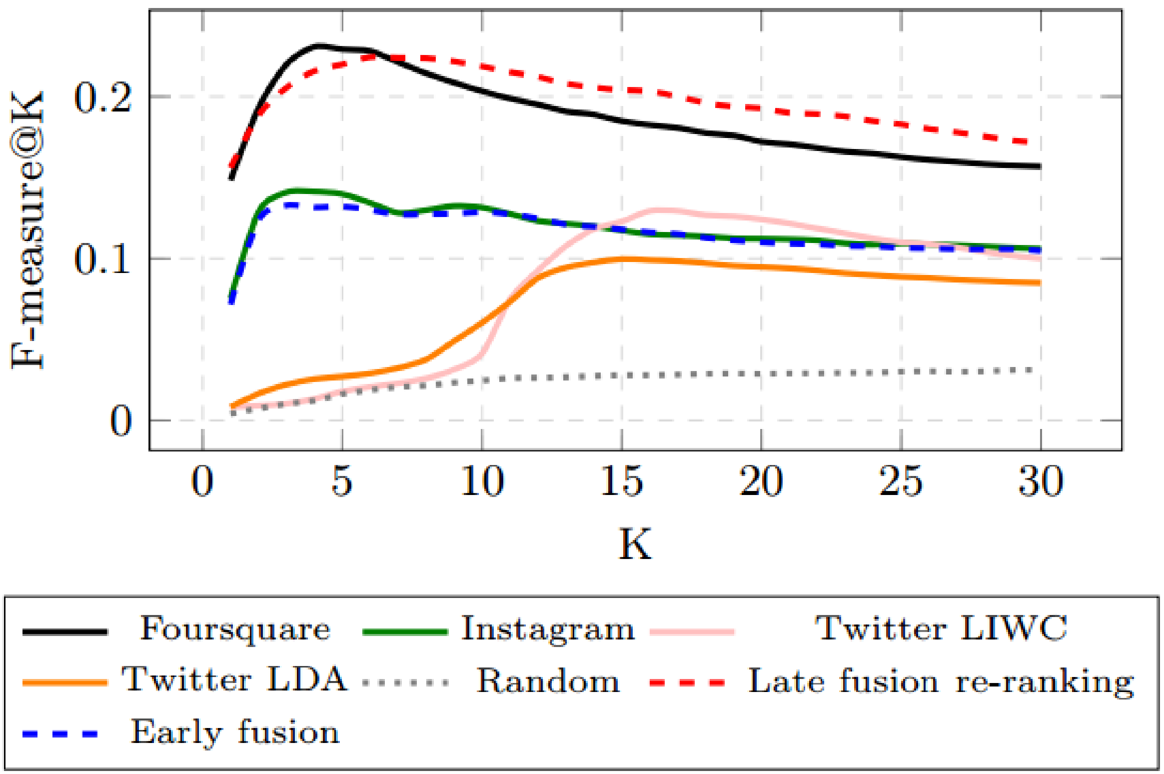
- Seeking to boost the recommendation performance, we developed late fusion re-ranking approach. We linearly combined the outputs from different sources, where the weight of each source is learned based on a stochastic hill climbing with random restart (SHCR)

$$Rank_f(item_i) = \frac{1}{n}\sum_{s=1}^{n}\frac{w_s}{Rank_s(item_i)}$$

- where $Rank_s(item_i)$ is the rank of ith item in recommendation list for source s; w_s corresponds to the weight of the source s; n is a total number of sources (in our case, n = 4). The venue categories in final recommendation list are sorted in increasing order according to their rank.

Results

$$F - measure@K = \frac{2 \cdot P@K \cdot R@K}{P@K + R@K}$$



- To measure the recommendation performance we use F-measure@K, where P@K and R@K are precision and recall at K, respectively, and K indicates the number of selected items from the top of the recommendation list.

What we are doing now?

Something much bigger…

You, actually can join us as Intern or Research Engineer

http://next.comp.nus.edu.sg/opportunities

Extended User Profiling

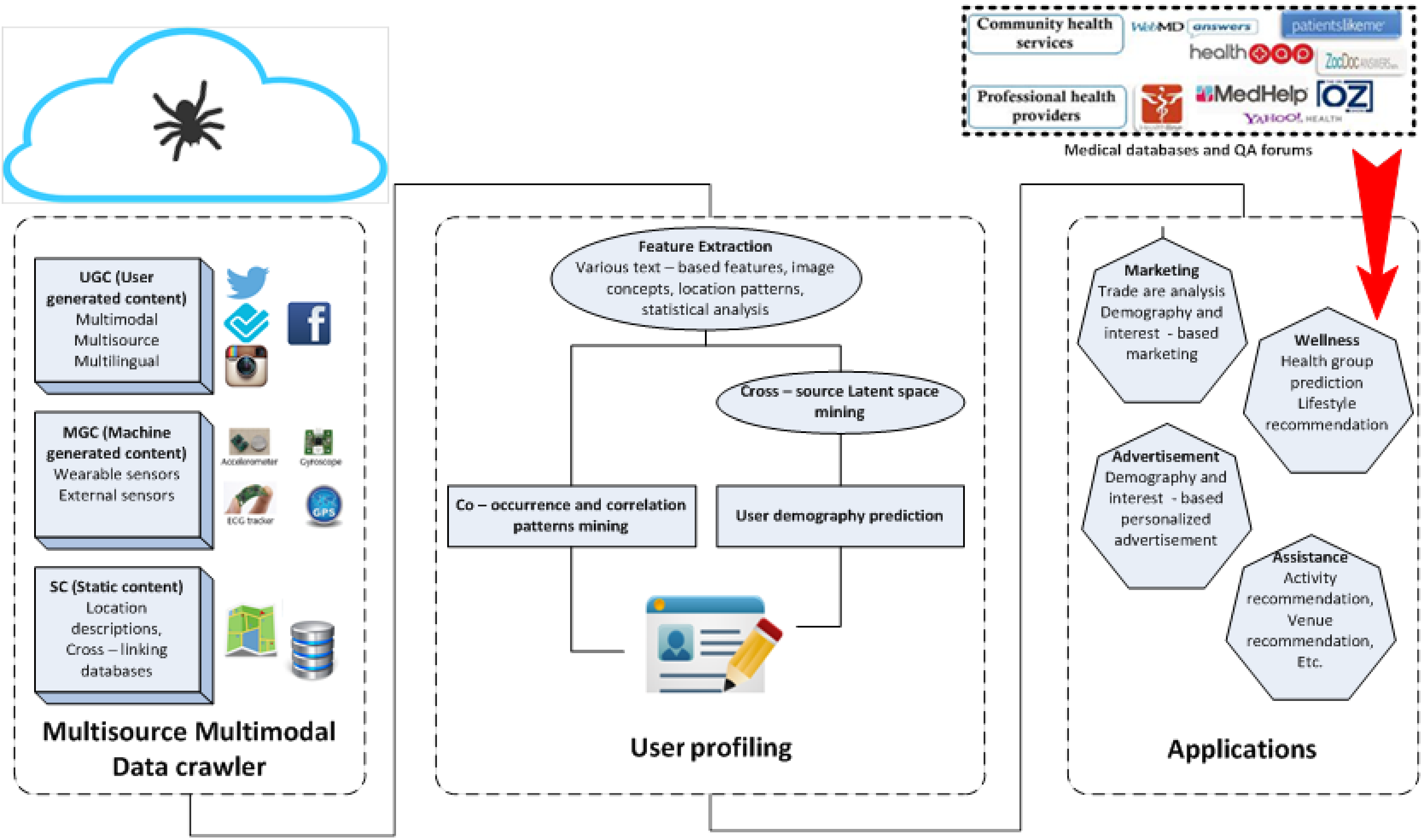
- Extended Demographic Profiling:
  - ~~Occupation detection~~;
  - Personality detection;
  - Social status detection.
- Extended Mobility Profiling :
  - User communities detection and profiling (In terms of demographics, movement patterns, multi-source interests) – in progress
  - Cross-region mobility profiling (comparison of users' mobility across different regions and cultures) – in progress

