Профилирование атрибутов пользователей

из множества источников данных различной модальности

Aleksandr Farseev

E-mail: farseev@u.nus.edu

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

Harvesting Multiple Sources for User Profile Learning: a Big Data Study

Aleksandr Farseev,
Liqiang Nie,
Mohammad Akbari,
and Tat-Seng Chua

@ ICMR 2015

National University of Singapore

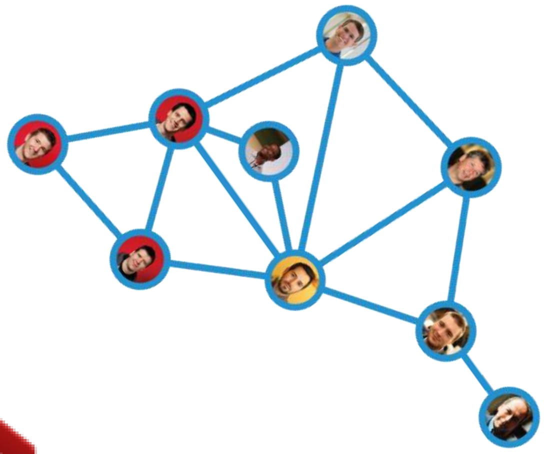
Lab for Media Search

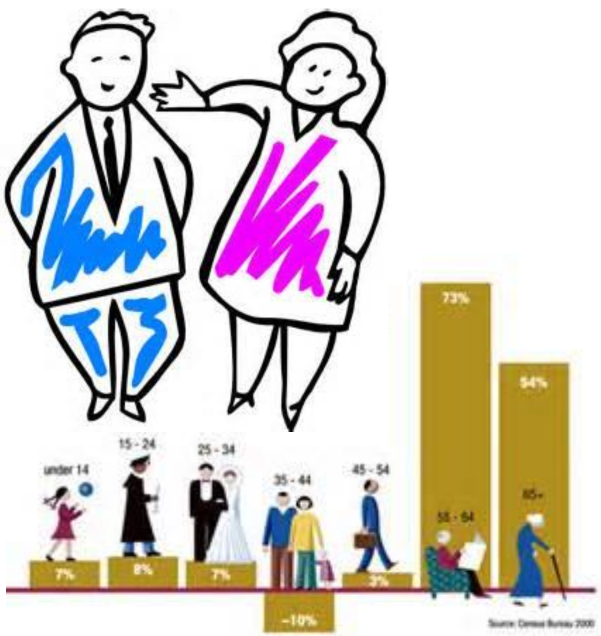
References

- 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),* GB, Oxford, June 28 – July 1, 2015.

What is user profile?





What is human mobility?

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

What is human mobility?

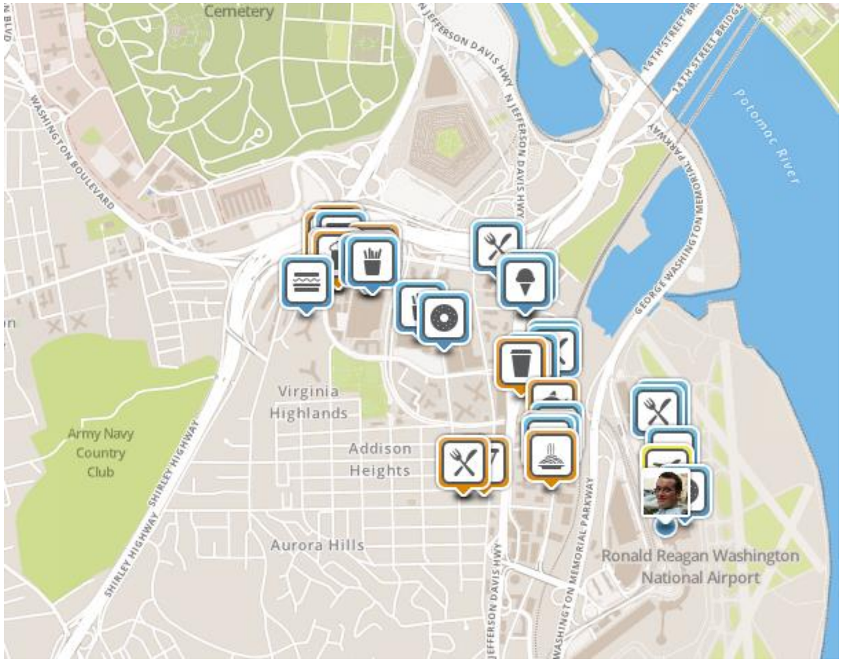
- Mobility - contemporary paradigm, which explores various types of people movement.

- 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

THE FREE DICTIONARY BY FARLEX

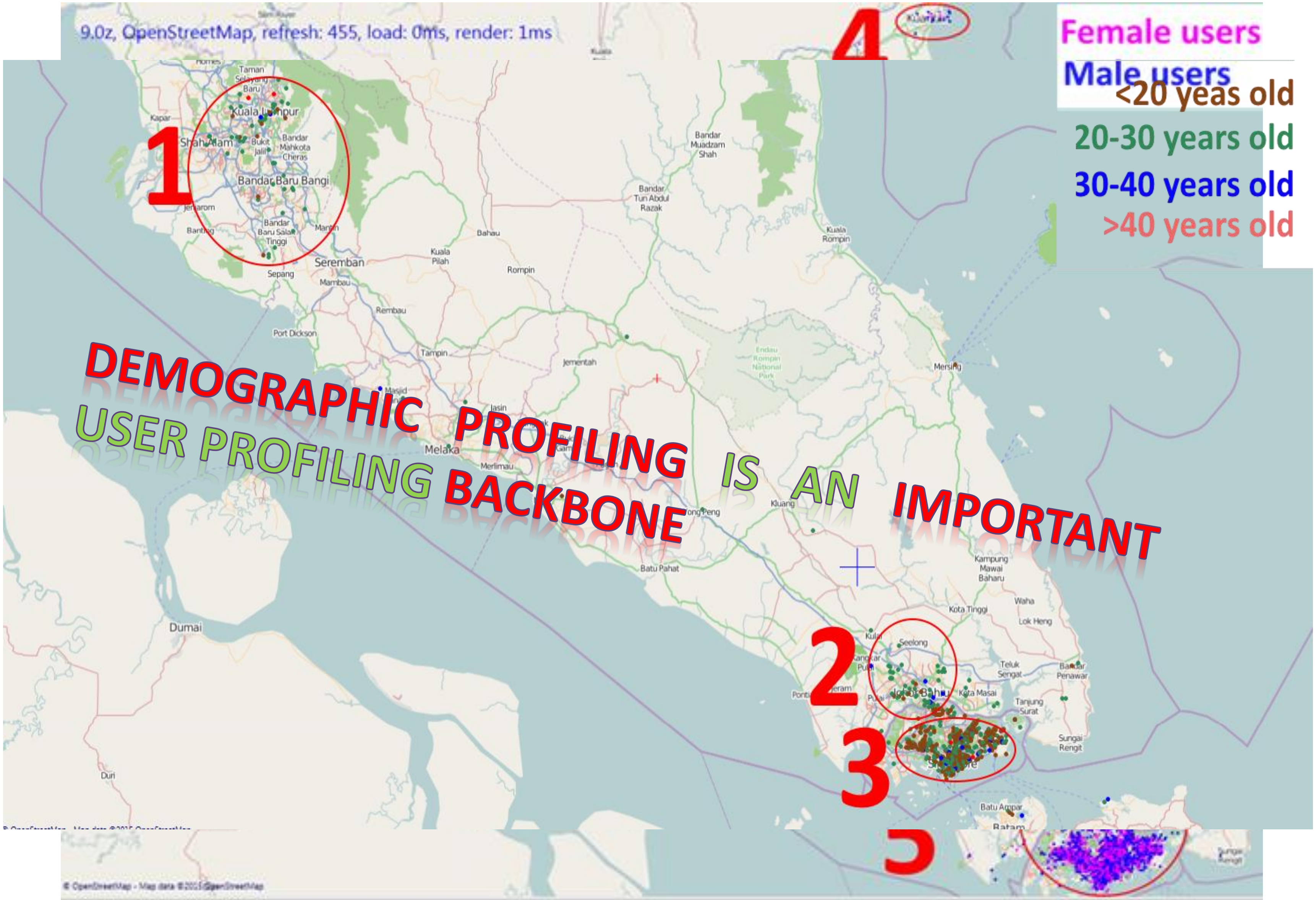
Why human mobility?

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



Mobility can describe people?

If we want to know more?



DEMOGRAPHIC PROFILING IS AN IMPORTANT

USER PROFILING BACKBONE

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

User profile

Mobility profile

- Location preference
- Movement patterns

Demographic profile

- Age
- Gender
- Personality
- Occupation

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://nusmulsitource.azurewebsites.net

NUS-MSS: Data sources

FOURSQUARE

BIGGEST LBSN

twitter

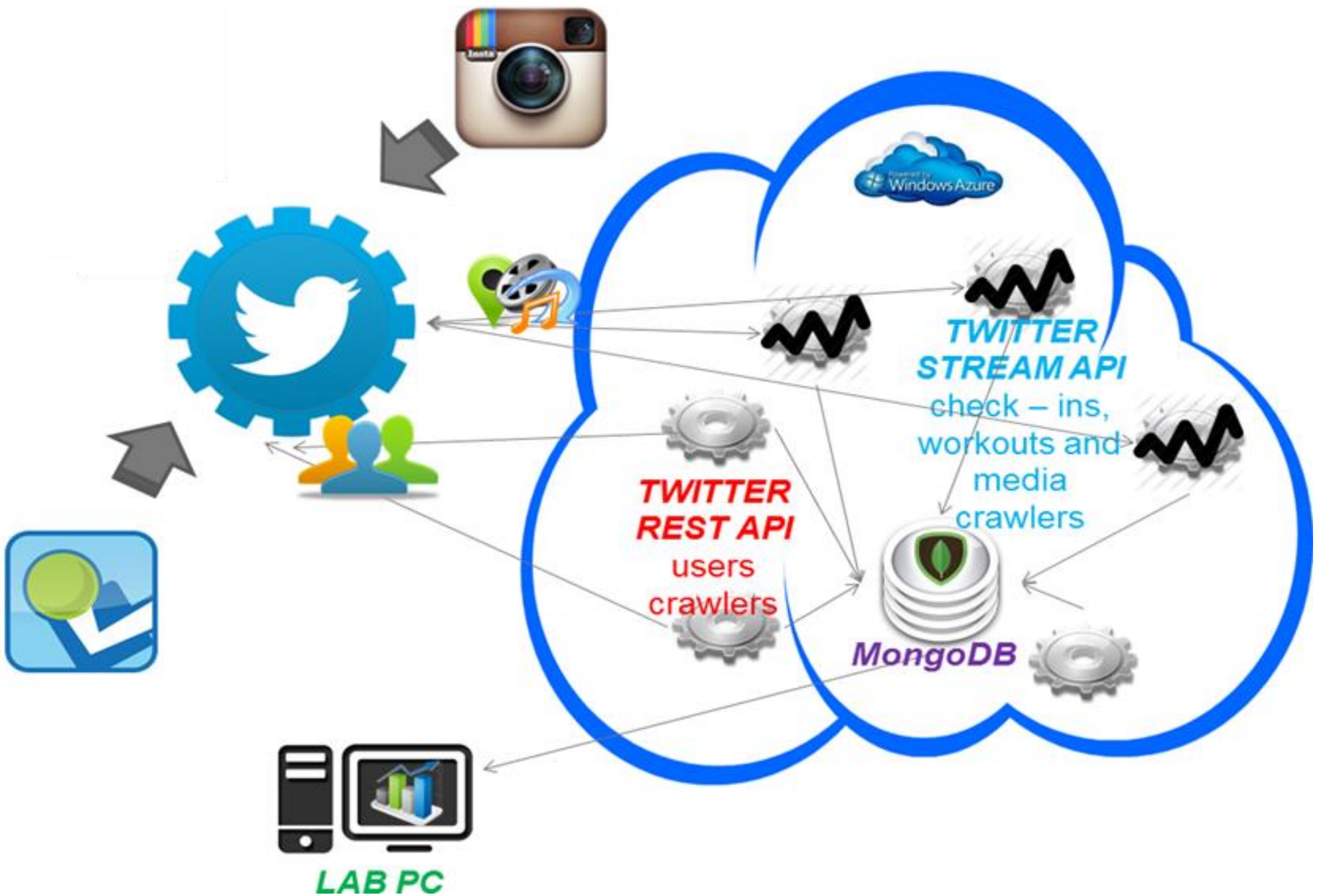
BIGGEST ENGLISH-SPEAKING MICROBLOG

Insta

Instagram

BIGGEST PHOTO SHARING SERVICE

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

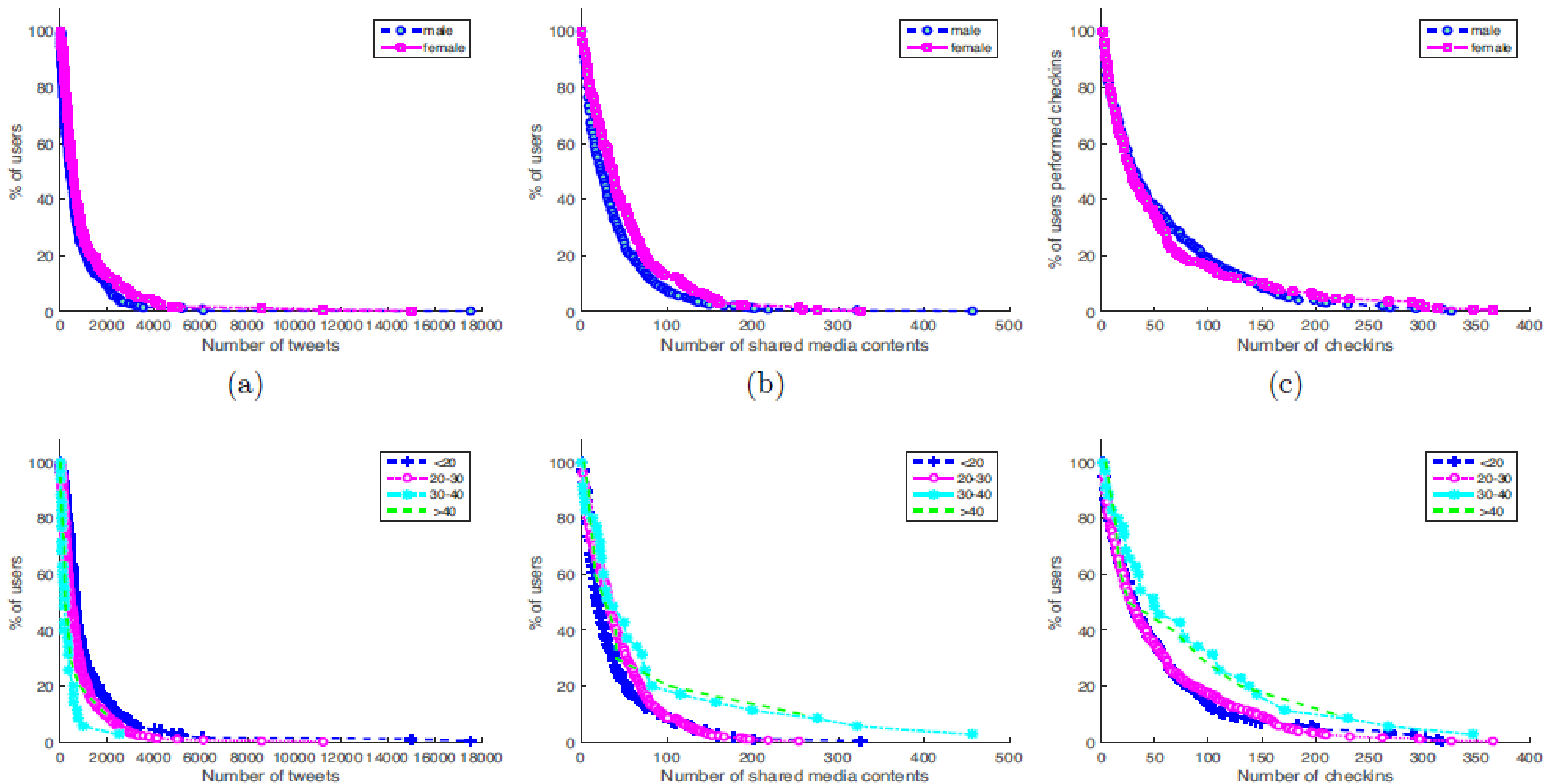
5,263,630 TWEETS

304,493 CHECK-INS

230,752 IMAGES

FROM 7,957 USERS

NUS-MSS: Dataset Statistics in Singapore



(a) (b) (c)

(d) (e) (f)

Demographic profiling

User profile: Mobility + Demography

User profile

Mobility profile

Demographic profile

Location preference

Movement patterns

Age

Gender

Personality

Occupation

Data representation

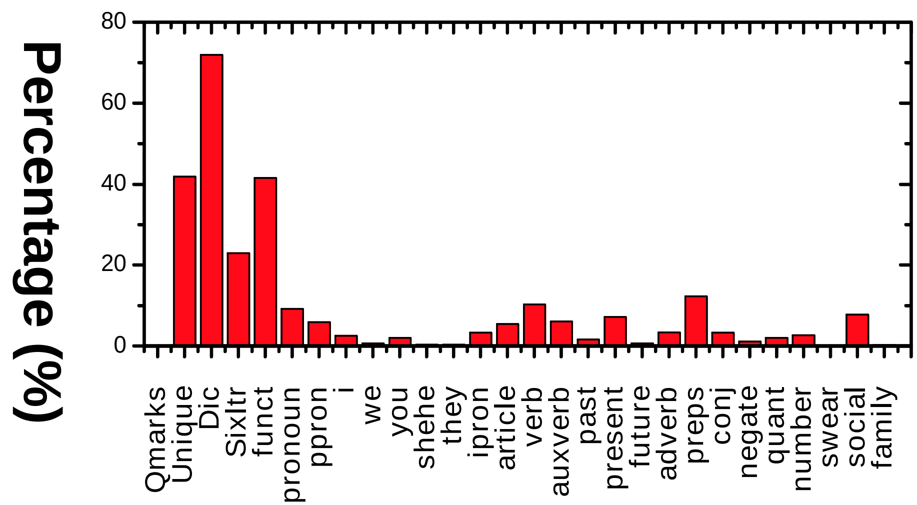
- Linguistic features
  - **LIWC**
  - User Topics
- Heuristic features
  - Writing behavior

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.

