Multi-Source Big Data

How to obtain and process various types of user-generated data.

by Aleksandr Farseev

nus.academia.edu/farseev

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Agenda

Brief summary of the lecture...

1 Textual Data

1. N-Grams
2. Topics
3. Dictionaries
4. Heuristics

2 Visual Data

1. Bag of Visual Words
2. Visual Concepts
3. Deep features

3 Location Data

1. Venue (POI) semantics
2. Spatial Features
3. Temporal Features
4. Mobility Features

4 Sensor Data

1. Exercise semantics
2. Statistical Features
3. Frequency Features

5 Data Gathering

1. Collecting Multi-Source Data

“

Big data is not actually about the data. The revolution is not that there’s more data available.

The revolution is that we know what to do with it now…

- Gary King

Multiple social networks describe users from multiple views

Some facts about social networks…

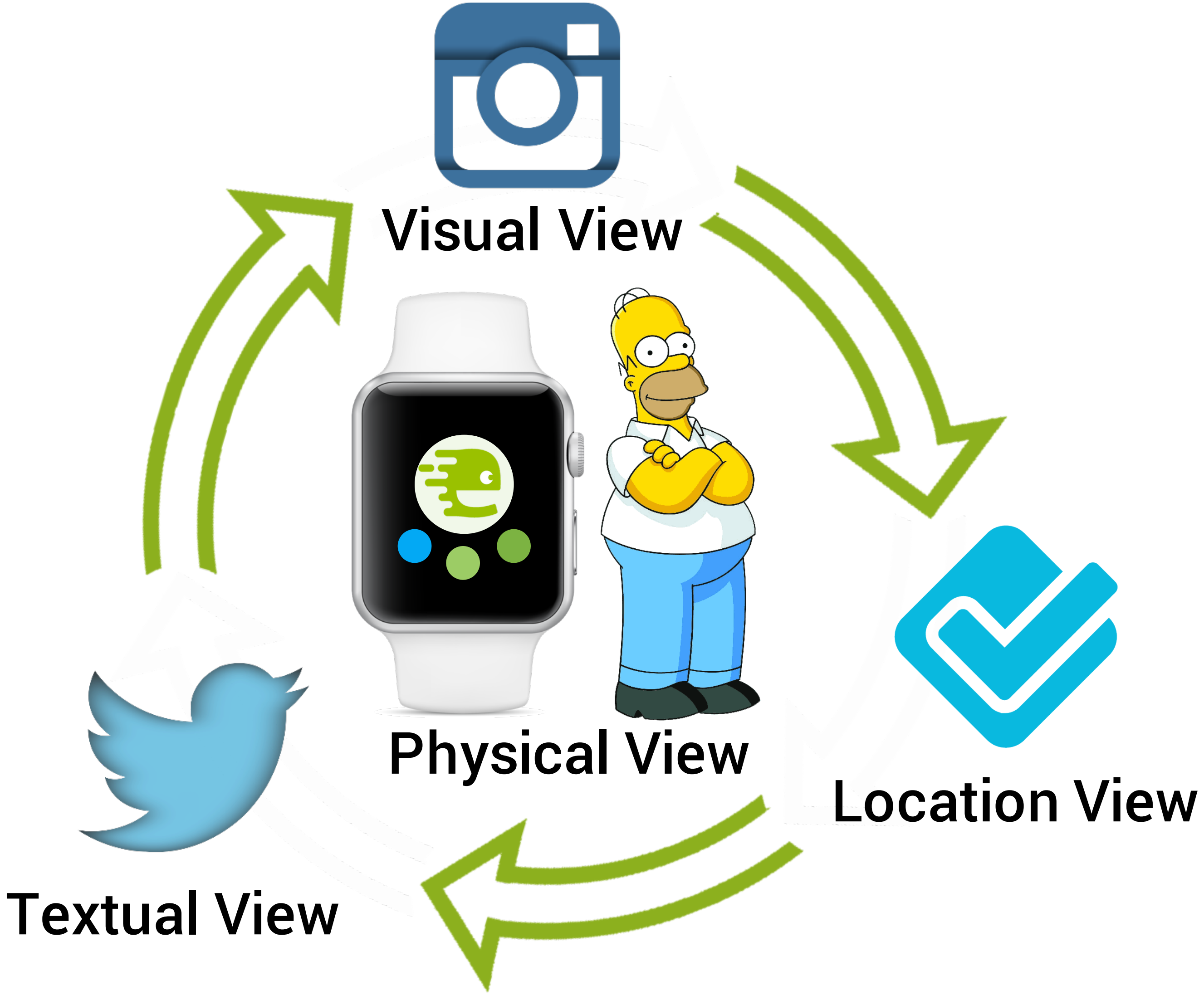
More than 50% of online-active adults use more than three social networks in their daily life*

*According Paw Research Internet Project's Social Media Update 2014 (www.pewinternet.org/fact-sheets/social-networking-fact-sheet/)

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Different data modalities describe users from multiple views

Indeed, they are:



Textual Data

Reference

*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.