Sensor Data Incorporation & Wellness Research

- Wellness lifestyle recommendation via:
  - Chronic diseases tendency prediction
  - Cross-source causality relationships analysis (just like Ramesh Jain proposed*)

*Ramesh Jain, Laleh Jalali: Objective Self. IEEE MultiMedia 21(4): 100-110 (2014)

Future work: How the framework may look like



Other tasks based could be approached

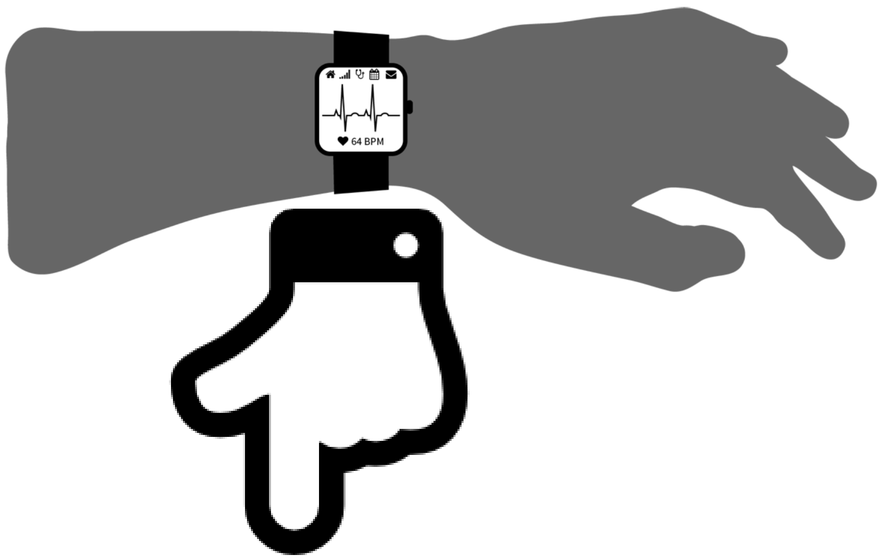
1. **Demographic profile learning**
2. **Multi-source data fusion**
3. **Individual and group mobility analysis**
4. **Cross-source user identification**
5. **Cross-region user community detection**
6. **Cross-source causality relationships extraction**
7. **Users' privacy-related and cross-disciplinary research**

User Profile Learning in Wellness Domain

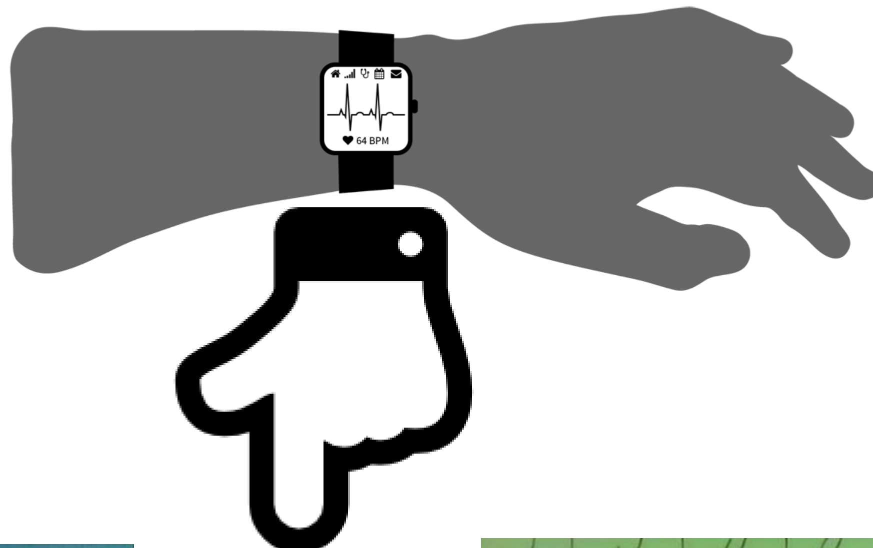
People are often now aware of their wellness problems