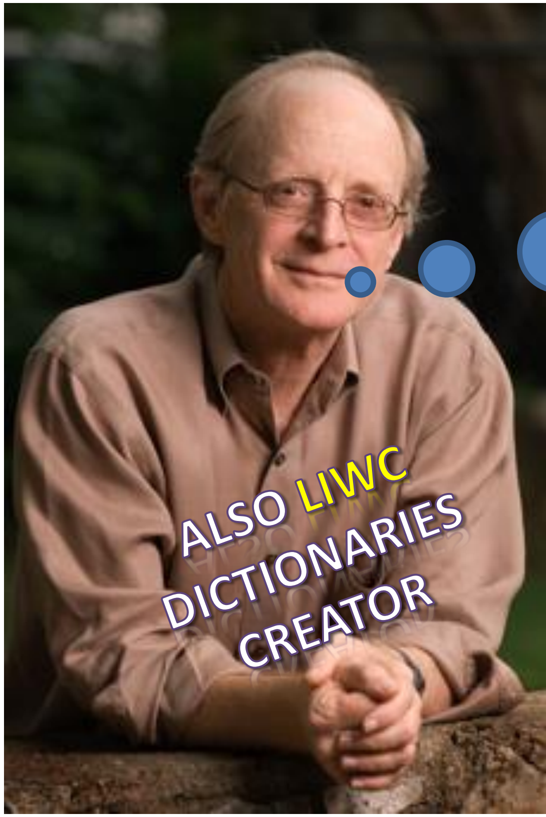
A text analysis software.

Dictionary

Word category



Words usage study for personality profiling



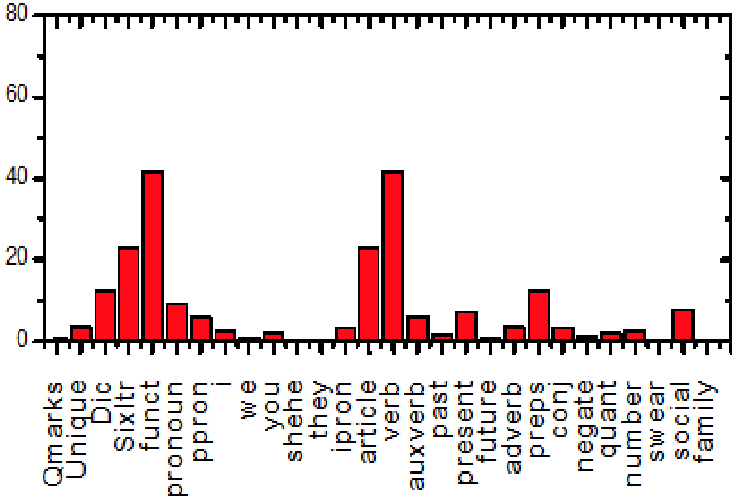
James W. Pennebaker

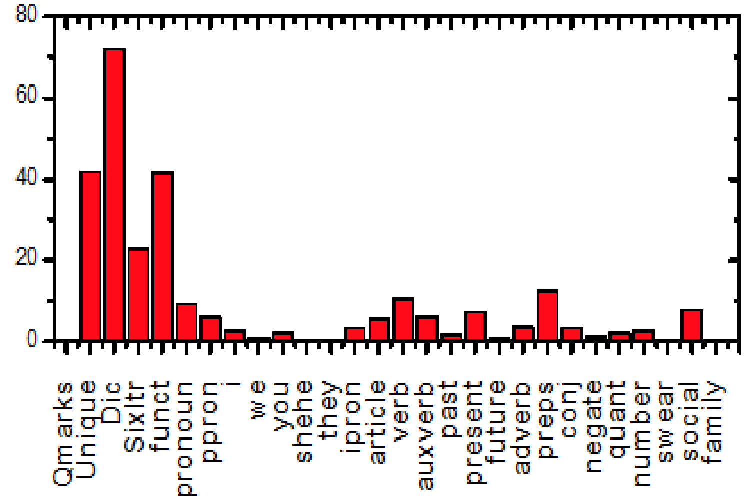
The smallest, most commonly used, most forgettable words serve as windows into our thoughts, emotions, and behaviors.

- Task – Word usage analysis* and correlation with personality
- Data – Various essays and questionnaires
- Approach – manual personality-related dictionaries construction
- Findings:
  - *Certain word usage statistics are good indicators for human personality profiling*

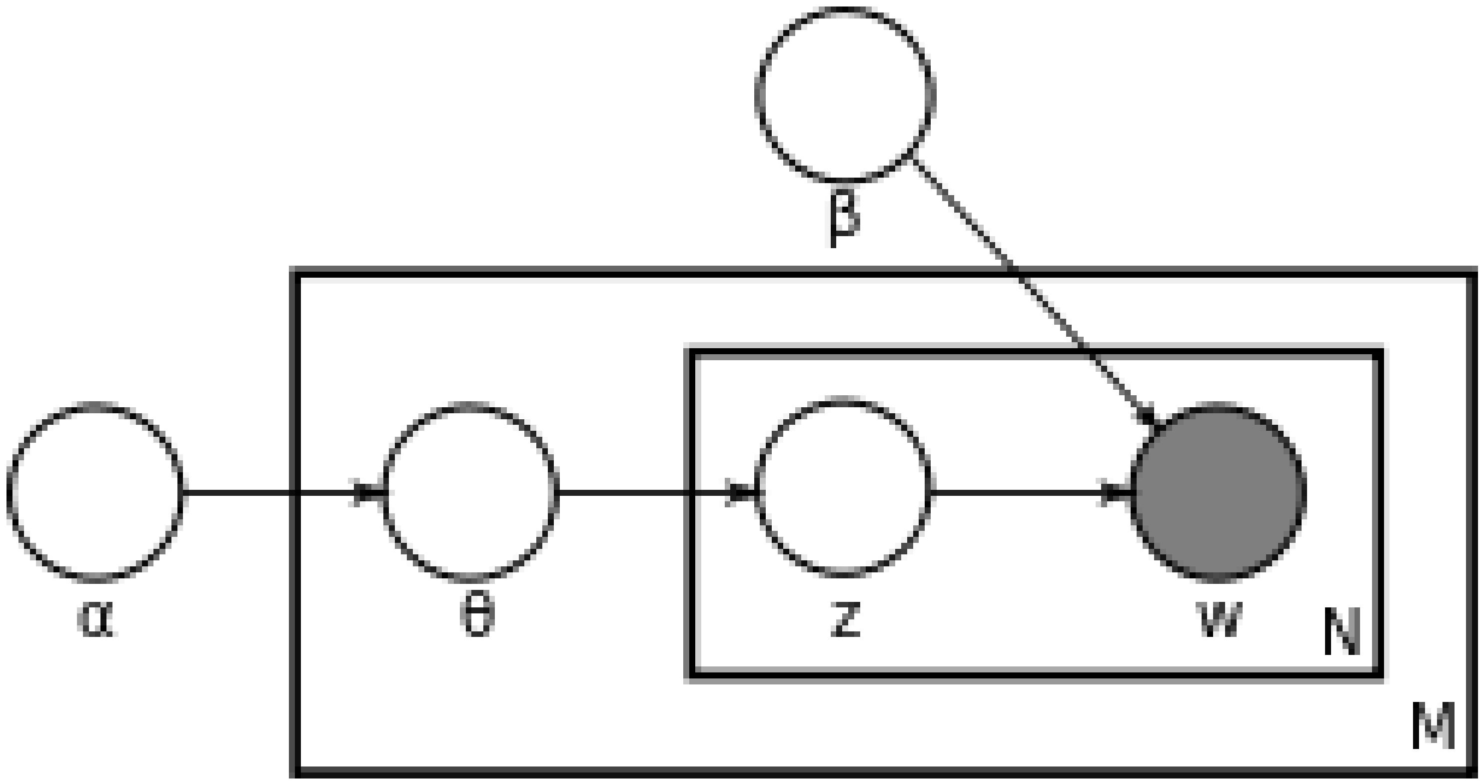
* Pennebaker, J. W. (2011). The secret life of pronouns.

LIWC





- Count occurrences of each LIWC category
- Each document D for user u is represented as a distribution among 74 LIWC categories: $D_u = \{\frac{LIWC_{1u}}{N}, \frac{LIWC_{2u}}{N}, \frac{LIWC_{3u}}{N}, \dots, \frac{LIWC_{74u}}{N})$

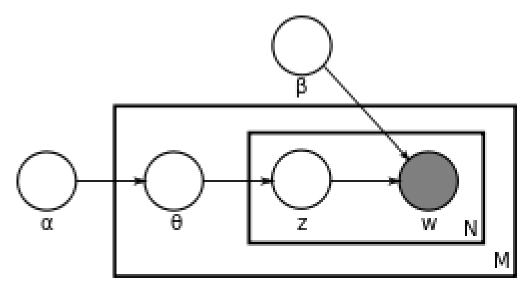

β
α
θ
z
w
N
M

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.



Topic Modeling -1

- Methods for automatically organizing, understanding, searching and summarizing large electronic archives.
  - Uncover hidden topical patterns in collections.
  - Annotate documents according to topics.
  - Using annotations to organize, summarize and search.
  - Widely poplar approach: Latent Dirichlet Allocation (LDA)*

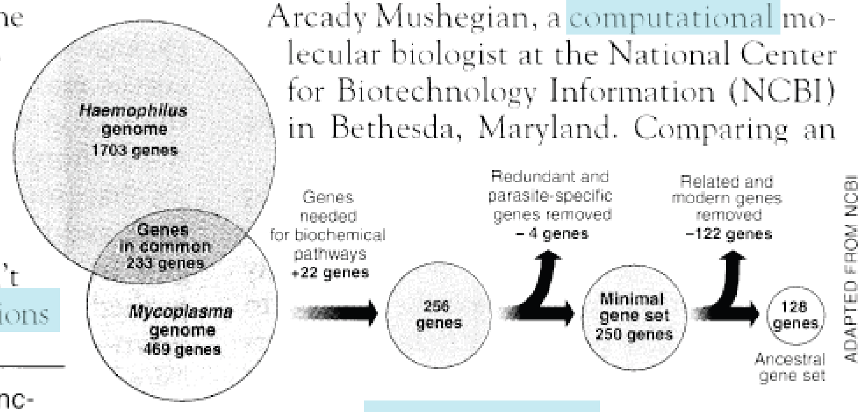
*D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

Topic Modeling -2

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



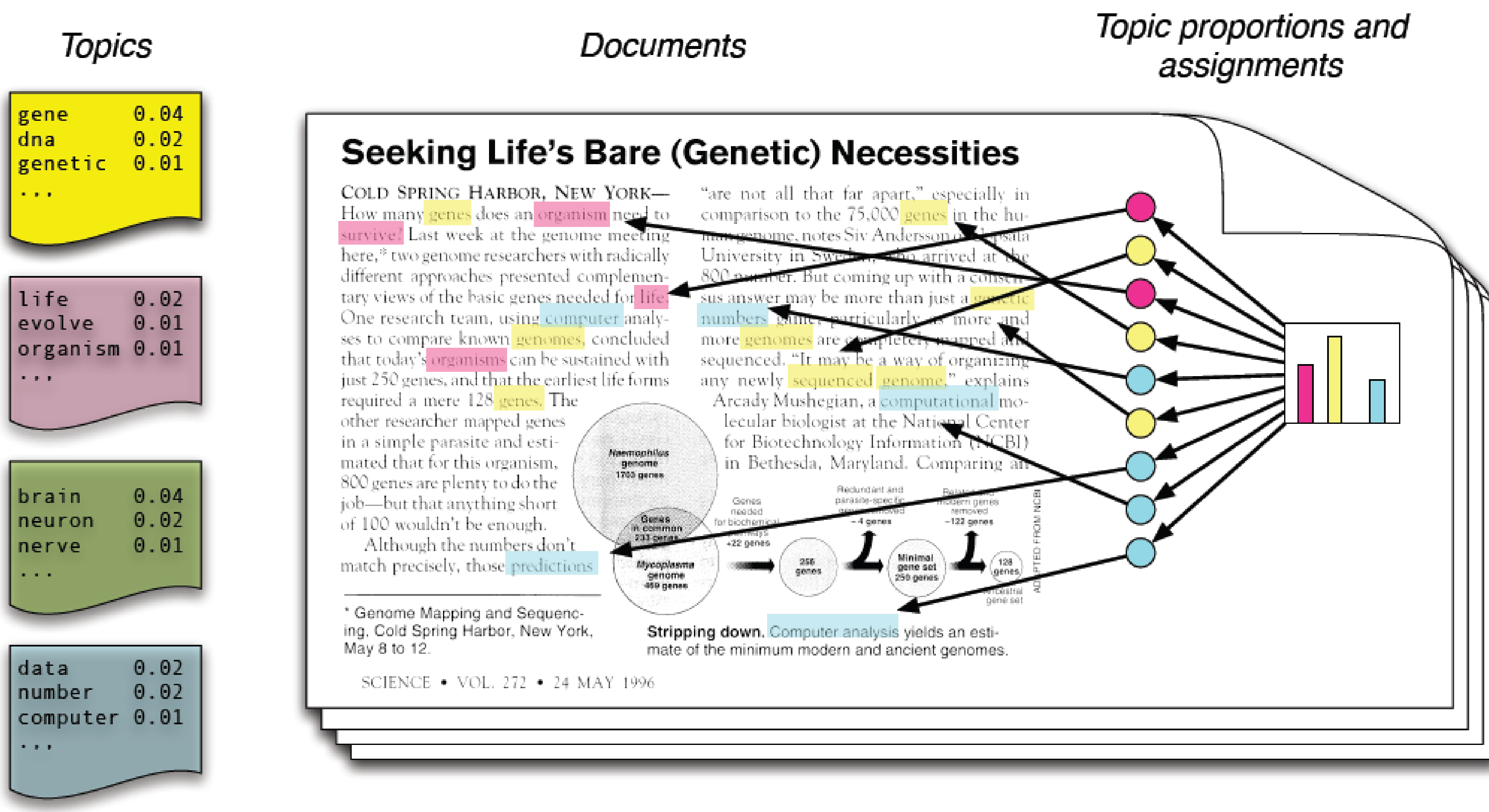
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