N-Gram Models of Language

- Use word sequences of length n = 1… k, called n-grams
- Language Model (LM)
  - unigrams (n = 1), bigrams (n = 2), trigrams,…
- How do we obtain such data representations?
  - Very large corpora – Why?

Simple N-Grams

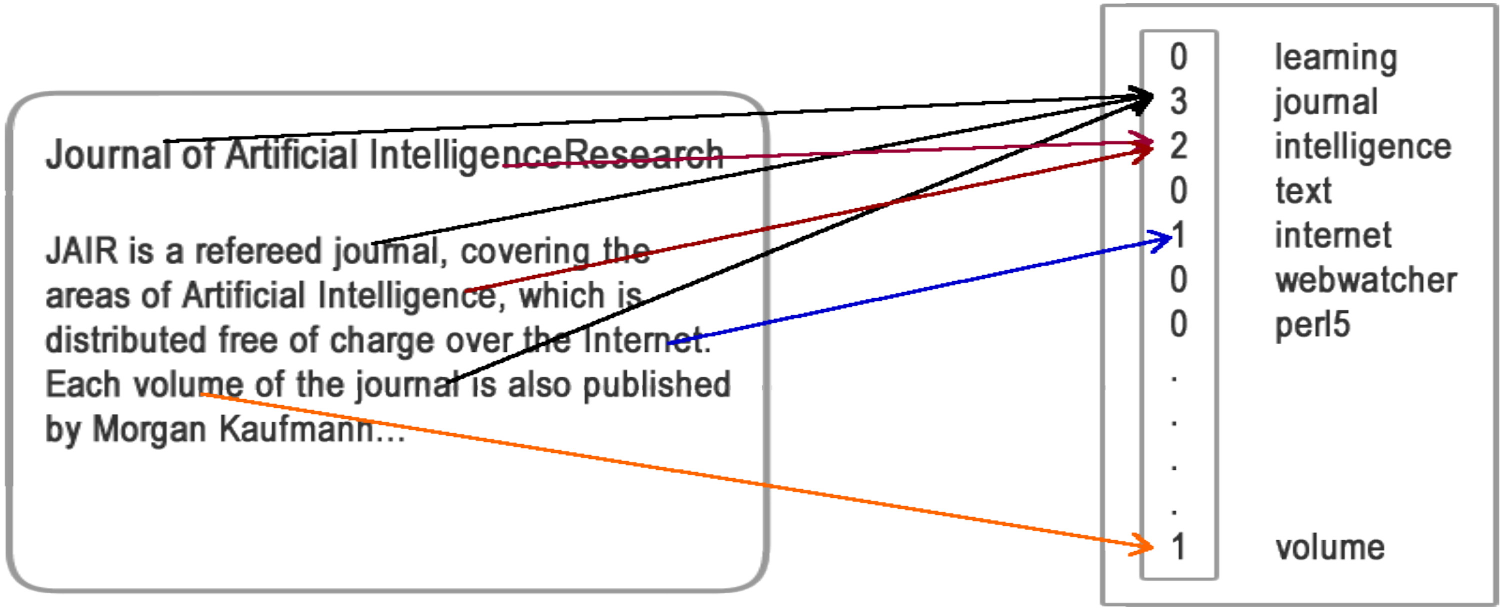
Assume a language has T words in its lexicon, how likely is word x to follow word y?

- Simplest model of word probability: 1/T
- Alternative 1: estimate likelihood of x occurring in new text based on its general frequency of occurrence estimated from a corpus (unigram probability)

  popcorn is more likely to occur than unicorn
- Alternative 2: condition the likelihood of x occurring in the context of previous words (bigrams, trigrams,…)

  mythical unicorn is more likely than mythical popcorn

Bag of N-Grams



- Count occurrences of each term (n-gram)
- Pick top N most frequent as a Bag of Terms (n-grams): T = $\{t_1, t_2, t_3, \ldots, t_N\}$
- Each document D for user u is represented by normalized terms (n-grams) distribution among N most frequent terms (n-grams): $D_u = \{\frac{t_{1u}}{N}, \frac{t_{2u}}{N}, \frac{t_{3u}}{N}, \ldots, \frac{t_{Nu}}{N}\}$

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.