It is not easy to follow doctor's prescriptions

Personal and continuous assistance is necessary



Continuous patients monitoring for better prescription

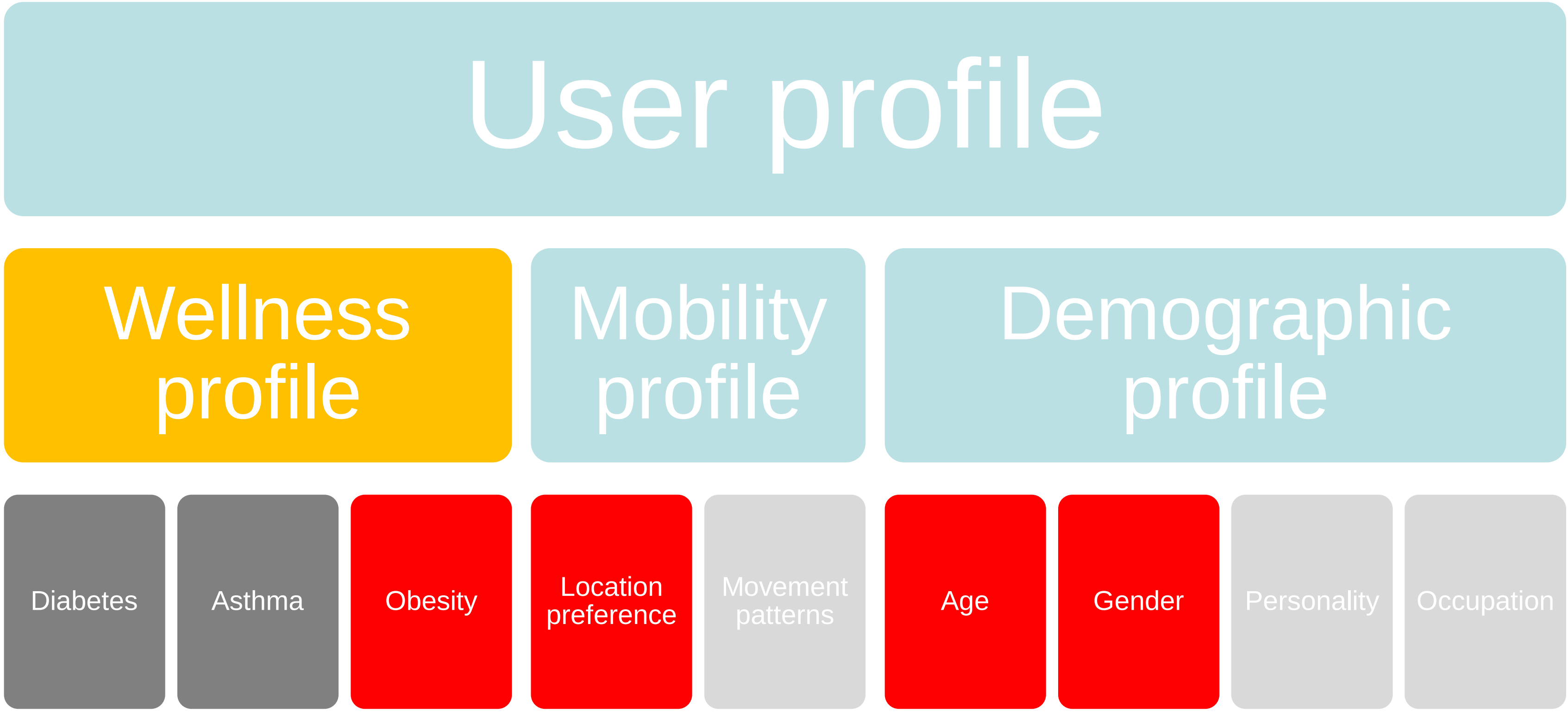


Weight Problems Consequences*

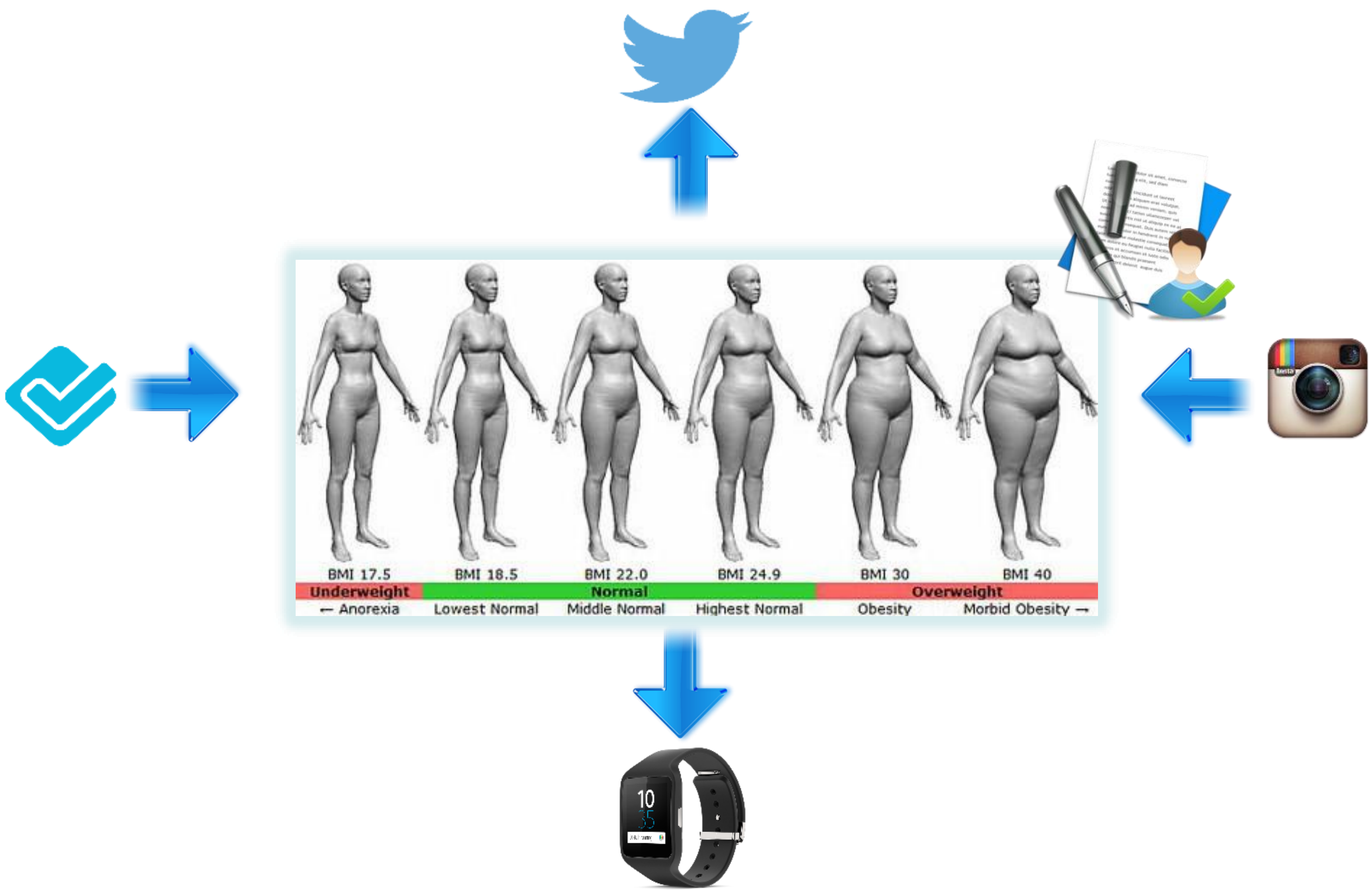
- All-causes of death (mortality)
- **High blood pressure (Hypertension)**
- High LDL cholesterol, low HDL cholesterol, or high levels of triglycerides (Dyslipidemia)
- **Type 2 diabetes**
- **Coronary heart disease**
- Stroke
- Gallbladder disease
- **Osteoarthritis (a breakdown of cartilage and bone within a joint)**
- Sleep apnea and breathing problems
- **Some cancers (endometrial, breast, colon, kidney, gallbladder, and liver)**
- Low quality of life
- **Mental illness such as clinical depression, anxiety, and other mental disorders**
- Body pain and difficulty with physical functioning[6]

*Health effect of overweight and obesity. *Center of disease control and prevention*. http://www.cdc.gov/healthyweight/effects/

User Profiling: Next Step



Data sources describe user in multiple views



Research Problems

- Multi-source user profiling:
  - Wellness profiling
    - Predict one's obesity level by leveraging multi-source multi-modal data (in other words – BMI prediction)
  - Data gathering, noise, sensitivity and incompleteness
  - Multi–source multi–modal data integration

Summary

- We constructed and released a large multi-source multi-modal cross-region “NUS-MSS” dataset;
- We conducted first-order and higher-order learning for user mobility and demographic profiling;
- New multi-modal features were proposed for a demographic profile learning.
- Based on our experimental results, we can conclude that multi-source data mutually complements each other and their appropriate fusion boosts the user profiling performance.
- We believe that we can predict one’s social media data and the data from wearable sensors.