*D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

Topic Modeling -3

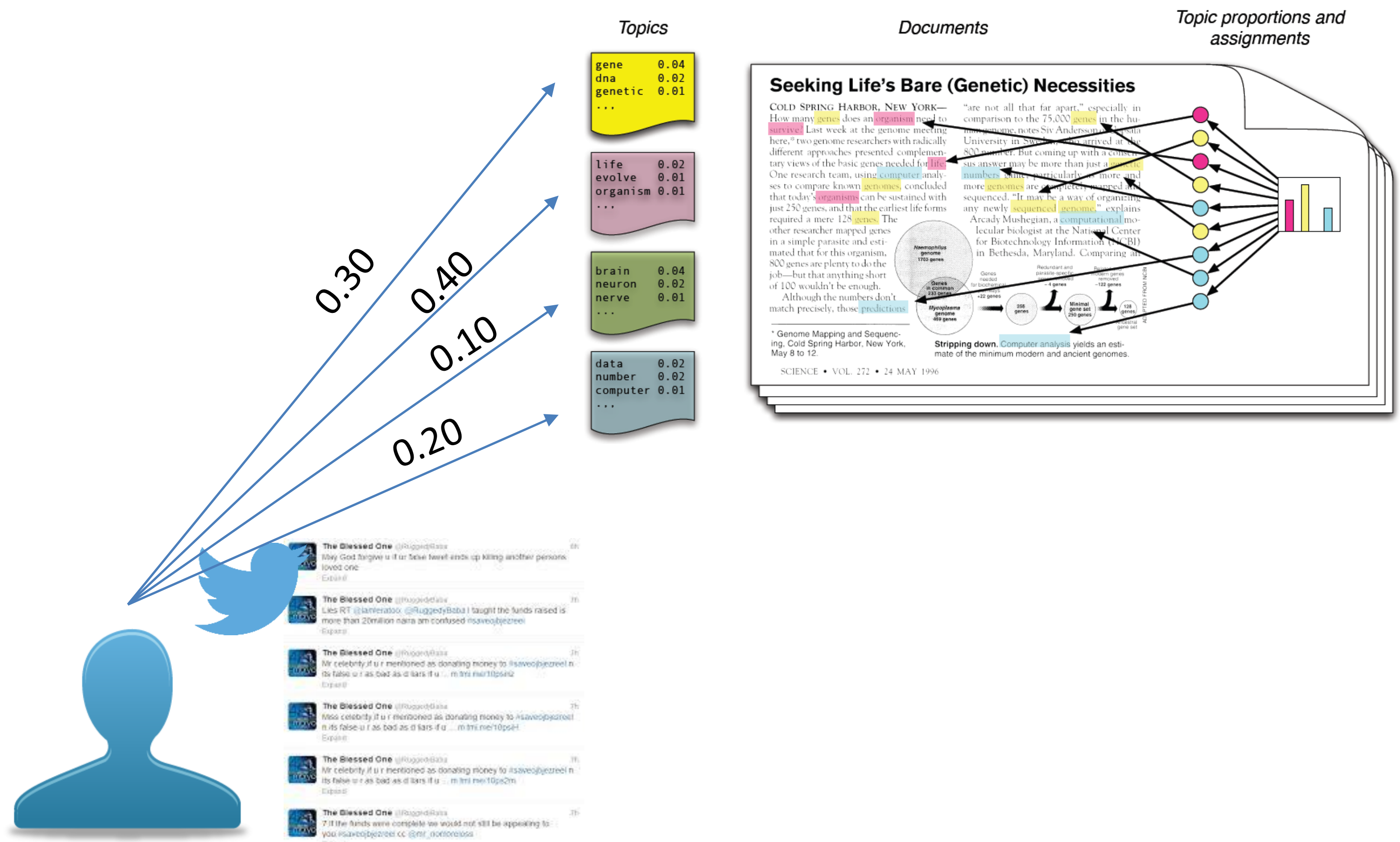


*D. M. Blei, A. Y. Ng, and M. I. Jordan, "[Latent dirichlet allocation](#)," *The Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

Topic Modeling -4

- Only documents are observable ***(All user's tweets are in one document for every user)***.
- Infer underlying topic structure:
  - Topics that generated the documents.
  - For each document, distribution of topics.
  - For each word, which topic generated the word.

Topic Modeling -5



- Each document D (one user - one document) is represented as a distribution among N LDA topics: $D_u = \{ LDA_{1_u}, LDA_{2_u}, LDA_{3_u}, \ldots, LDA_{N_u})$

Data representation

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

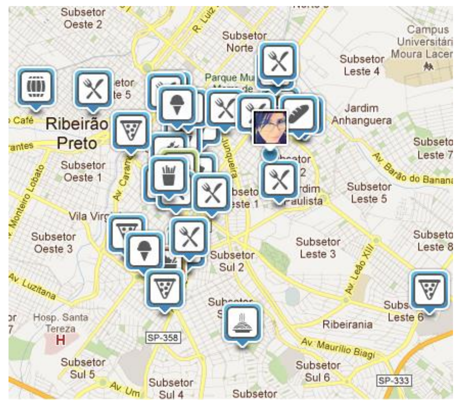
As we mention from our research – user's writing behavioral patterns are highly correlated with e.g. age (individuals from 10 – 20 years old are making two times less grammatical errors than 20 -30 years old individuals)

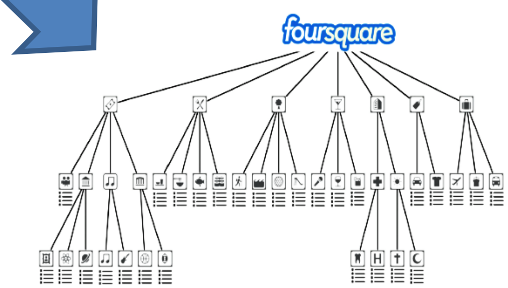
Feature name	Description
Number of hash tags	Number of hash tags mentioned in message
Number of slang words	Number of slang words one use in his tweets. We calculate number of slang words / tweet and compute average slang usage
Number of URLs	Number of URL's one usually use in his/her tweets
Number of 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**

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





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.

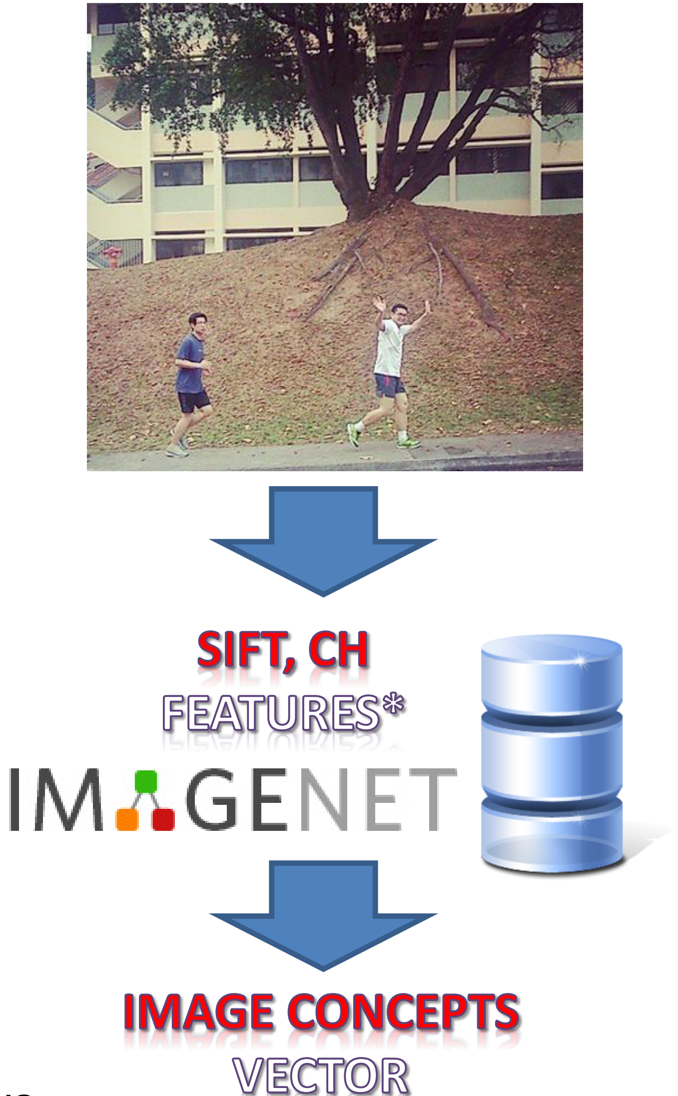
For case when user performed check-ins in two restaurants and airport but did not perform check-ins in other venues:

	cat_1	...	$cat_{rest.}$	...	$cat_{air.}$	...	cat_{517}
u_1	0	0	2	0	1	0	0
...	*	*	*	*	*	*	*
u_N	*	*	*	*	*	*	*

Data representation

- Image features
  - **Image concept learning**

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.



*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

Color Histogram -1

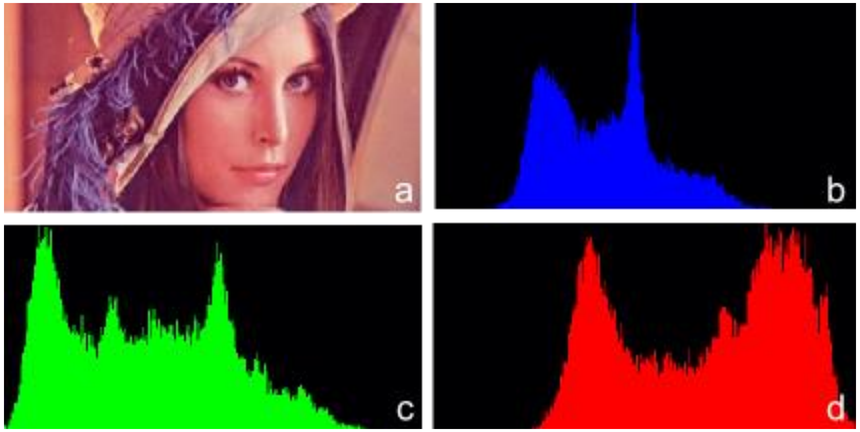
- Let image *I* be of dimension *p x q*
  - For ease in representation, need to quantize *p x q* potential colors into *m* colors (for m << p x q)
  - For pixel p = (x,y), the color of pixel is denoted by *I*(p) = c_k

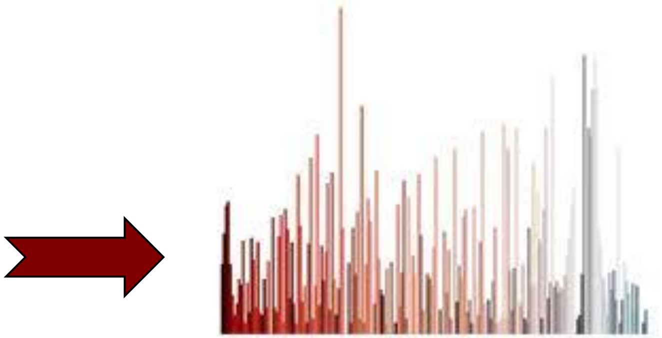
- Construction of Color Histogram
  - Extract color value for each pixel in image
  - Quantize color value into one of *m* quantization levels

  - Collect frequency of color values in each quantization level, where each bin corresponds to a color in the quantized color space

Color Histogram -2

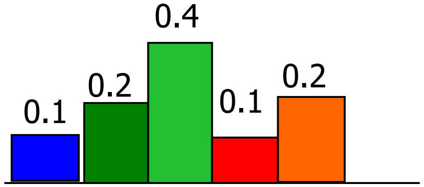
- Thus, image is represented as a color histogram H of size *m*
  - where H[i] gives # of pixels at intensity level I
- For example:





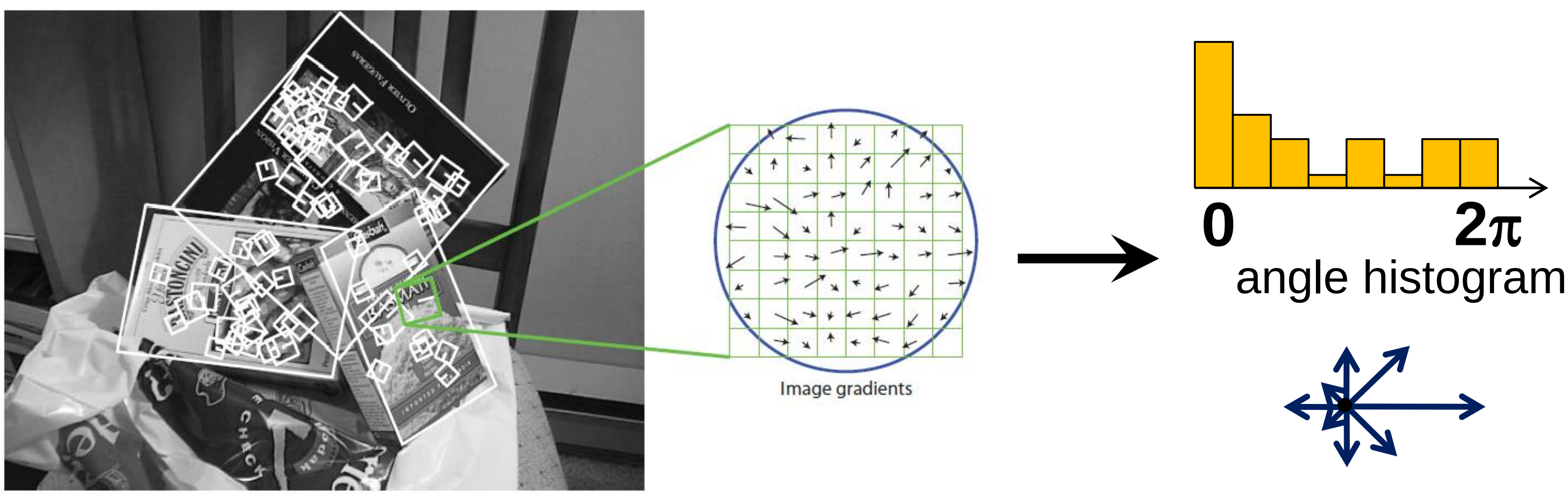
Into a single histogram

- Normalize H to NH by dividing each entry by size of image p*q



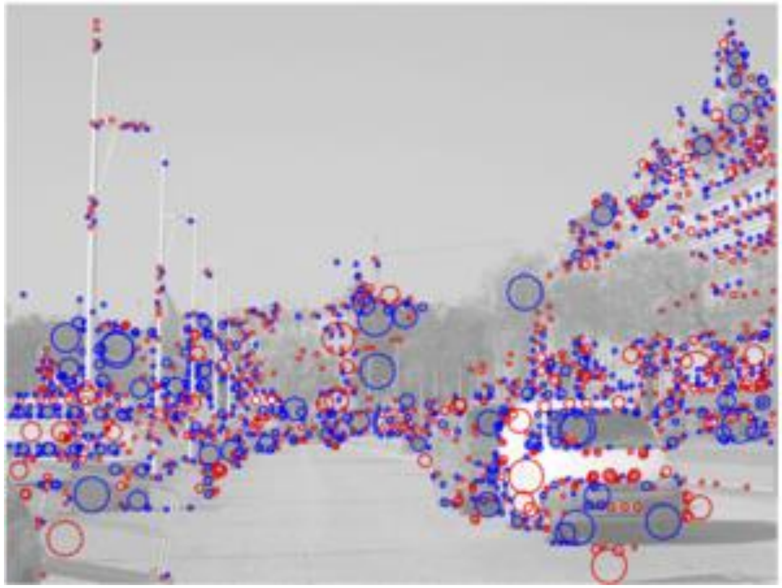
Scale Invariant Feature Transform (SIFT) descriptor -1

- Basic idea: use edge orientation representation
  - Obtain interest points from scale-space extrema of differences-of-Gaussians (DoG)
  - Take 16x16 square window around detected interest point
  - Compute edge orientation for each pixel
  - Throw out weak edges (threshold gradient magnitude)
  - Create histogram of surviving edge orientations



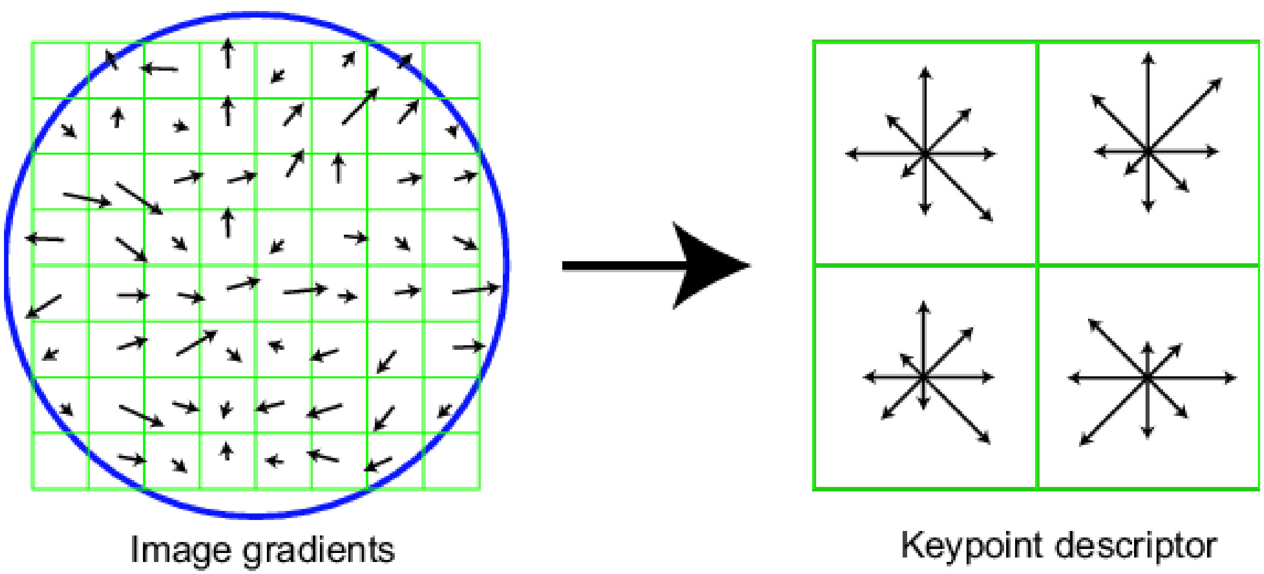
http://www.scholarpedia.org/article/Scale_Invariant_Feature_Transform