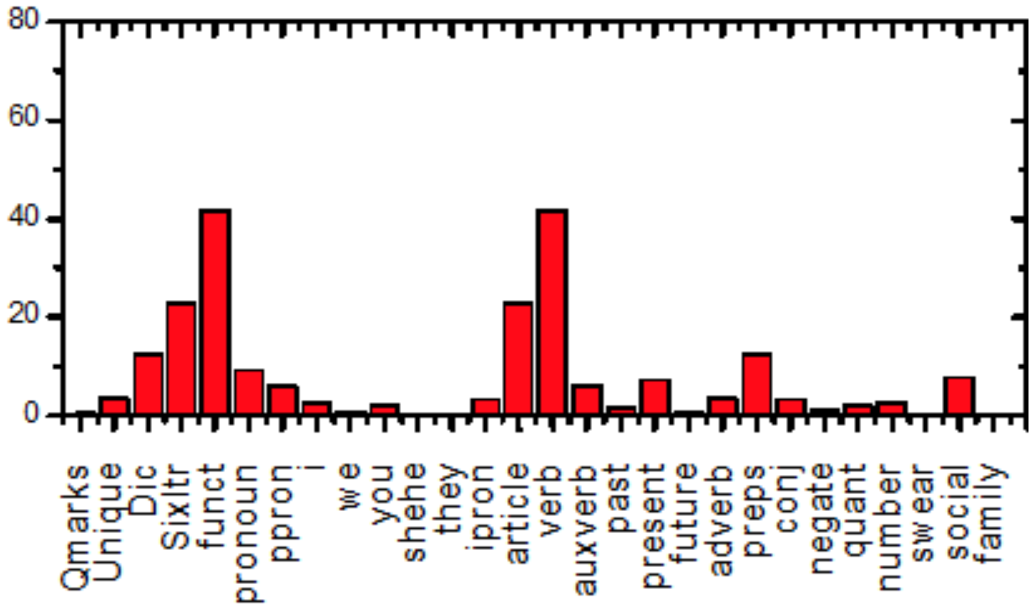
- Task – Word usage analysis* and correlation with personality
- Data – Various essays and questionnaires
- Approach – manual personality-related dictionaries construction
- Findings:

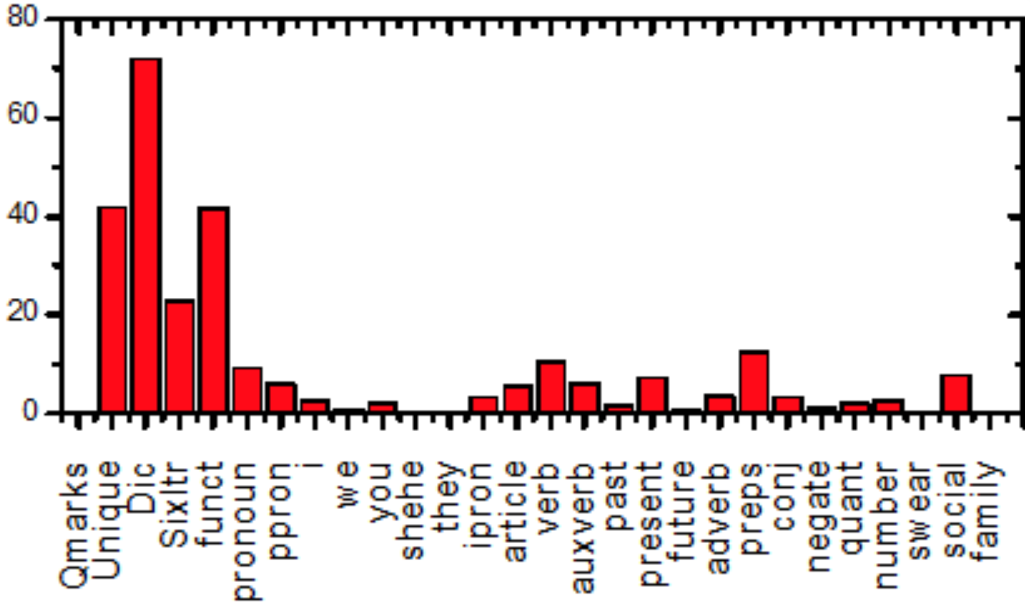
  Certain word usage statistics are good indicators for human personality profiling