Next Lesson

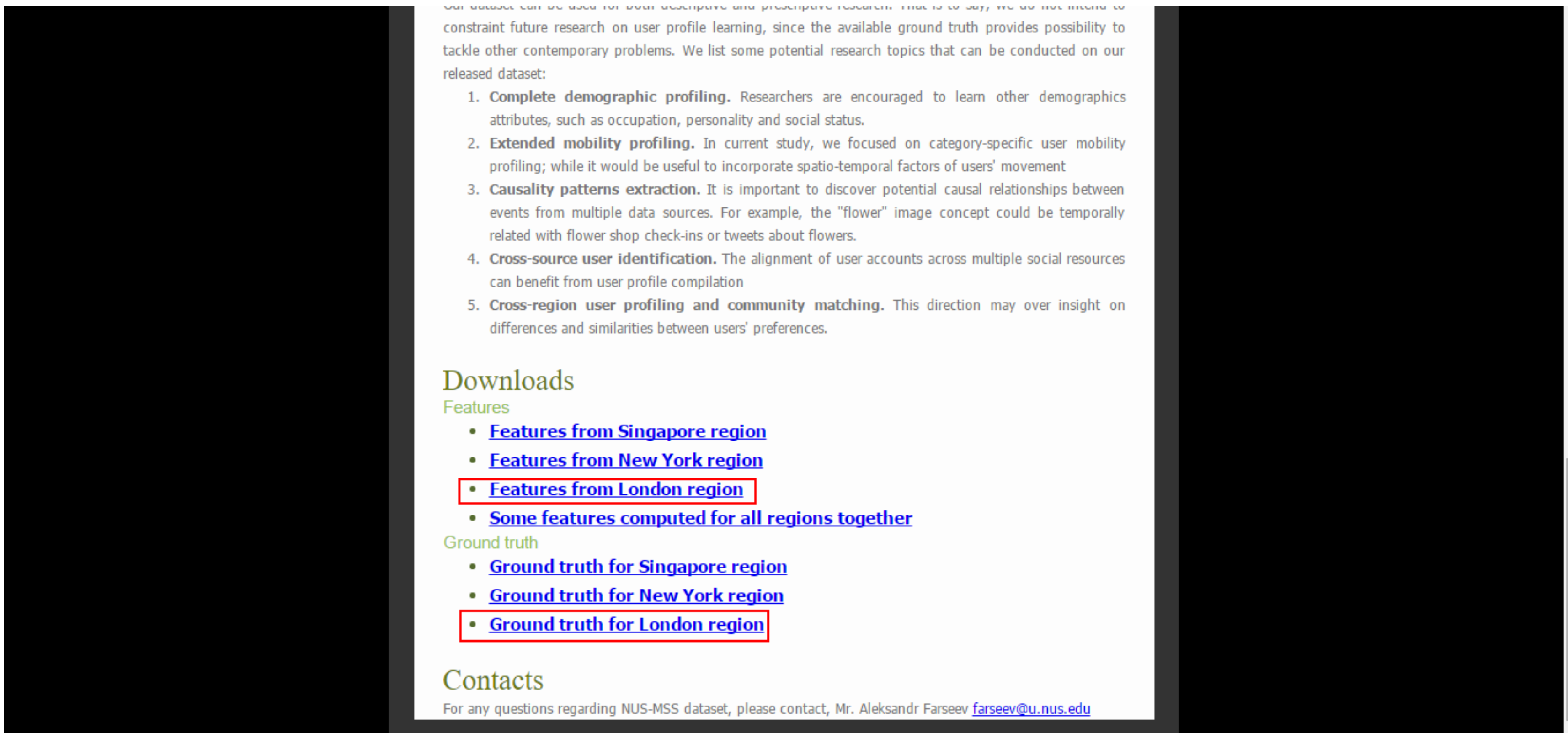
- **Wrap Up**

Short KNIME* Tutorial

1. *We select Knime since it could be used even without any programming experience.

2. Download and Install Knime together with all extensions from here: http://knime.org
3. Go to http://nusmultisource.azurewebsites.net
4. Download all the Features and Ground Truth from 3 cities: Singapore, London. New York

Further steps are based on London dataset, but it is applicable to all the other sources.



constraint future research on user profile learning, since the available ground truth provides possibility to tackle other contemporary problems. We list some potential research topics that can be conducted on our released dataset:

1. **Complete demographic profiling.** Researchers are encouraged to learn other demographics attributes, such as occupation, personality and social status.
2. **Extended mobility profiling.** In current study, we focused on category-specific user mobility profiling; while it would be useful to incorporate spatio-temporal factors of users' movement
3. **Causality patterns extraction.** It is important to discover potential causal relationships between events from multiple data sources. For example, the "flower" image concept could be temporally related with flower shop check-ins or tweets about flowers.
4. **Cross-source user identification.** The alignment of user accounts across multiple social resources can benefit from user profile compilation
5. **Cross-region user profiling and community matching.** This direction may over insight on differences and similarities between users' preferences.

Downloads

Features

- Features from Singapore region
- Features from New York region
- Features from London region
- Some features computed for all regions together

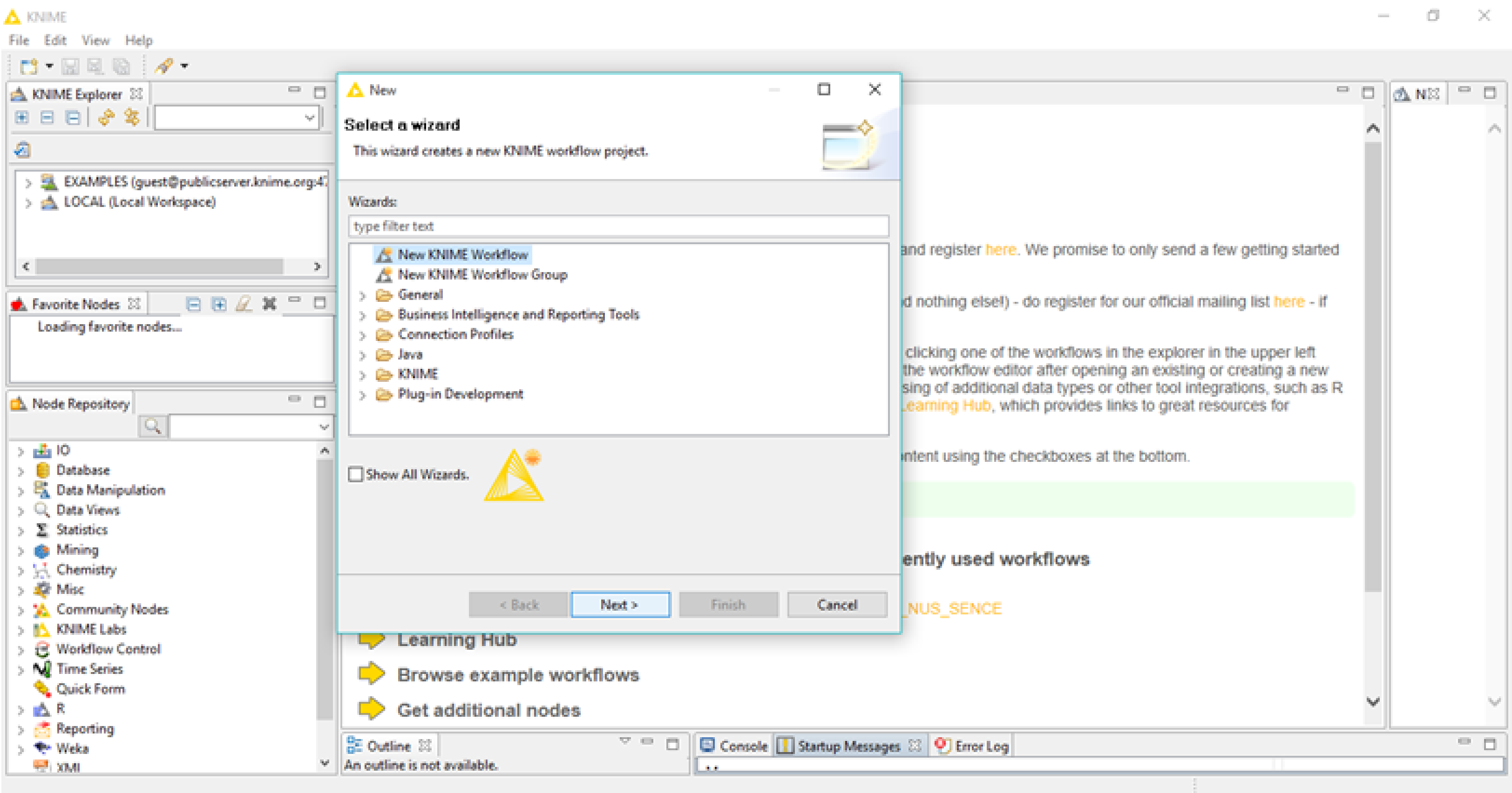
Ground truth

- Ground truth for Singapore region
- Ground truth for New York region
- Ground truth for London region