Detected Interest Points



Scale Invariant Feature Transform (SIFT) descriptor -2

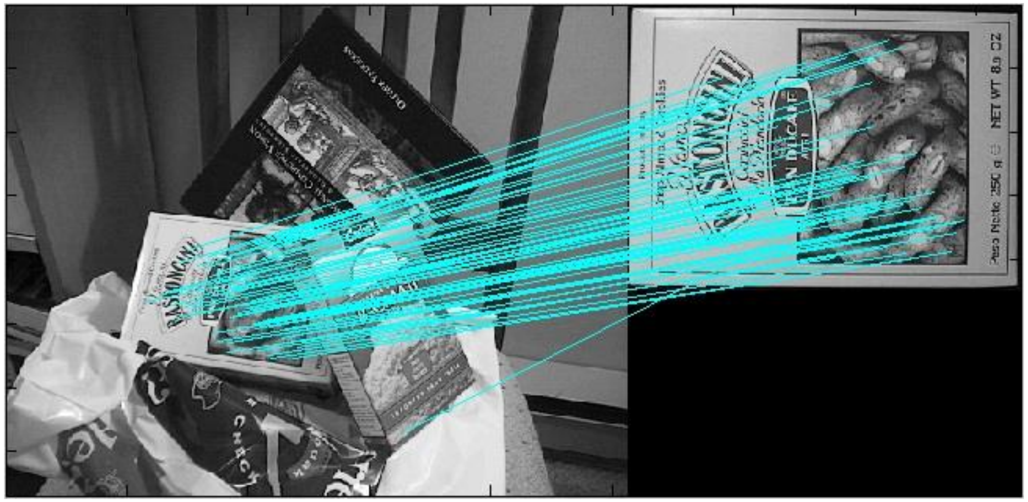
- A popular descriptor:
  - Divide the 16x16 window into a 4x4 grid of cells (we show the 2x2 case below for simplicity)
  - Compute an orientation histogram for each cell
  - 16 cells X 8 orientations = 128 dimensional descriptor



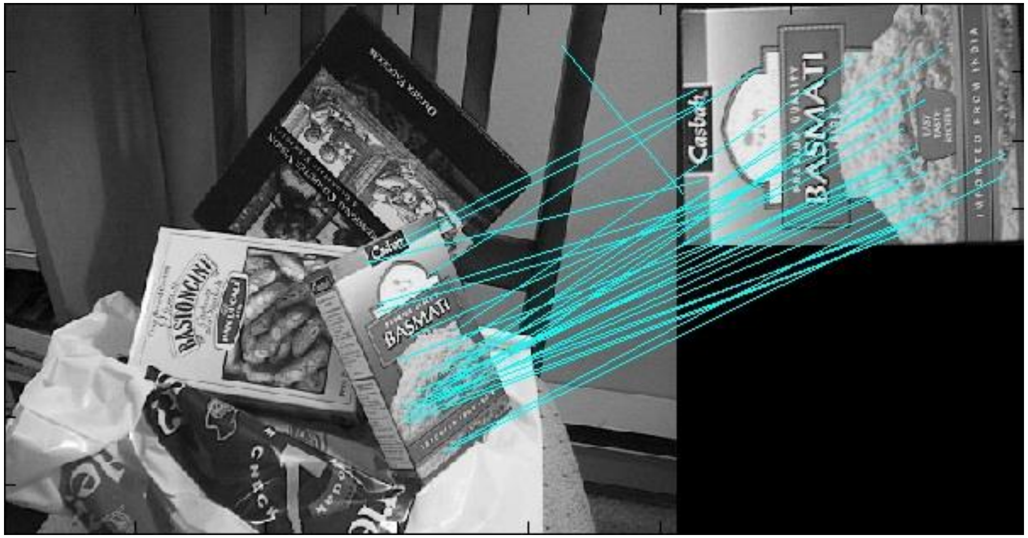
Scale Invariant Feature Transform (SIFT) descriptor -3

- Invariant to
  - Scale
  - Rotation
- Partially invariant to
  - Illumination changes
  - Camera viewpoint
  - Occlusion, clutter

Examples of SIFT matching



80 matches

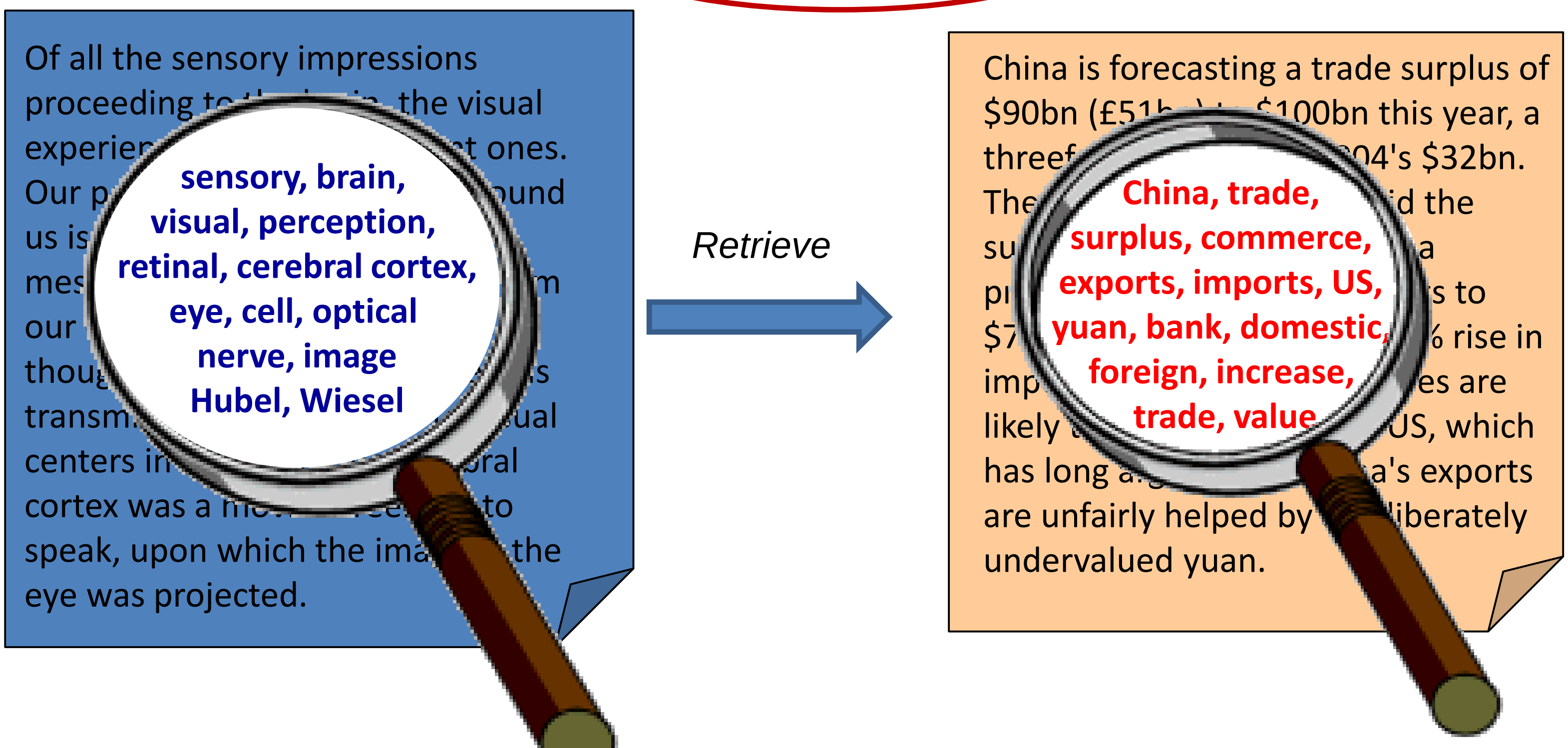


34 matches

Overall Representation: as Bag of Visual Words -1

- Text Words in Information Retrieval (IR)
  - Compactness
  - Descriptiveness

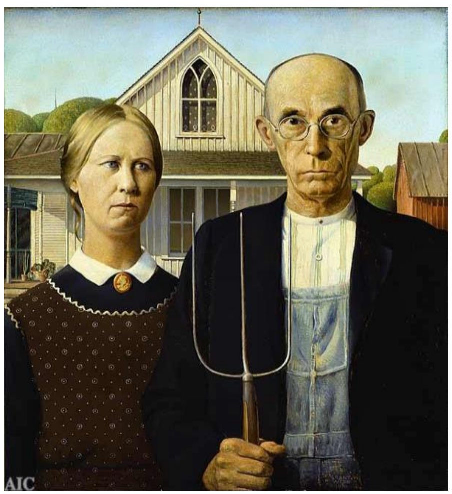
Bag-of-Word model



Overall Representation: as Bag of Visual Words -2

- Can images be represented as Bag-of-Visual Words?

Image → **Bag of ‘visual words’**

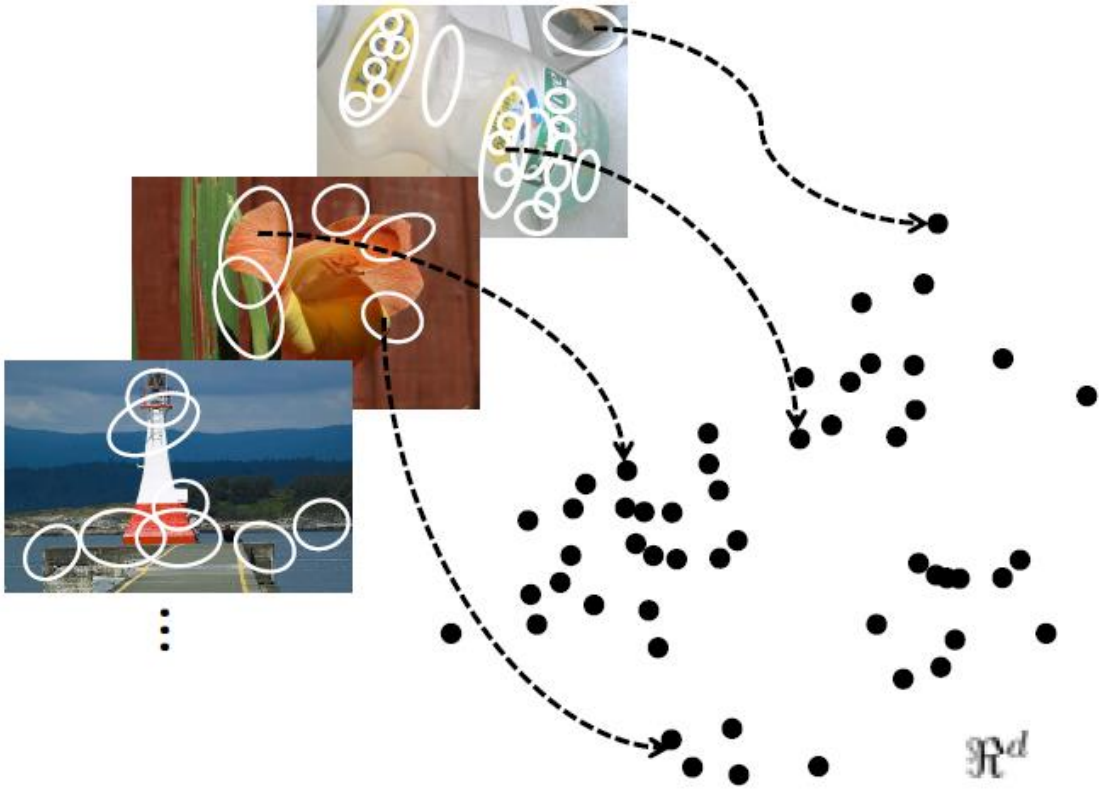


?

- Idea: quantize SIFT descriptors of all training images to extract representative visual words!

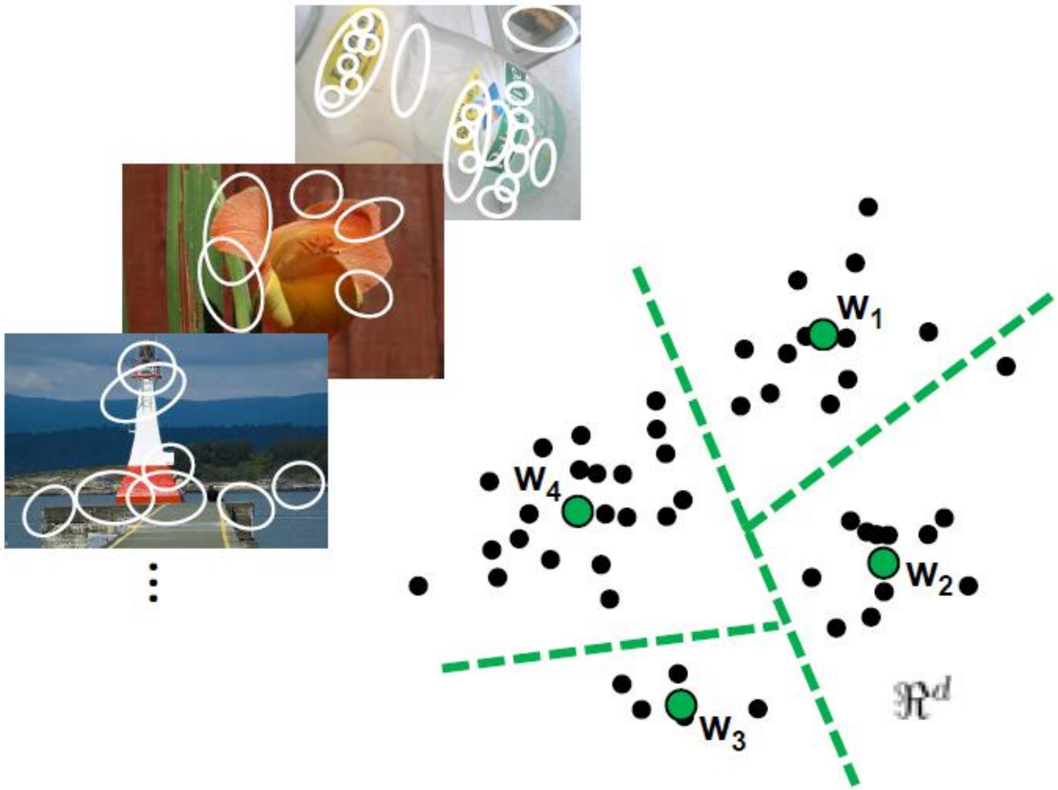
Overall Representation: as Bag of Visual Words -3

Step 1: Extract interest points of all training images

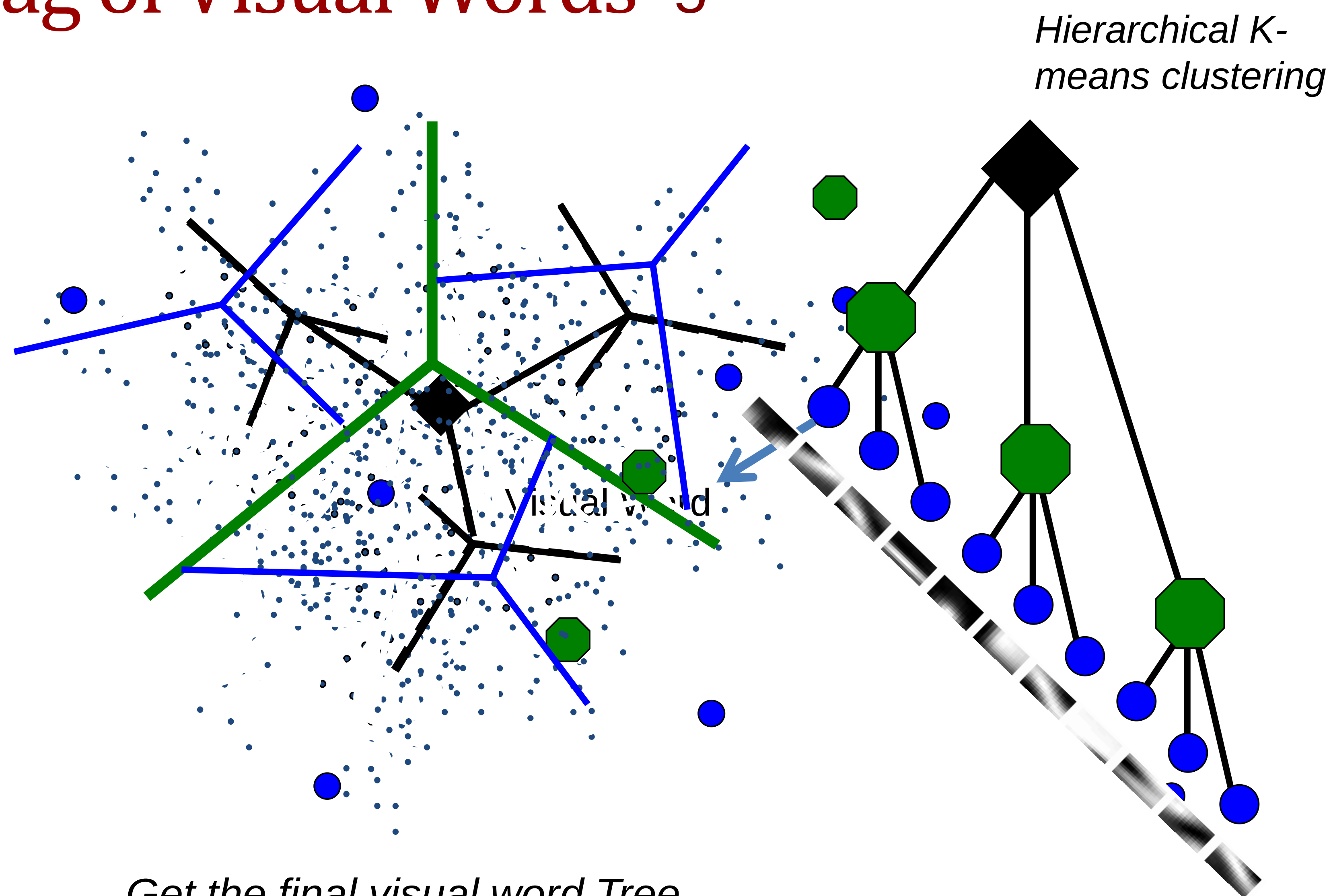


Overall Representation: as Bag of Visual Words -4

Step 2: Features are clustered to quantize the space into a discrete number of visual words.