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

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

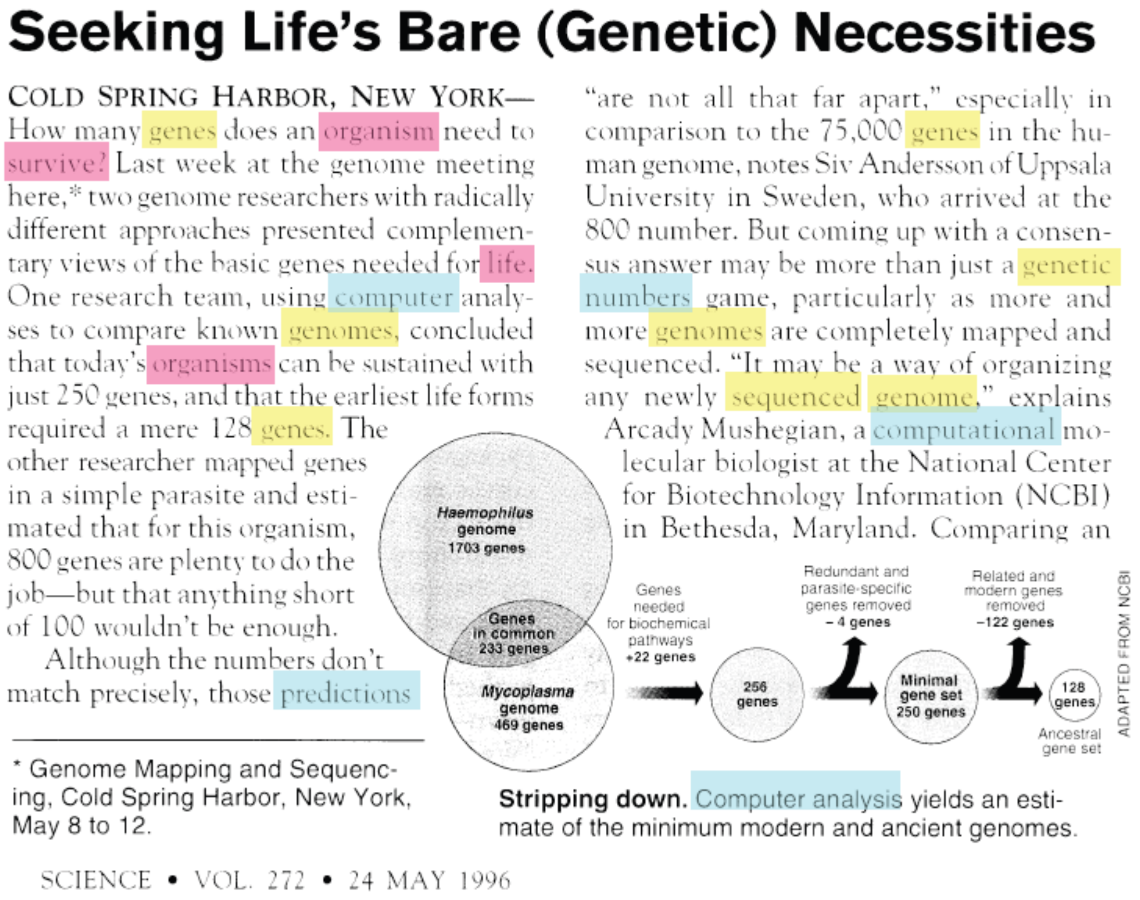
LIWC





- Count occurrences of each words that belong to each LIWC category
- Each document D for user u is represented as a distribution among 74 LIWC categories: $D_u = \{\frac{\mathrm{LIWC}_u}{N}, \frac{\mathrm{LIWC}_{2u}}{N}, \frac{\mathrm{LIWC}_u}{N}, \dots, \frac{\mathrm{LIWC}_u}{N}\}$

Topic Modeling (1)



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.