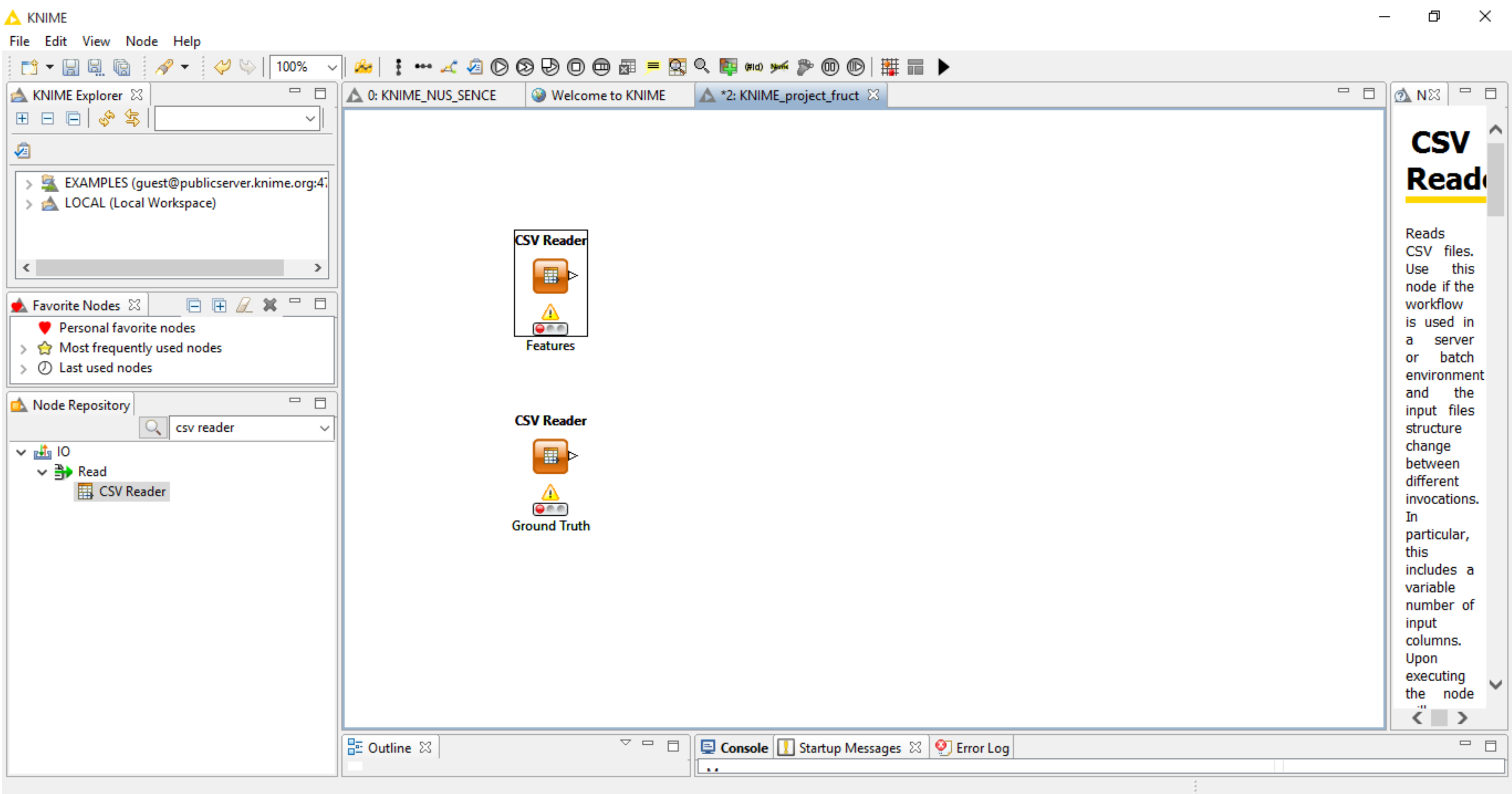
Contacts

For any questions regarding NUS-MSS dataset, please contact, Mr. Aleksandr Farseev farseev@u.nus.edu