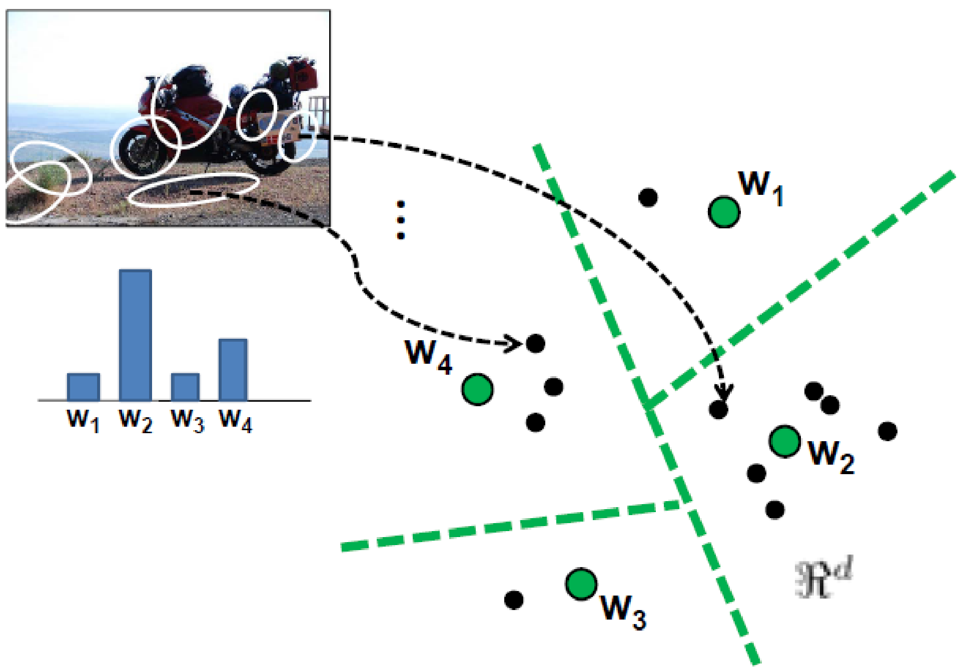
Overall Representation: as Bag of Visual Words -5



Get the final visual word Tree

Overall Representation: as Bag of Visual Words -6

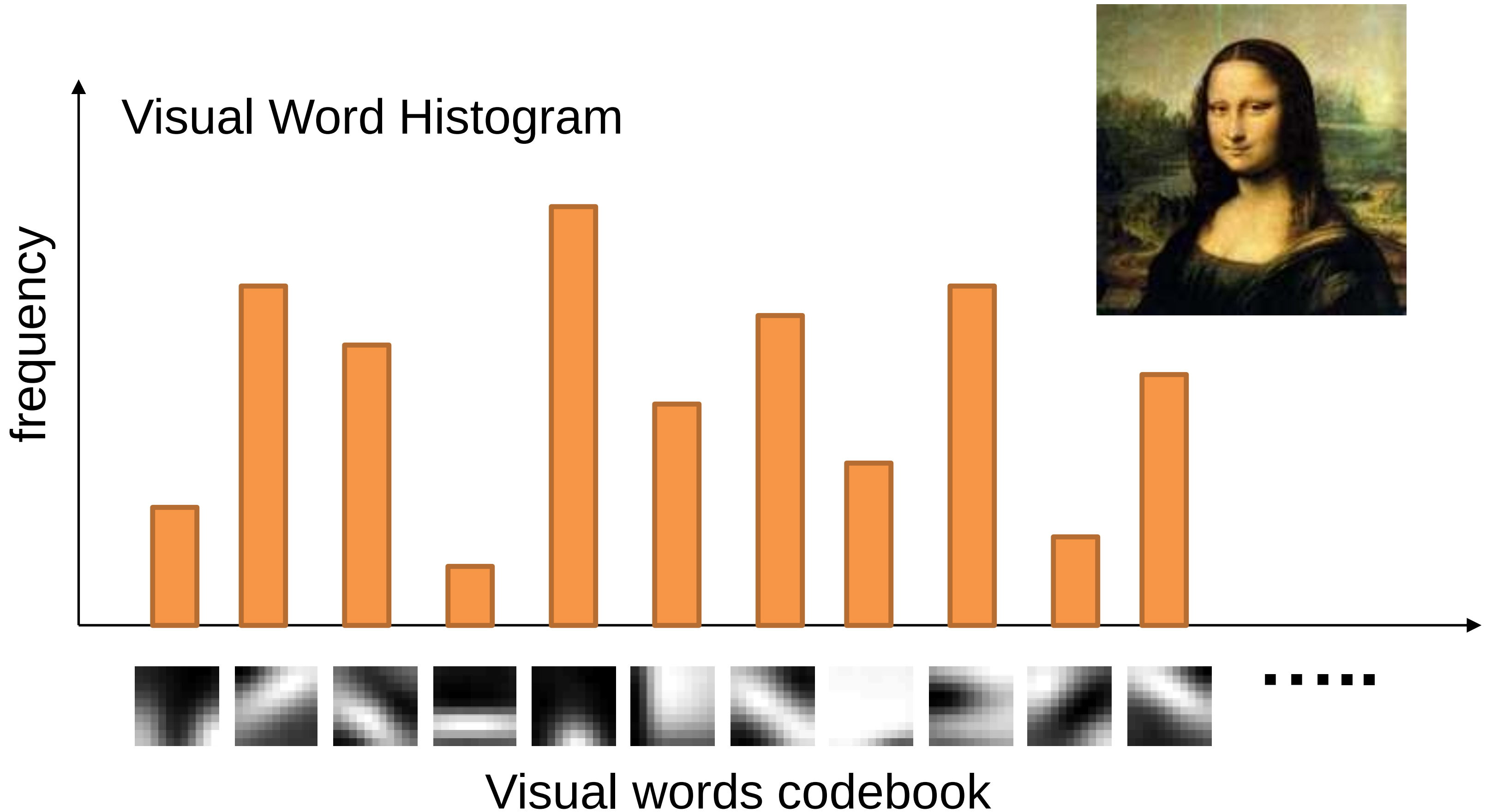
Step 3: Summarize (represent) each image as histogram of visual words



and use as basis for matching and retrieval!

Overall Representation: as Bag of Visual Words -7

- Another example:

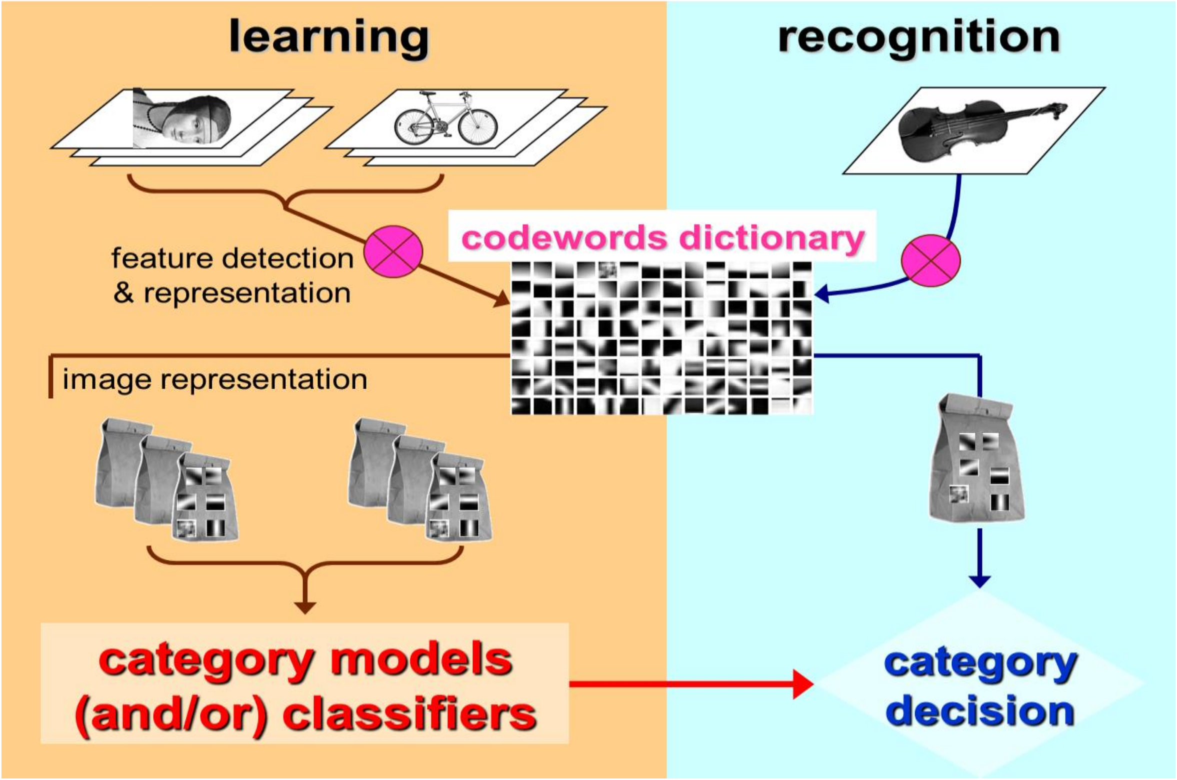


Concept Recognition: Bag-of-Word Model

BASIC IDEA:

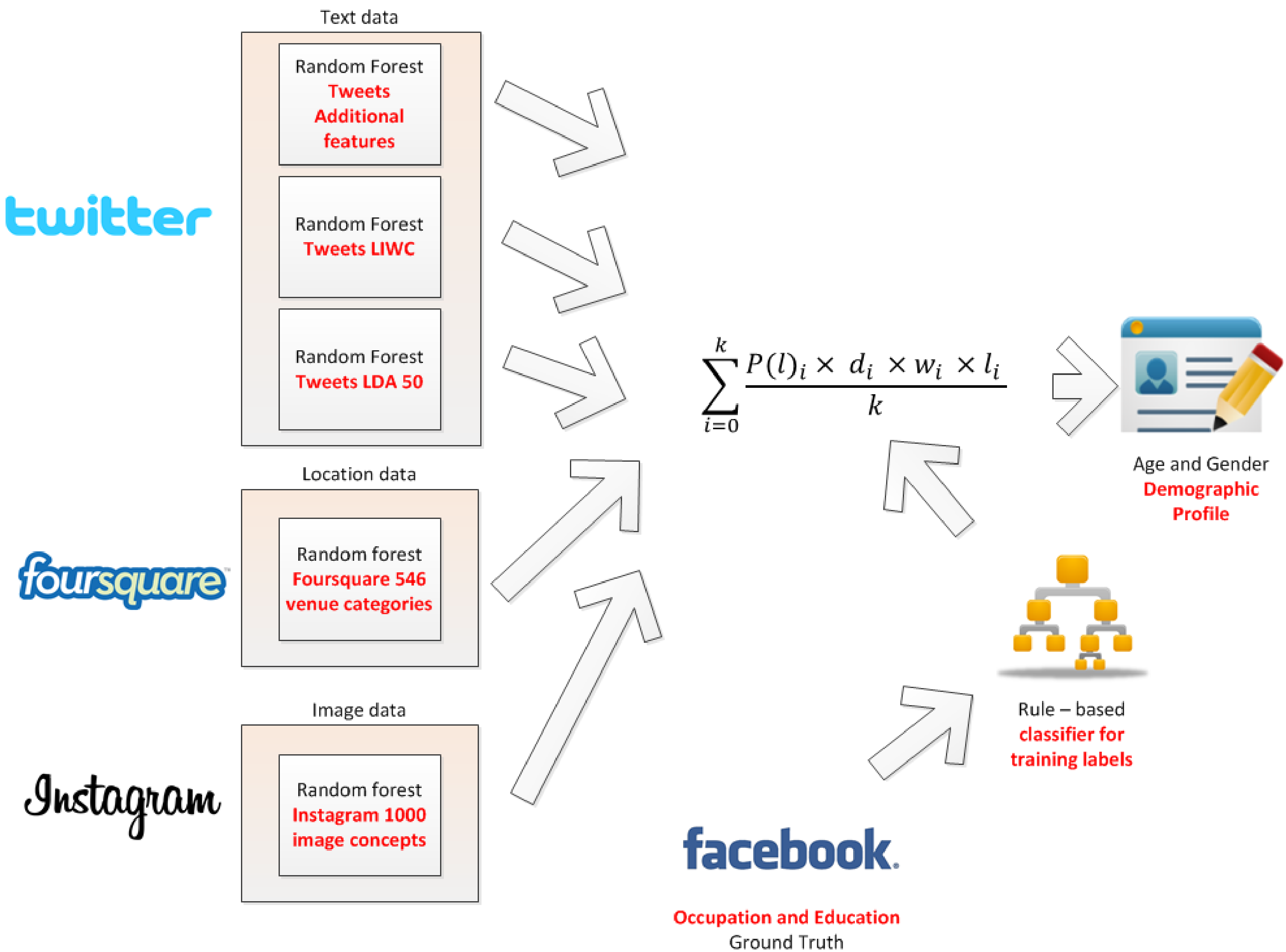
- Representative set of images in each category is collected
- An image is represented by a collection of "visual words"