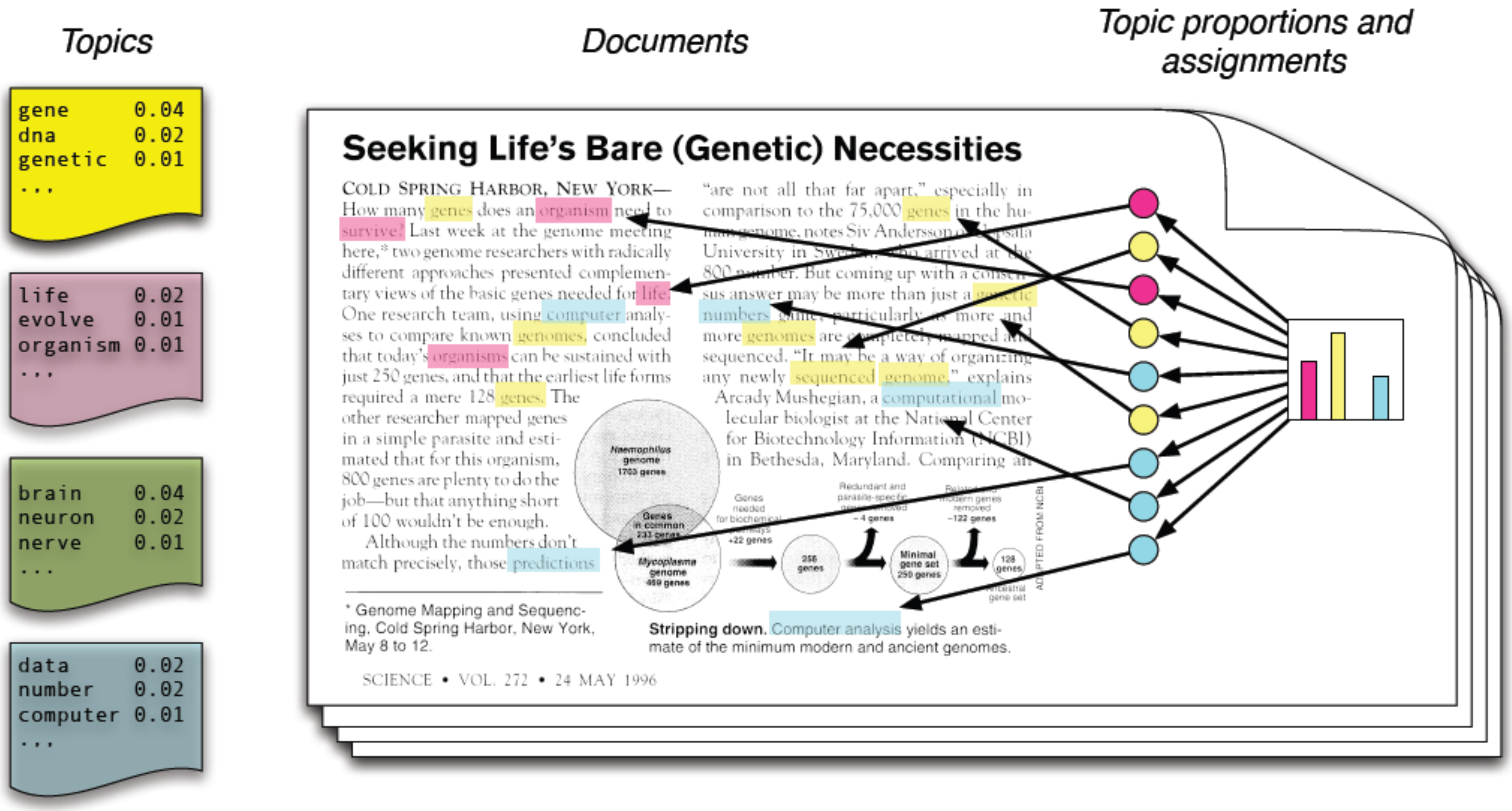
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

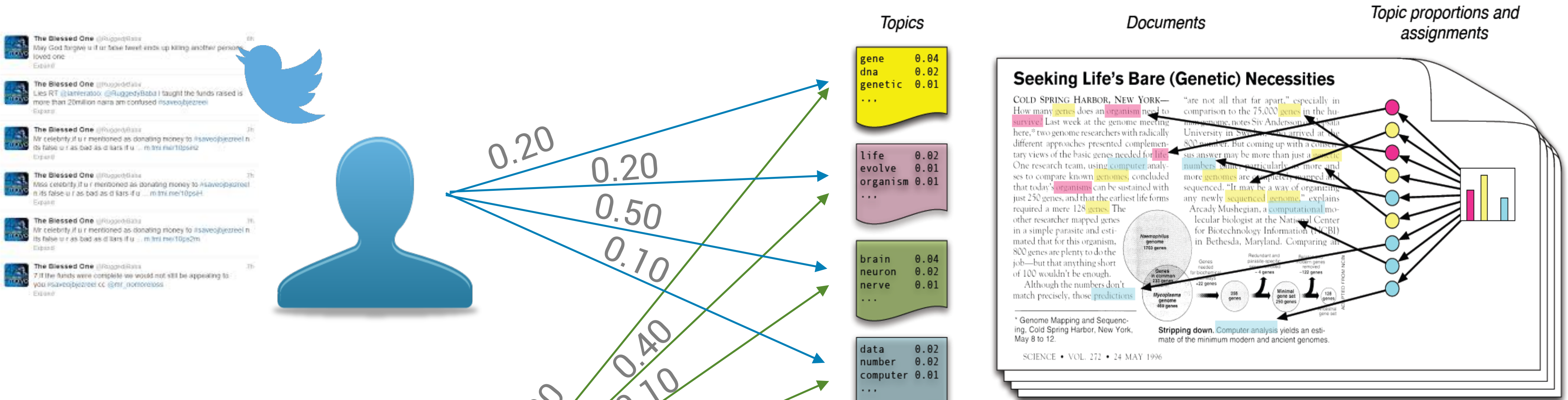
- Methods for automatically organizing, understanding, searching and summarizing large electronic archives.
- Uncover hidden topical patterns in collections.
- Annotate documents according to topics.
- Using annotations to organize, summarize and search.
- Widely poplar approach: Latent Dirichlet Allocation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