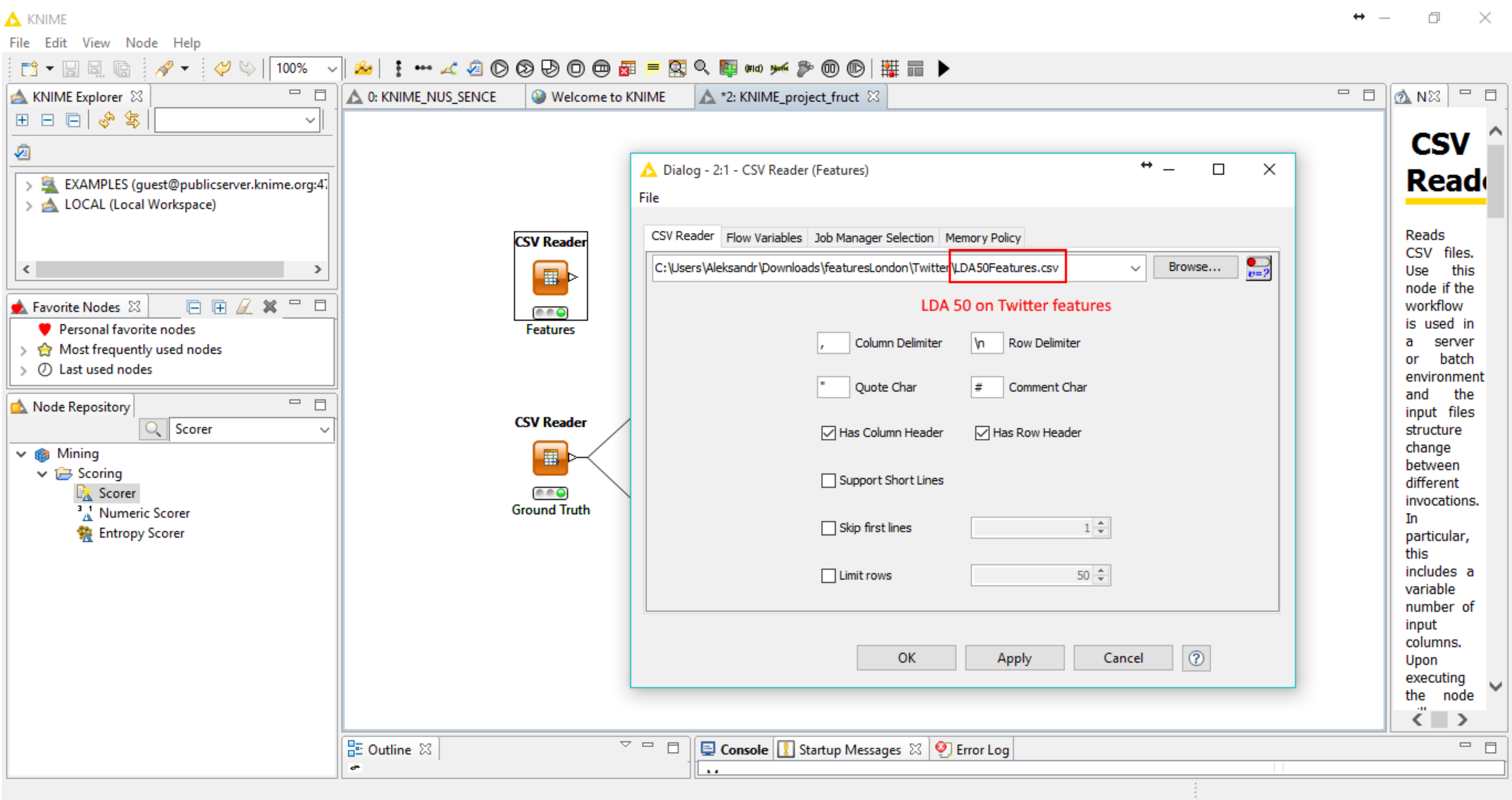
Open Knime and Create New Workflow



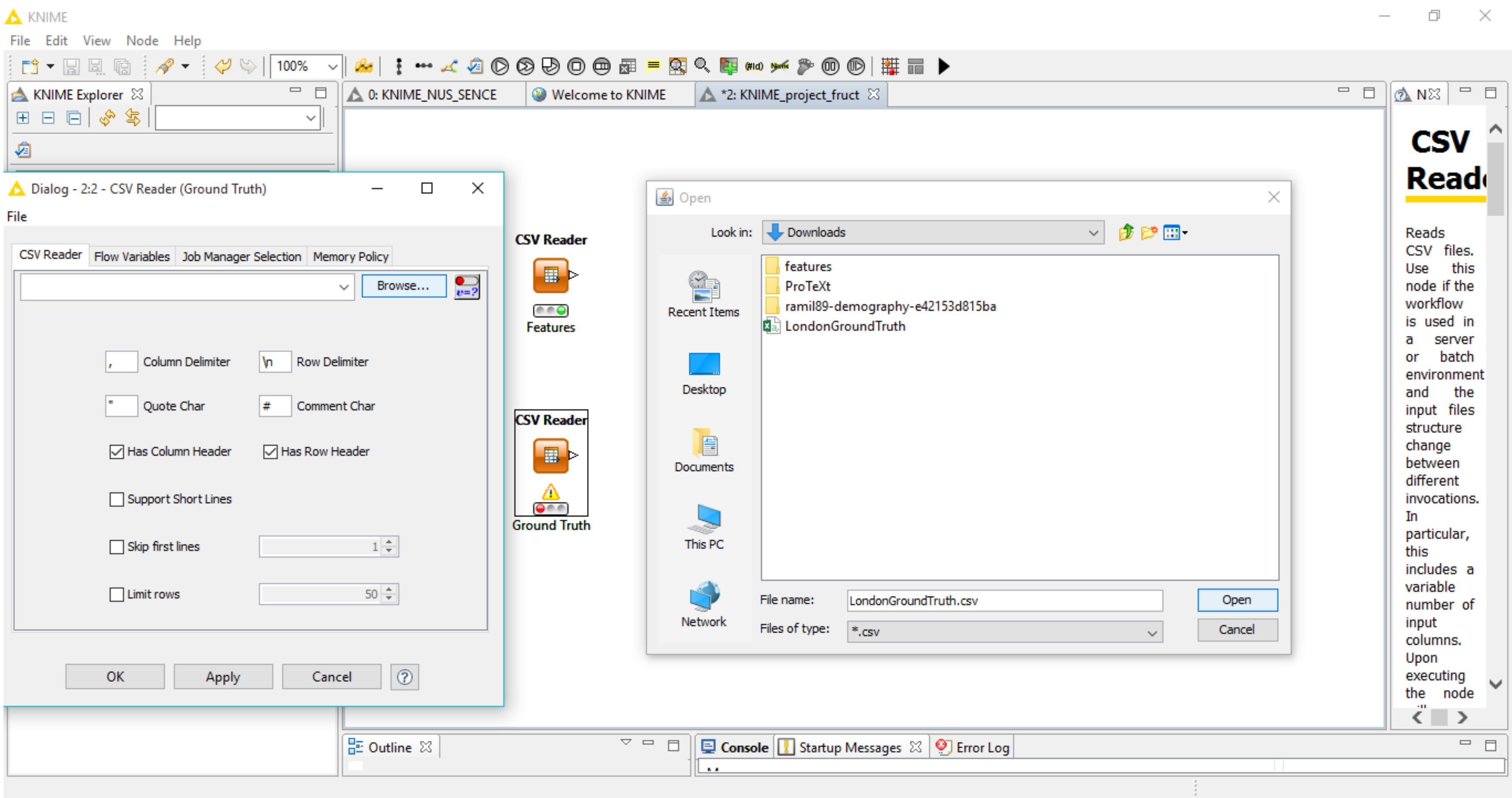
Add two “CSV Reader” nodes – one for features, one for ground truth



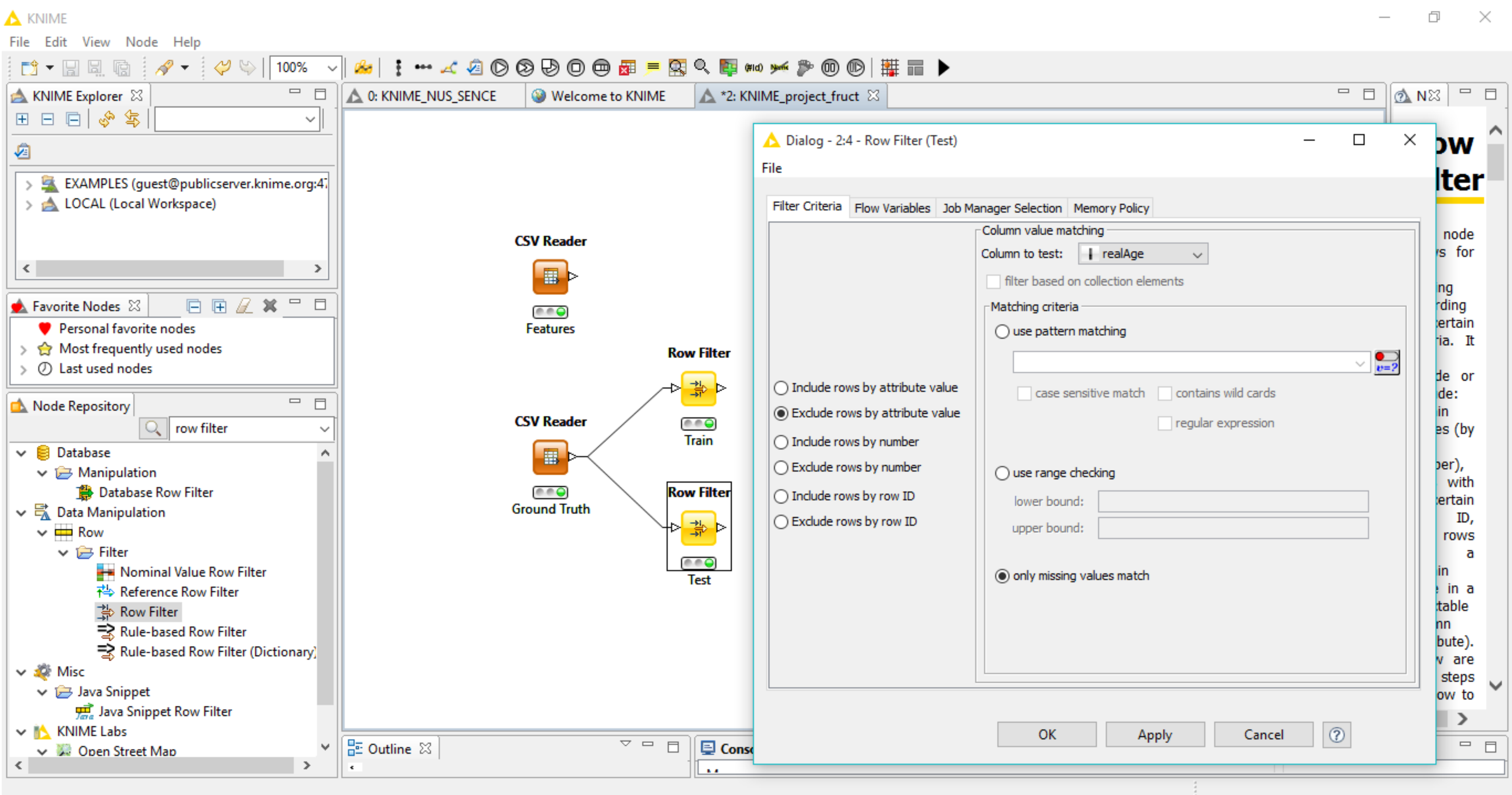
Set up CSV readers to read features file and ground truth file, execute the workflow.