- Object categories are modeled by the distributions of these visual words

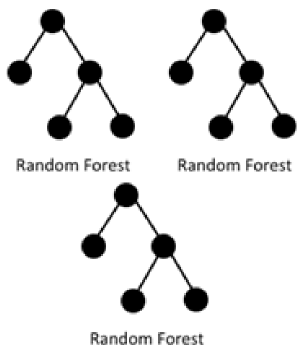



~10,000 to 30,000

Ensemble learning

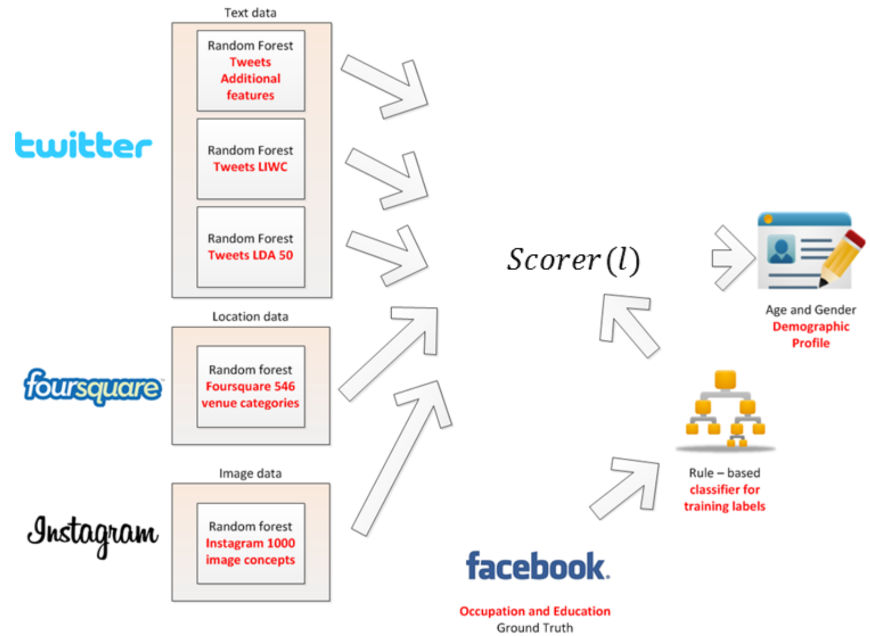


Ensemble learning



$$d_i \times w_i \times l_i$$

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

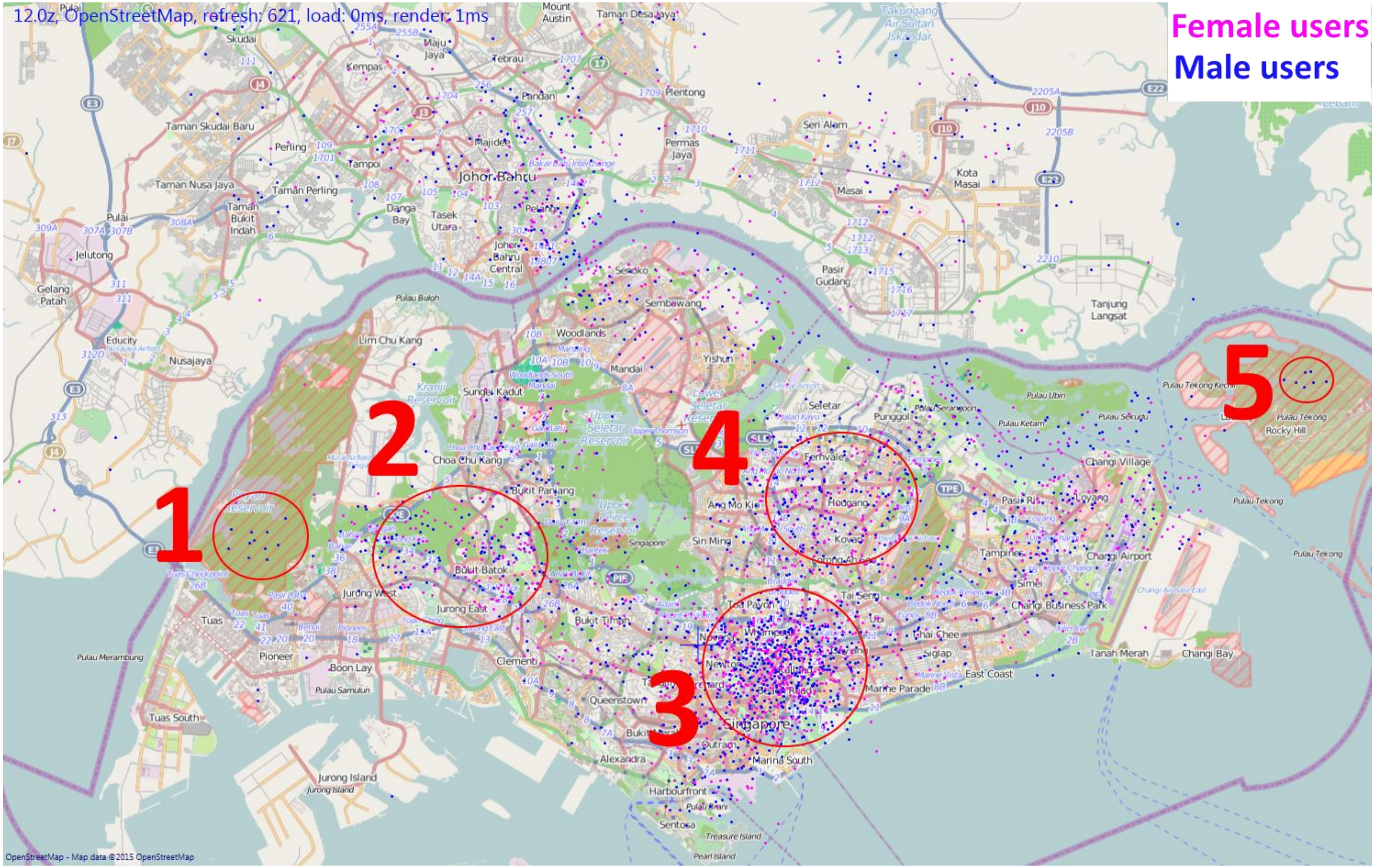
Demographic mobility

User profile: Mobility + Demography

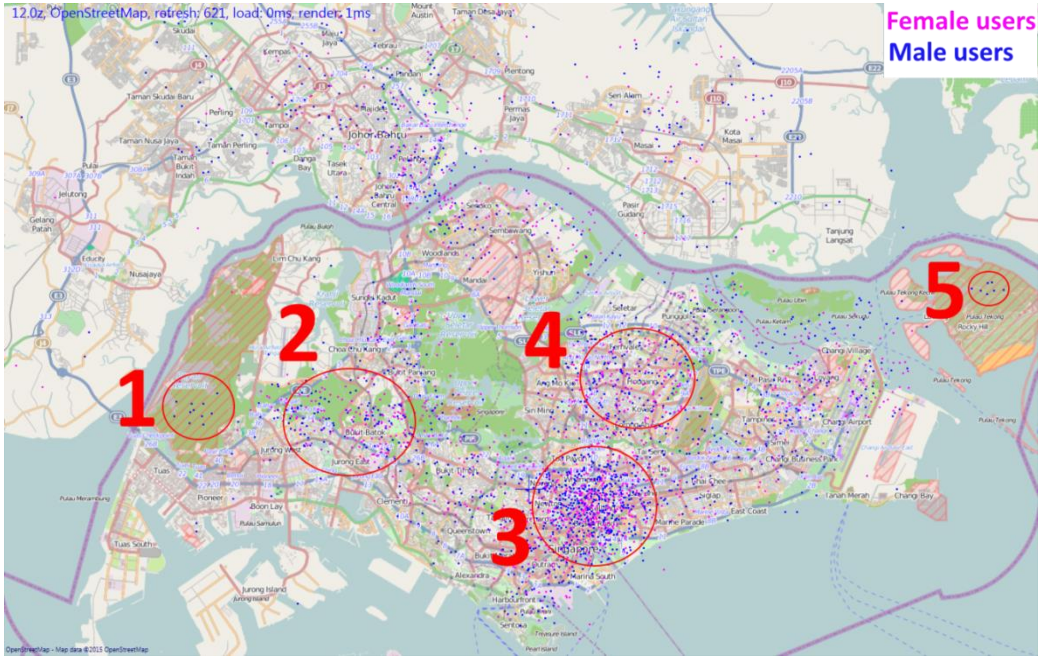
User profile

- Mobility profile
  - Location preference
  - Movement patterns
- Demographic profile
  - Age
  - Gender
  - Personality
  - Occupation

Geographical user mobility: users movement (city level)

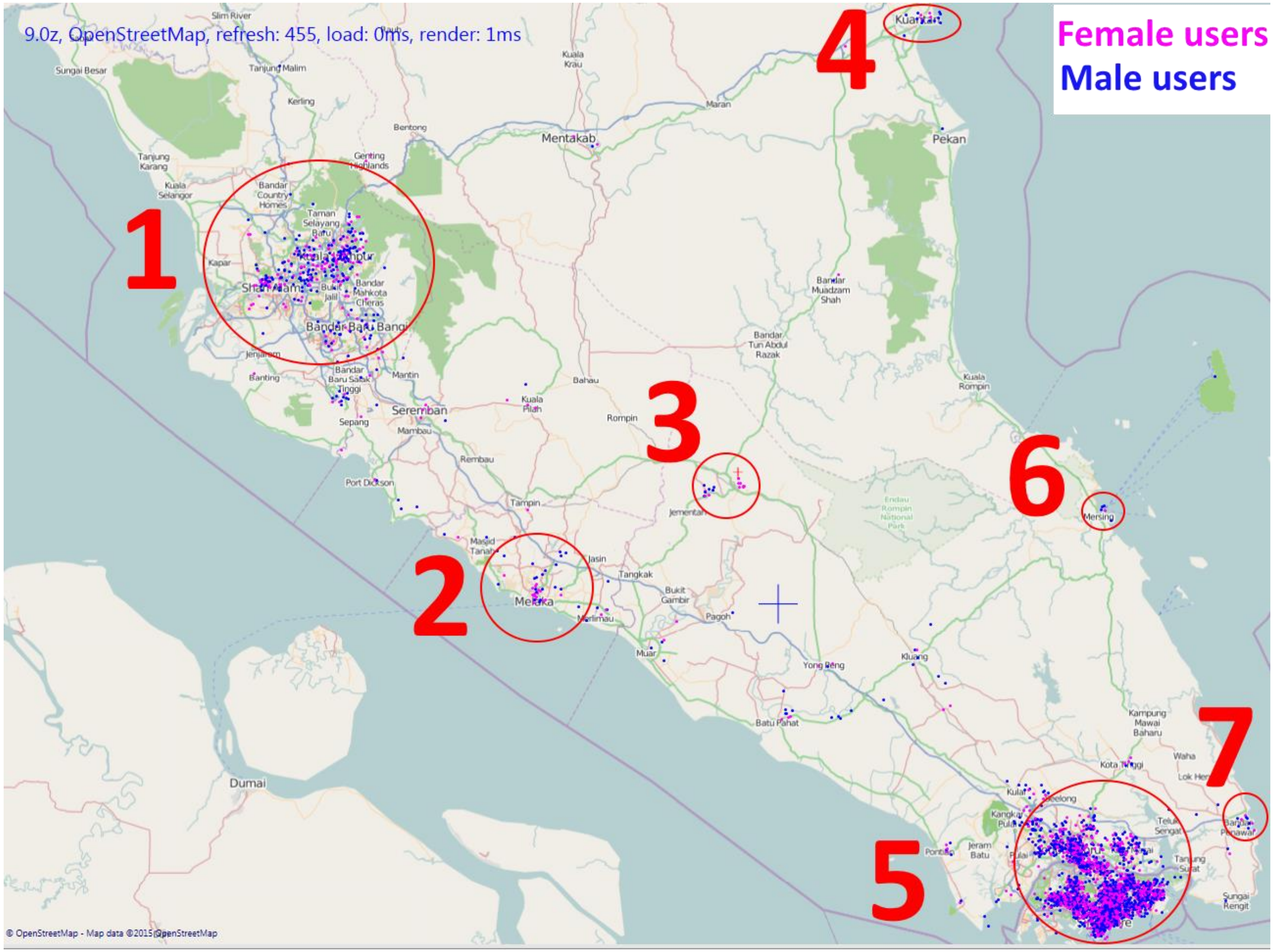


Geographical user mobility: users movement (city level)

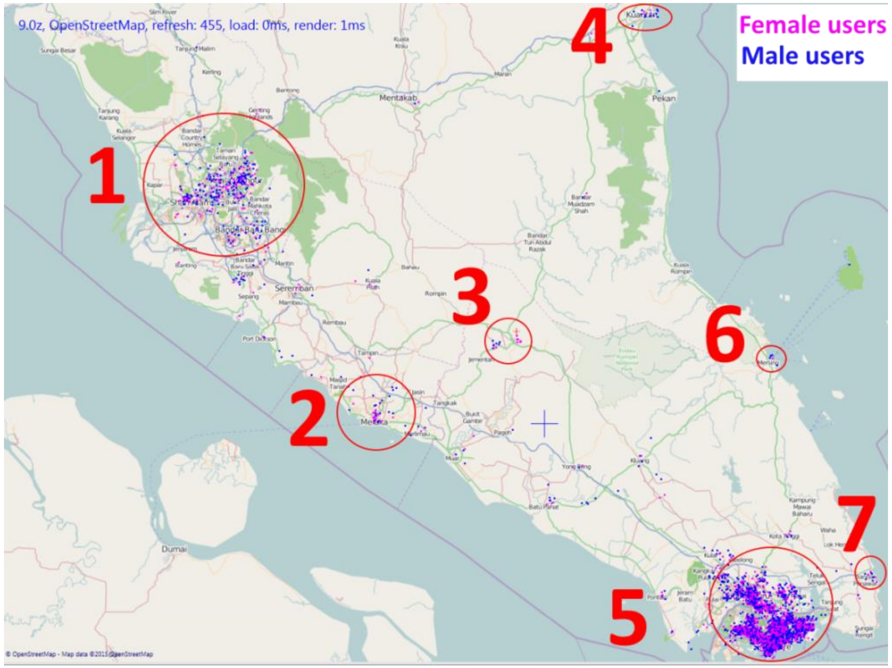


- Singapore population is concentrated in several regions, which represent peoples' housing (Regions 2 and 3) and working (Region 3) areas.
- There are some regions where male (Blue markers) user check-in density is much higher than female (Pink markers).

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)

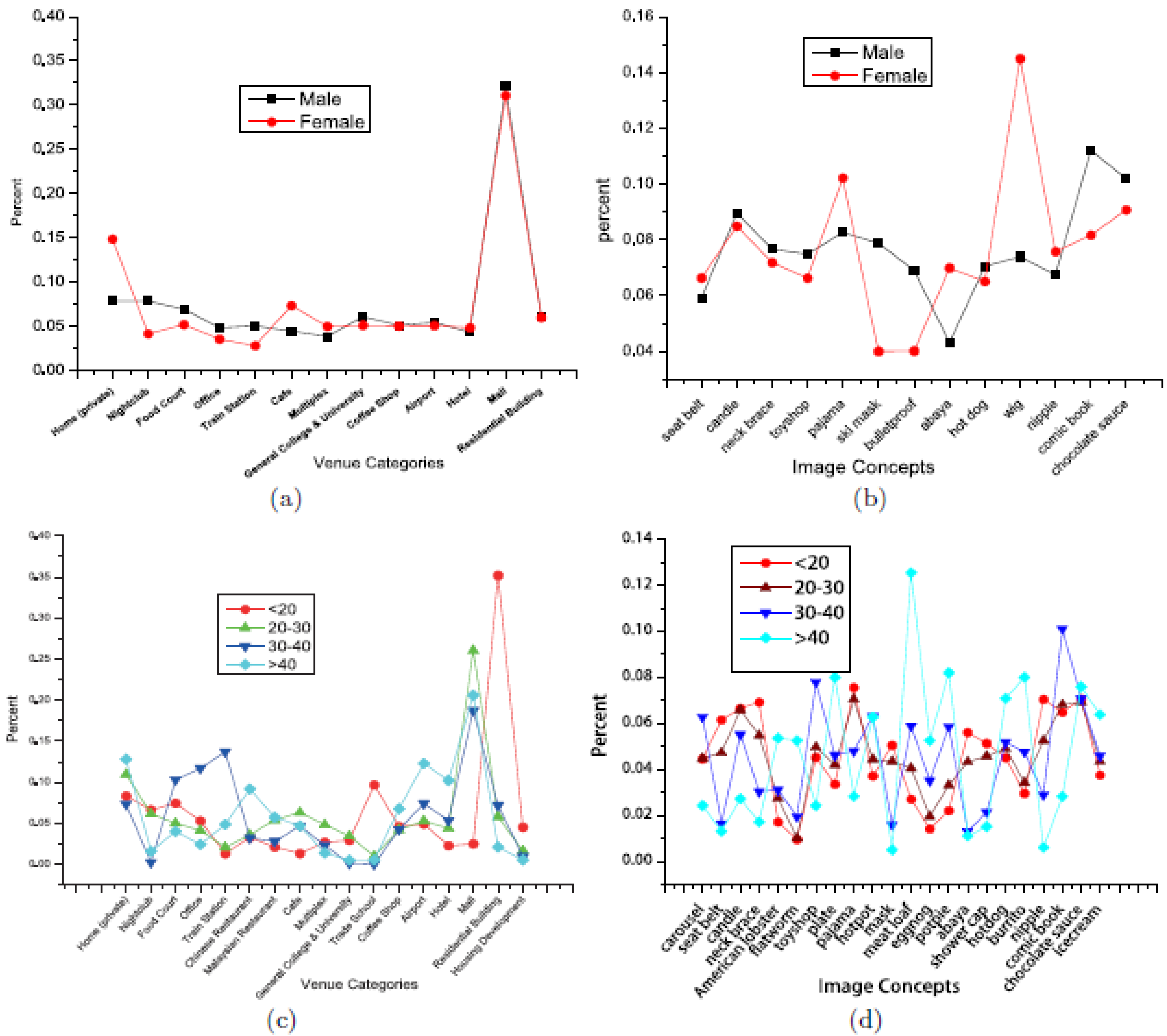
- Teenagers and children (Brown markers) mostly perform check-ins in housing city areas and around schools (Regions 1,2,3,5).
- Students (Green markers) and working professionals (Blue and Red markers) are concentrated in city center (Region 4).

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



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.