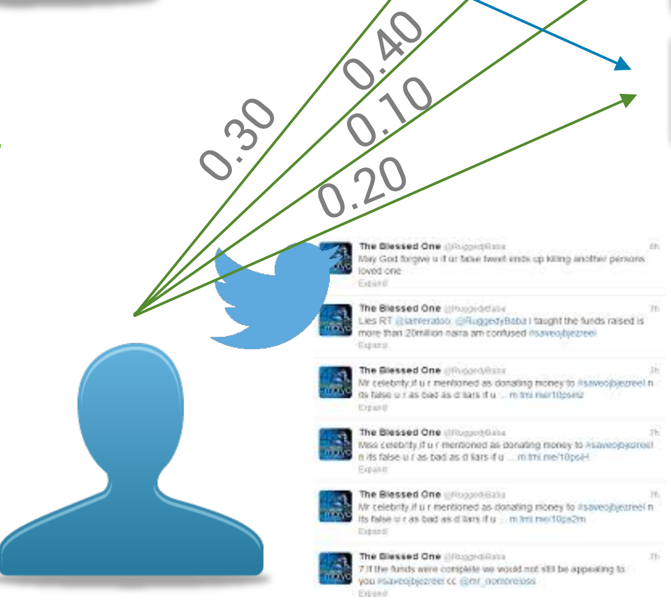
*D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation" The Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

Topic Modeling (2)



Topic Modeling (3)





Each document D_u for user u is represented by normalized LDA topic distribution:

$D_u = \{P(LDA_1|u), \ldots, P(LDA_K|u)\}$

- 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.

Behavioral Features

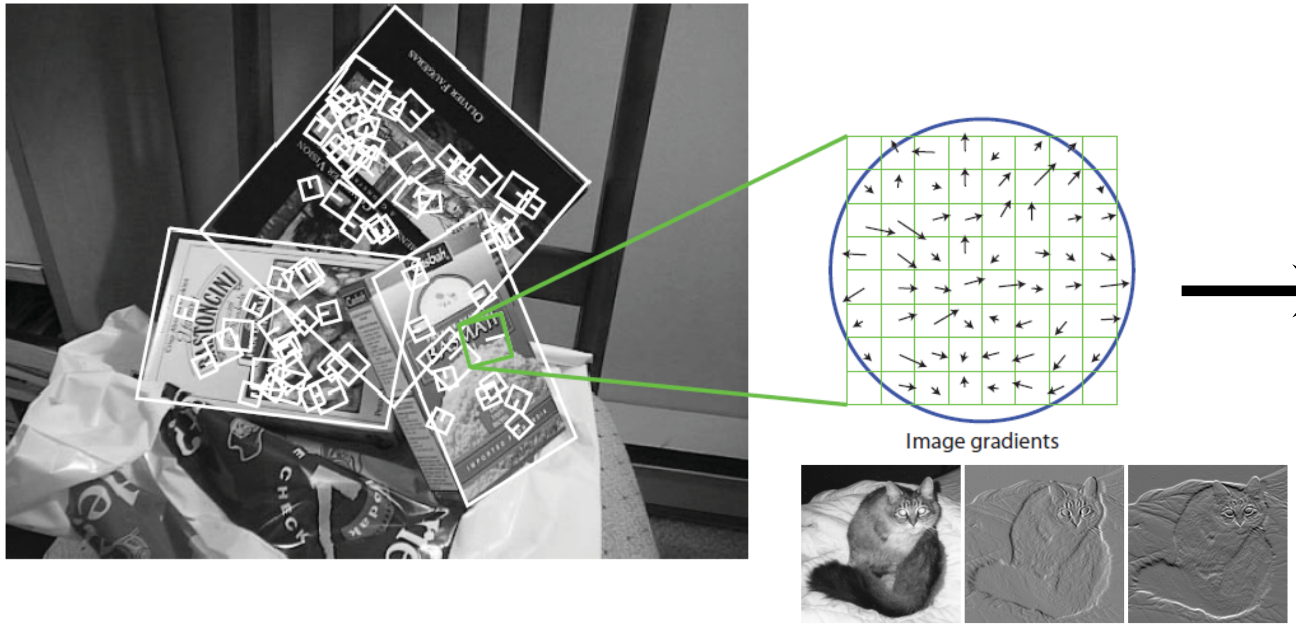
Data representation	Description
Number of hash tags	Number of hash tags mentioned in message
Number of slang words	We calculate number of slang words / tweet and compute average slang usage
Number of URLs	Number of URL's one usually use in his/her tweets
Number of user mentions	Number of user mentions – 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 texts	Number of tweets where 70% of words either not in English or cannot be fixed by spell checker
Number of terms average	Average number of terms per / tweet