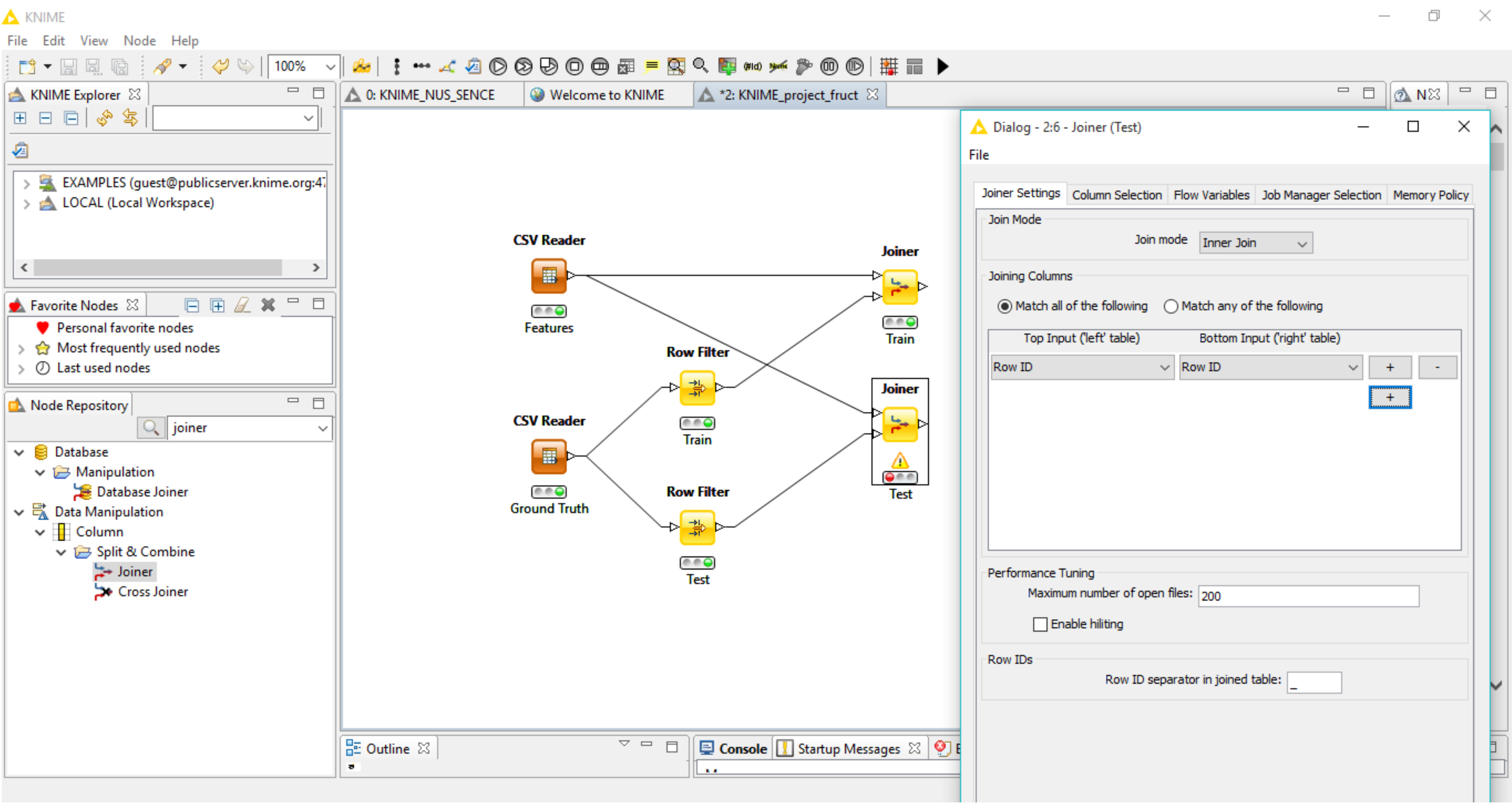
Set up CSV readers to read features file and ground truth file, execute the workflow.



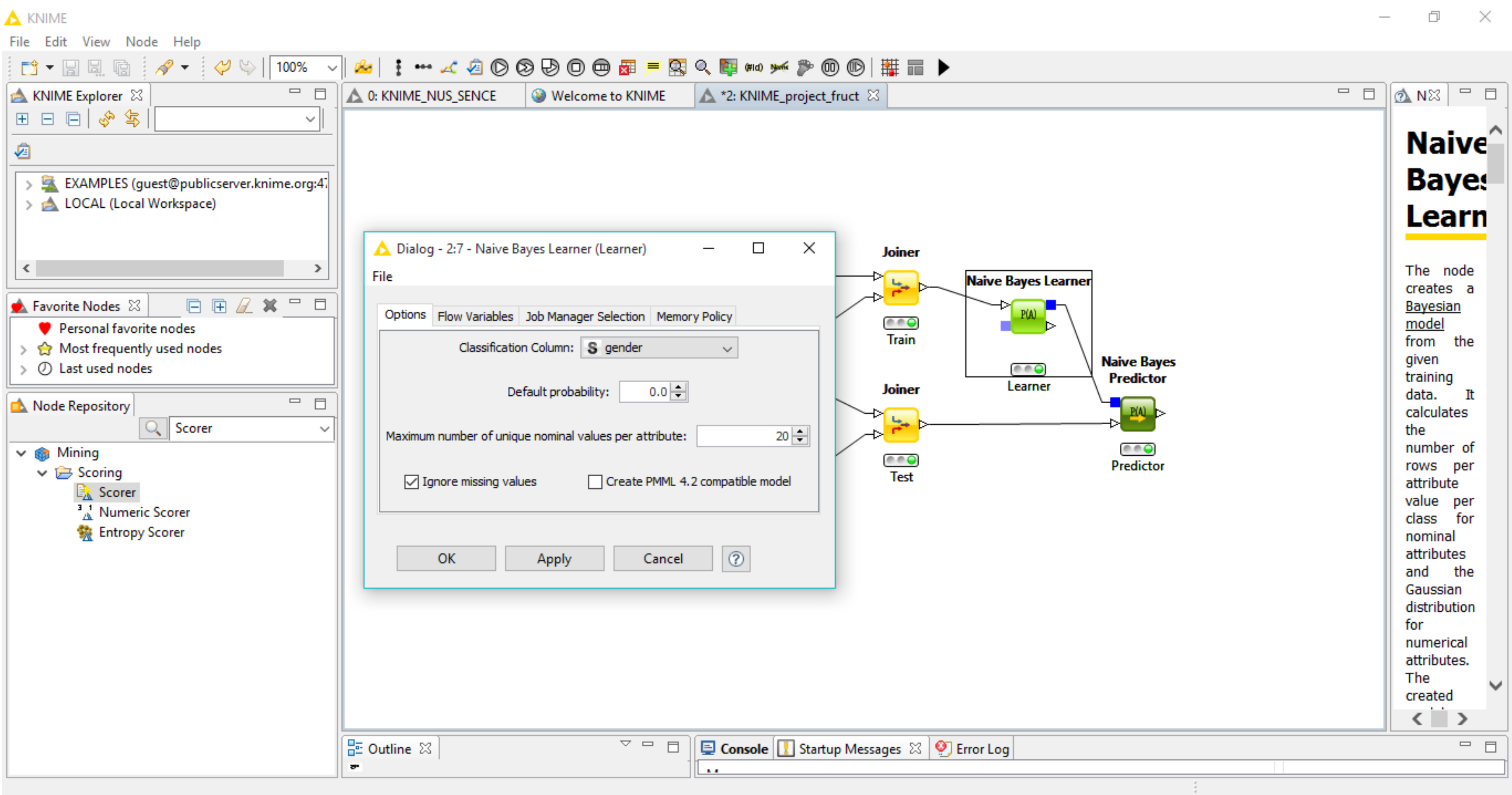
Add two “Row Filter” nodes – to separate users with real age indicated and without by excluding and including missing rows, respectively.