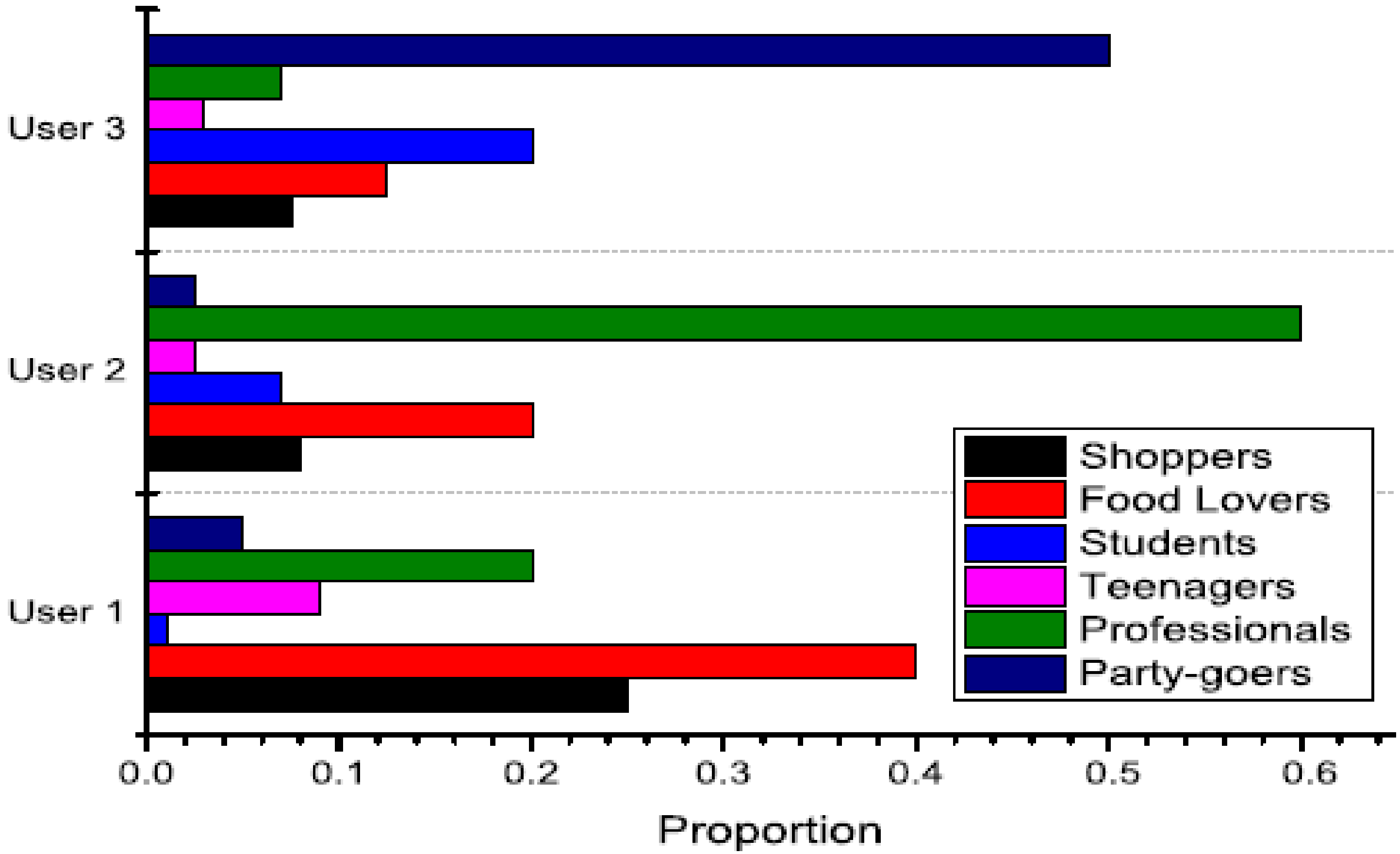
Every venue category

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

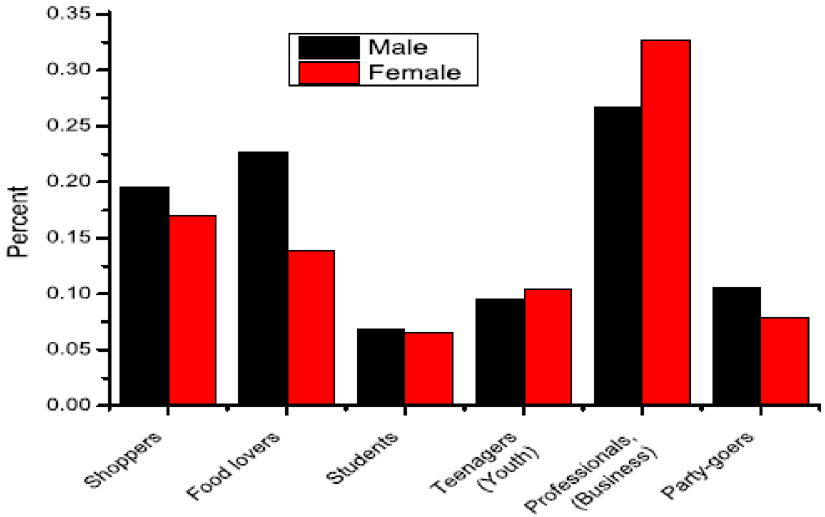
Table 2: Category distribution among LDA topics

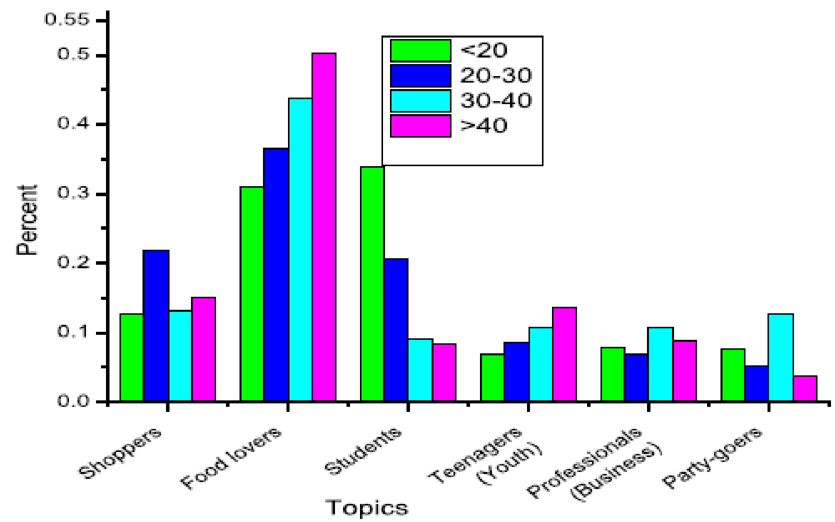
ID	Categories	LDA Topics
T1	Malay Res-t, Mall, University, Indian Res-t, Aisian Res-t	Food Lovers
T2	Cafe, Airport, Hotel, Coffee Shop, Chinese Res-t	Travelers (Business)
T3	Nightclub, Mall, Food Court, Trade School, Res-t, Coffee Shop	Party Goers
T4	Home, Office, Build., Neighbor-d, Gov. Build., Factory	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



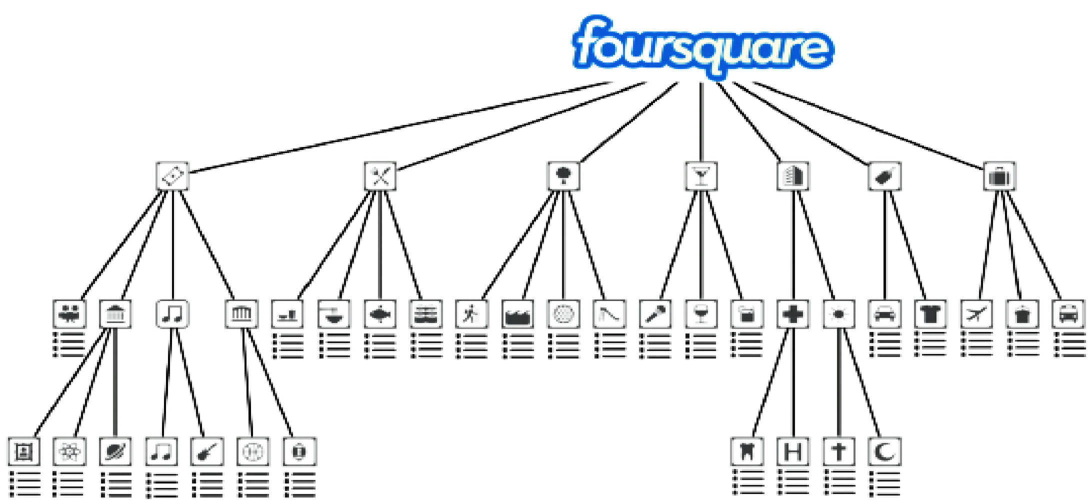


- Male users more often do shopping than male, while female users often show-up in job-related venues.
- > 30 years old users often show-up in dining-related places, while < 20 – often visit education-related venues.

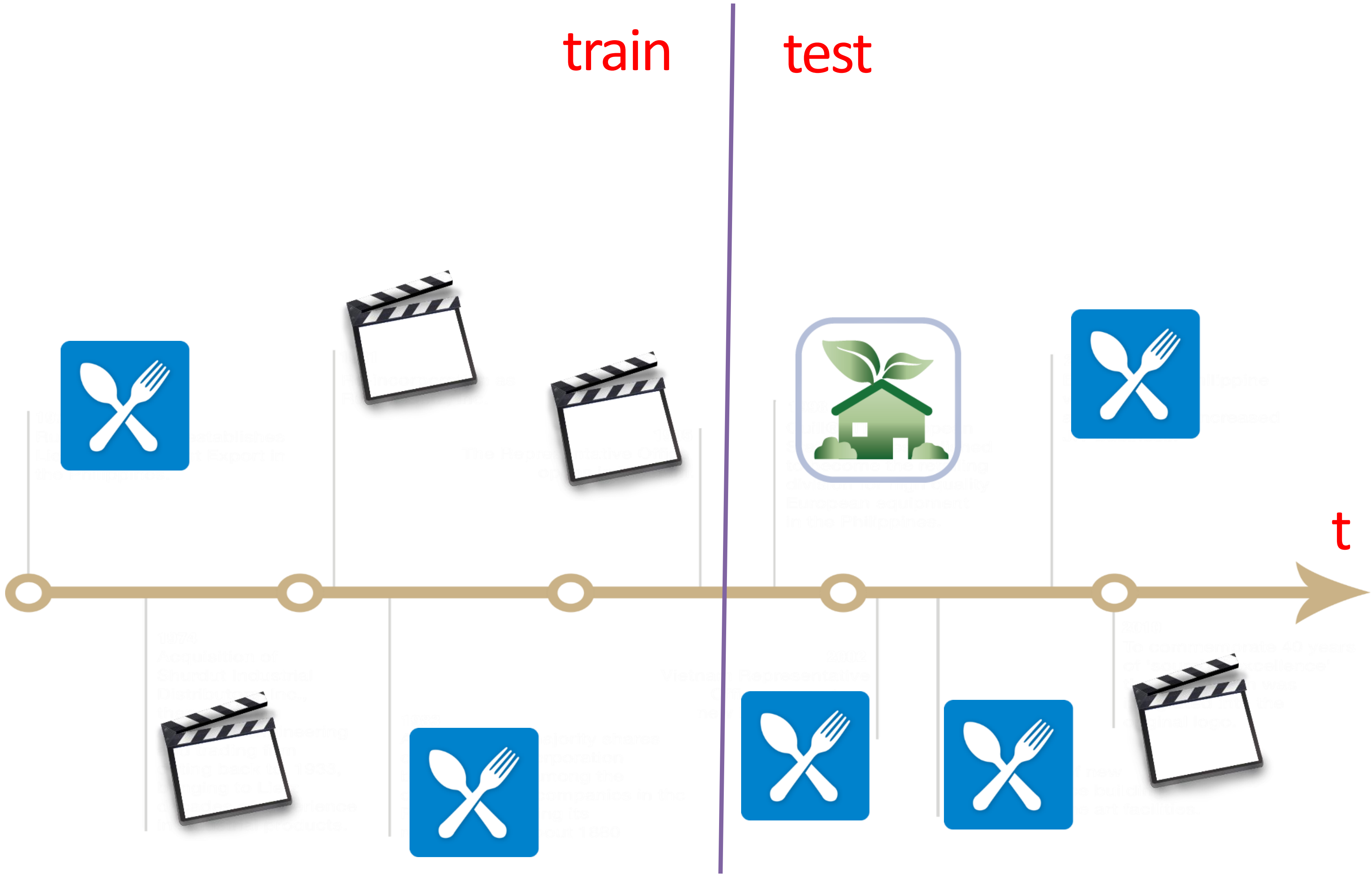
Semantic user mobility:

Venue Category Recommendation

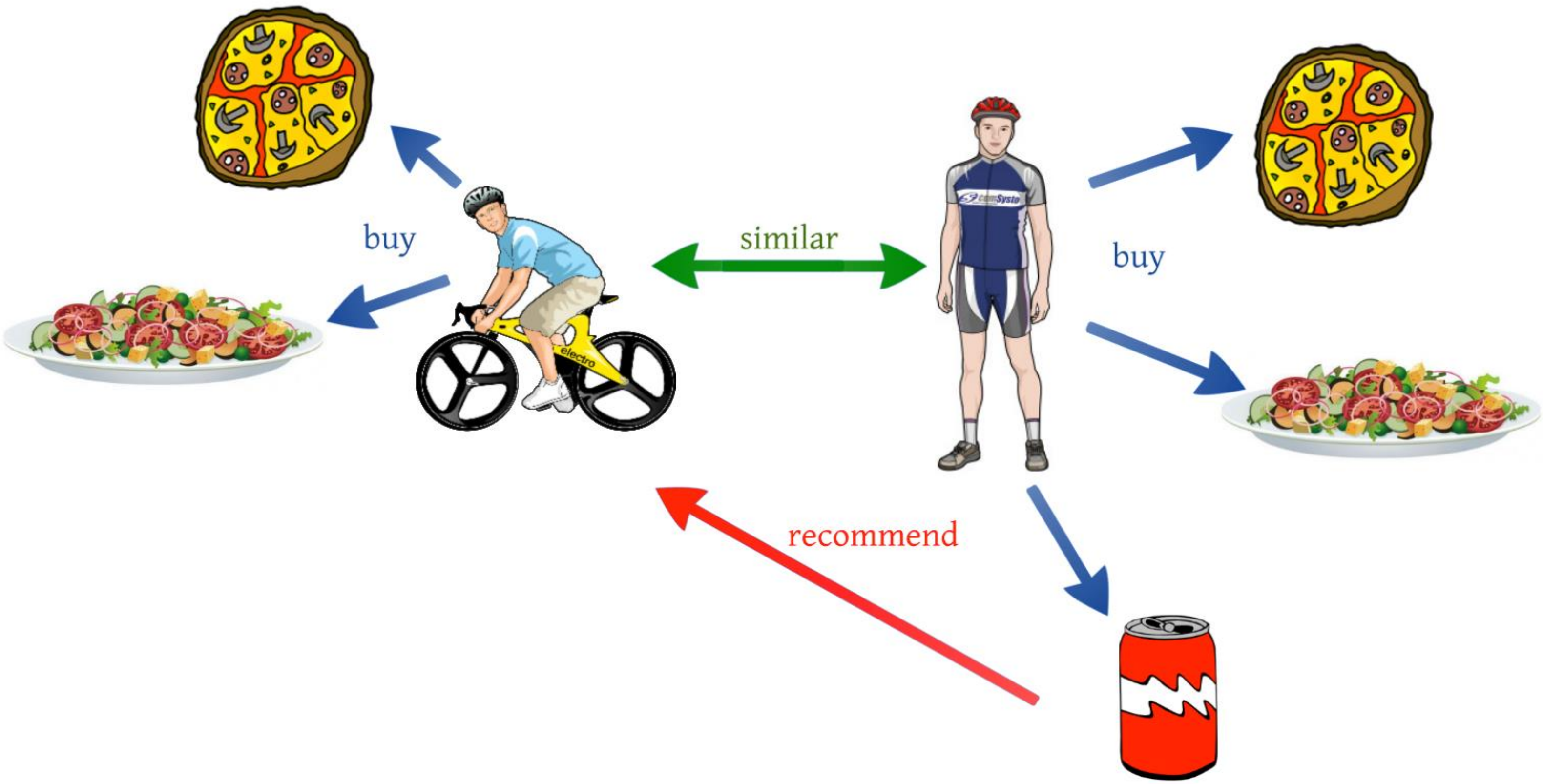
Which category(s) of 4sq venues to go next?



Evaluation – split time on train and test periods



We use Collaborative Filtering (CF)



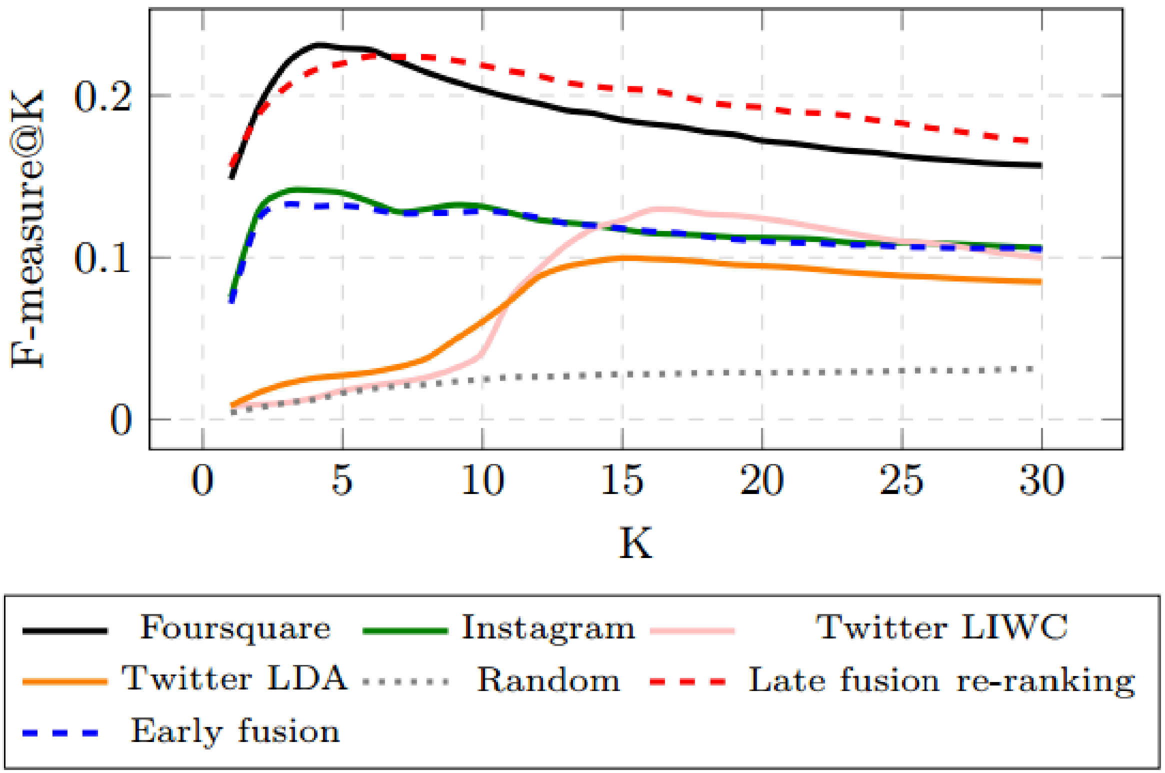
Multi-Source re-ranking

- 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 else can be done?

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

Other tasks based could be approached

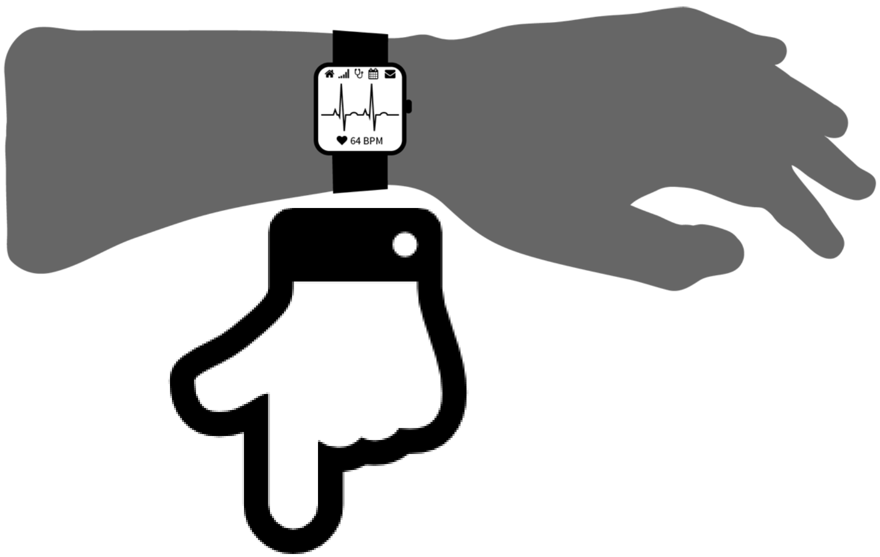
1. **Demographic profile learning**
2. **Multi-source data fusion**
3. **Individual and group mobility analysis**
4. **Cross-source user identification**
5. **Cross-region user community detection**
6. **Cross-source causality relationships extraction**
7. **Users' privacy-related and cross-disciplinary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

User Profile Learning in Wellness Domain

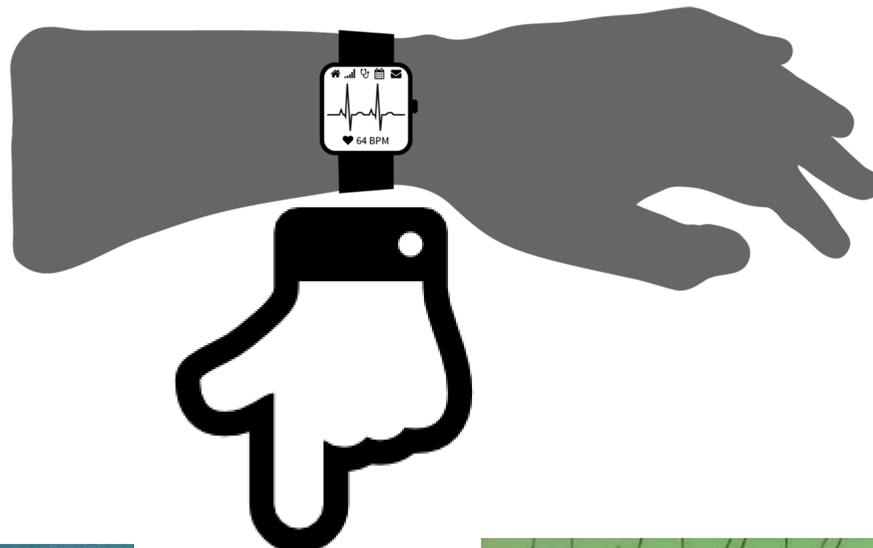
People are often now aware of their wellness problems

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

User profile

Wellness profile

Mobility profile

Demographic profile

Diabetes

Asthma

Obesity

Location preference

Movement patterns

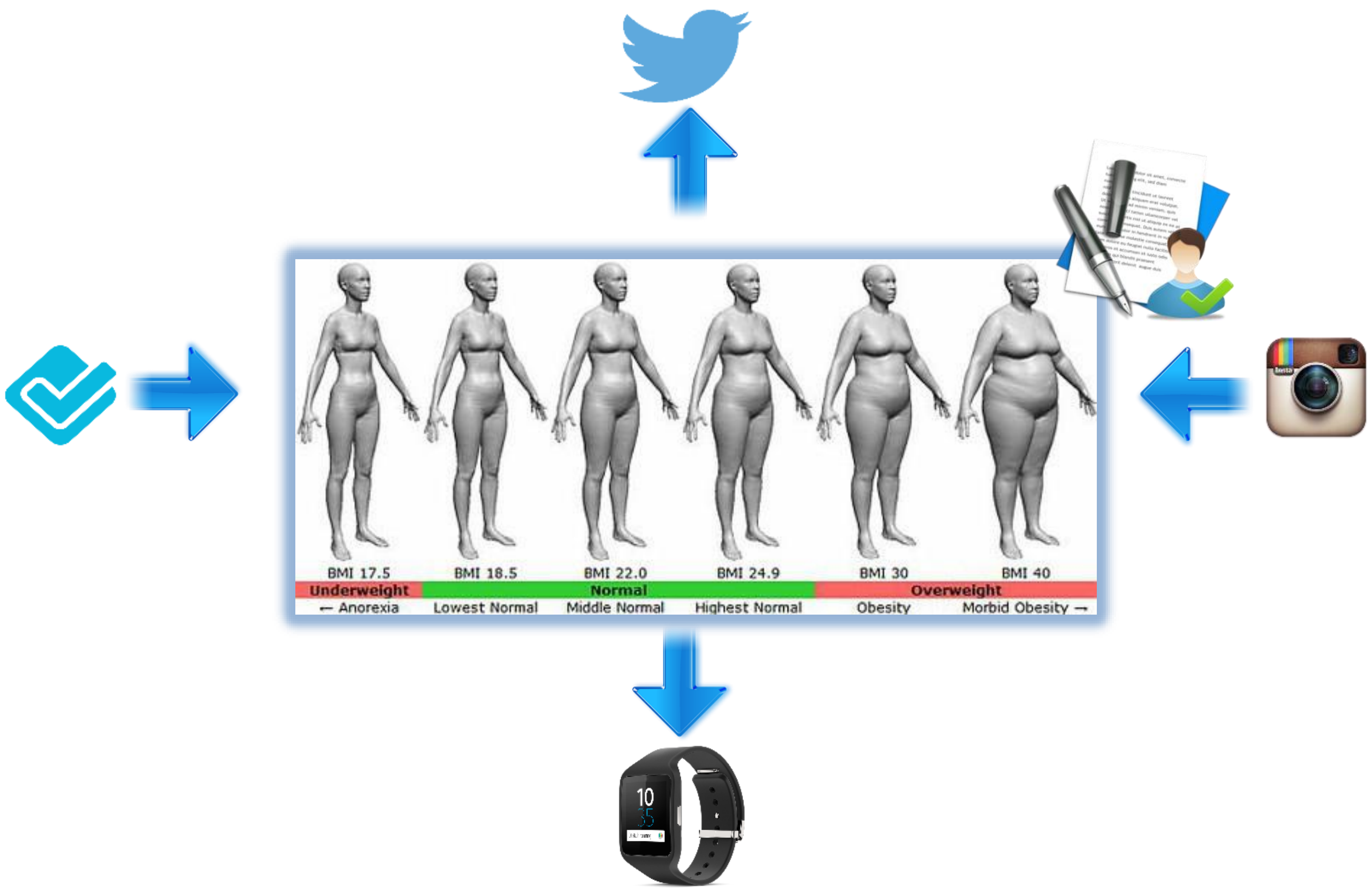
Age

Gender

Personality

Occupation

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

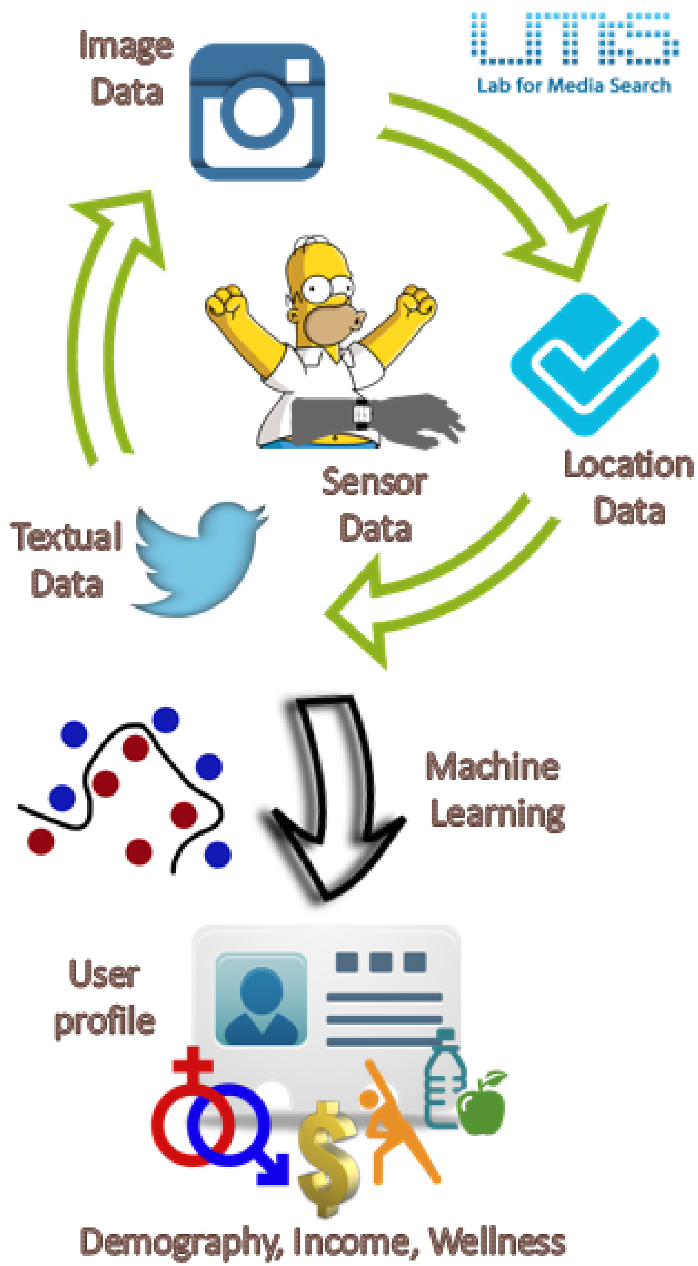
- We constructed and released a large multi-source multi-modal cross-region "NUS-MSS" dataset;
- We conducted first-order and higher-order learning for user mobility and demographic profiling;
- New multi-modal features were proposed for a demographic profile learning.
- Based on our experimental results,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

You, actually can join us as Intern or Research Engineer http://next.comp.nus.edu.sg/opportunities

E-mail to:

farseev@u.nus.edu

As Research Intern



The Project

User profile learning, such as **mobility, wellness, or demographic profile** learning, is of **great importance** to various applications. Meanwhile, the rapid growth of multiple social platforms makes it possible to perform a **comprehensive user profile learning from different views.** In our project, **we construct large-scale multi-source multi-modal datasets,** apply **machine learning techniques** on it to **infer various user profile attributes**, deploy large-scale data analytics platforms.

Requirements and Benefits

Requirements:

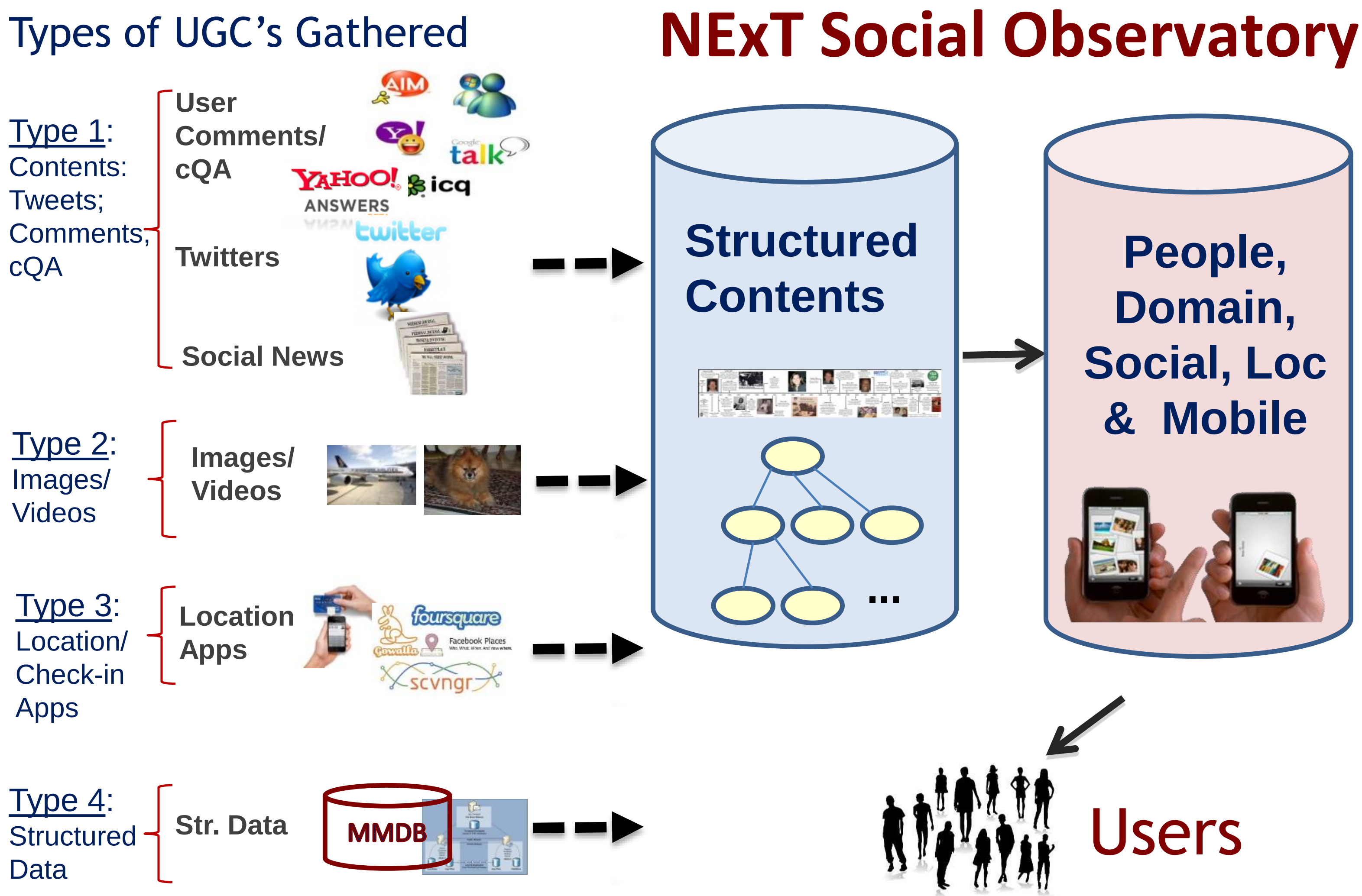
- Bachelor/Master/PhD full-time student (or graduate)
- **Strong programming background** (C#, Java, Python, R, MathLab, etc.)
- **Strong mathematical background** (Probability Theory, Linear Algebra, Convex Optimization)
- **Machine learning background**

Benefits:

- **1 year (extendable) Internship in** NUS - leading Asia's university and of Top10 in the world
- Internship in **leading Asia's Social Media Lab** LMS@NUS
- **Opportunity to publish** your work in Top-ranked journals and conferences
- **Finance allowance** provided.
- Opportunity to apply for PhD in NUS based on the internship results

http://nusmulsitource.azurewebsites.net

As Java Research Engineer



http://live.nextcenter.org

You, actually can join us as Intern or Research Engineer

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

E-mail to:

farseev@u.nus.edu

AINL-ISMW FRUCT OPEN DAY

Tat-Seng CHUA
Chair Professor School of Computing, National University of Singapore

Social Media Analytics: What has changed over the last 5 years.

Registration:
You Name and Job Place to:
office@ainlfruct.com or
+7 (921) 438-80-77

7-9, Universitetskaya nab.
(Здание Двенадцати Коллегий)
Start: 11:00 AM