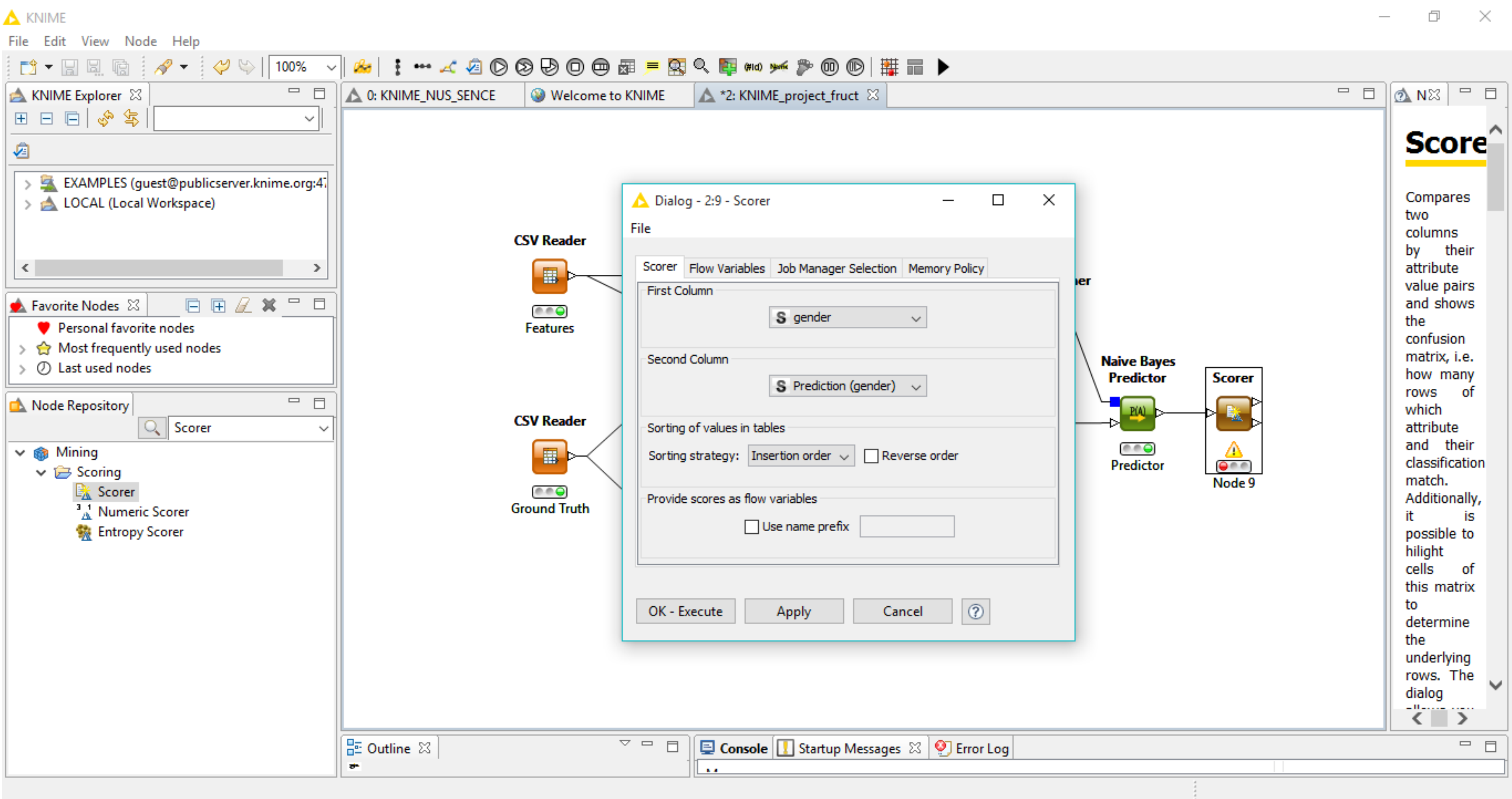
Add two "Joiner" nodes – to join (Inner Join by RowId) ground truth and features in one table for train and one table for test set.



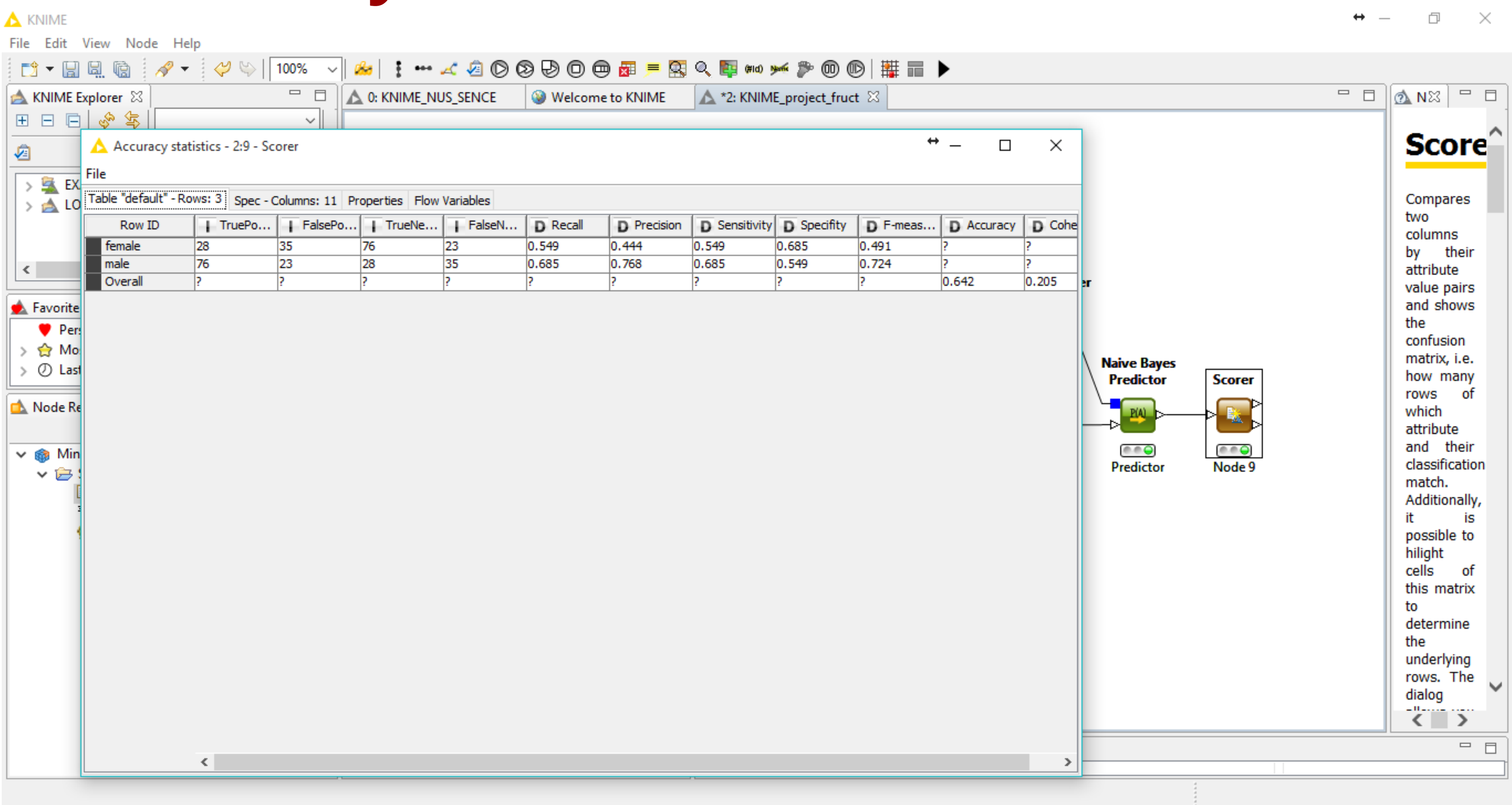
And i.e. “Naïve Bayes Learner” and “Naïve Bayes Predictor” nodes to train and test data flows, respectively. Set up learner to train based on i.e. “gender”.



And “Scorer” node to the output of “Predictor” and set it up to compare predicted results and ground truth. Execute the workflow.



That’s All! The evaluation metrics are computed in “Scorer” node and can be flushed to file (“CSV Writer”) or to UI. Try it different features.



Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Cohe
female	28	35	76	23	0.549	0.444	0.549	0.685	0.491	?	?
male	76	23	28	35	0.685	0.768	0.685	0.549	0.724	?	?
Overall	?	?	?	?	?	?	?	?	?	0.642	0.205

Summary

- Knime is easy to use but you must understand the principles of each node you used.
- Knime is not capable to solve custom tasks easily, but very helpful to test assumptions or run baselines.
- Sometimes it is useful to implement a model from scratch. It may help to understand results better, so we encourage it.
- You have two days to implement your assignment and prepare presentations. You can use whatever software (language) you like. Just make it work on time and present to us.