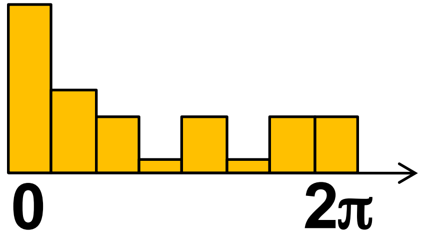
Visual Data

Reference

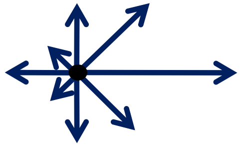
*A. Farseev, N. Liqiang, M. Akbari, and T.-S. Chua. Harvesting multiple sources for user profile learning: a Big data study. ACM International Conference on Multimedia Retrieval (ICMR). China. June 23-26, 2015.

Scale Invariant Feature Transform (SIFT) descriptor (1)





angle histogram



Basic idea: use edge orientation representation:

- Take 16x16 square window around detected feature
- Compute edge orientation for each pixel
- Filter out weak edges (threshold gradient magnitude)
- Create histogram of surviving edge orientations

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

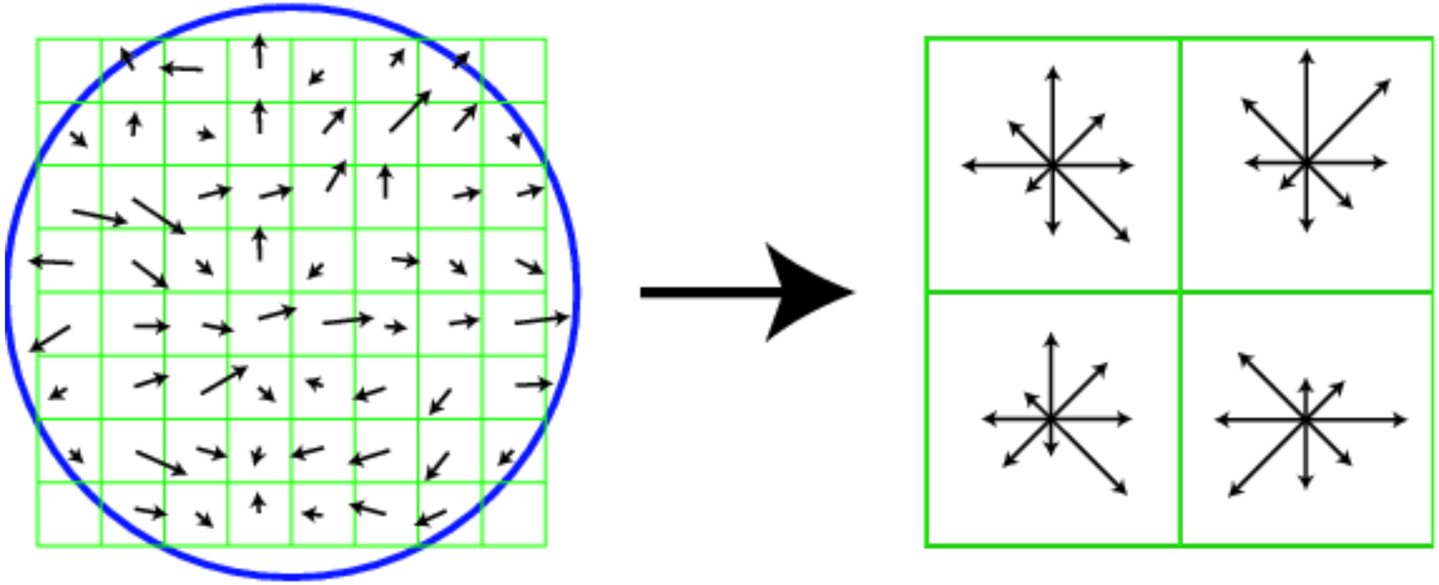


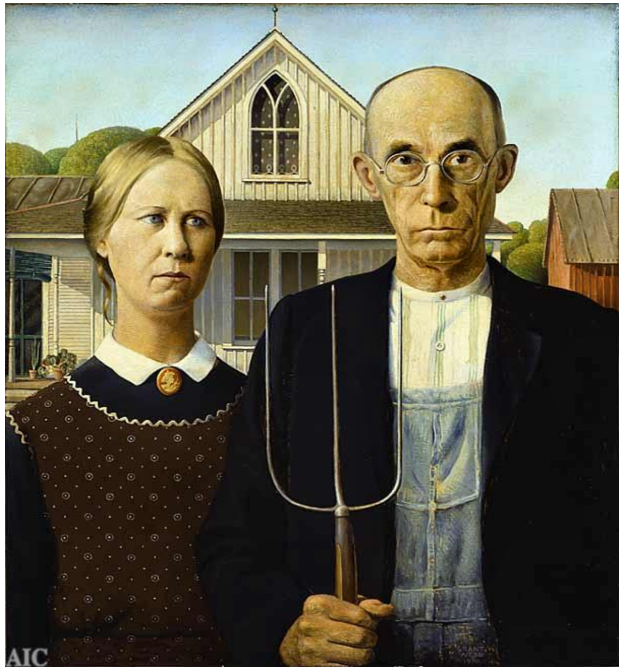
Image gradients

Keypoint descriptor

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

Bag of visual words (1)

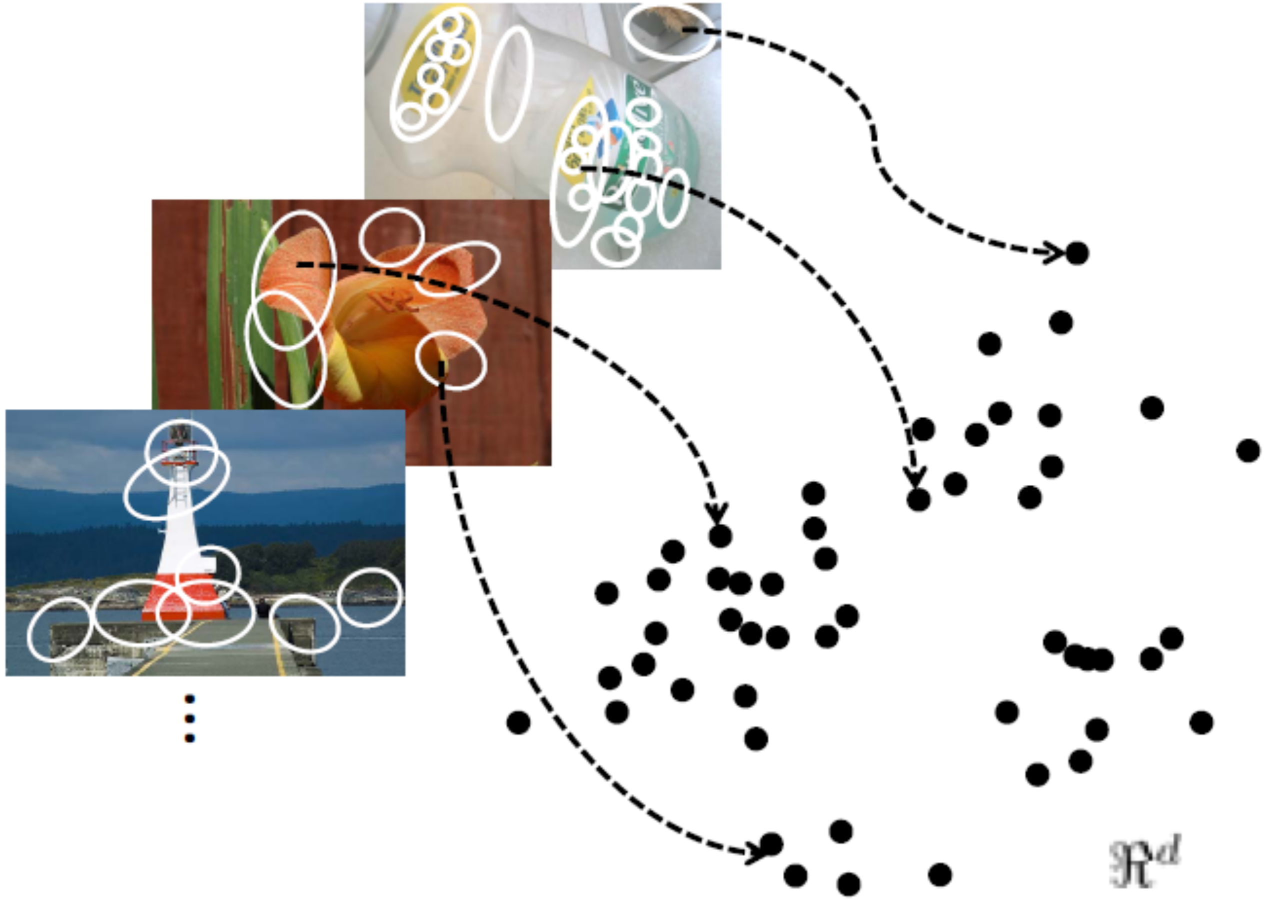
Can images be represented as Bag-of-Visual Words?



?

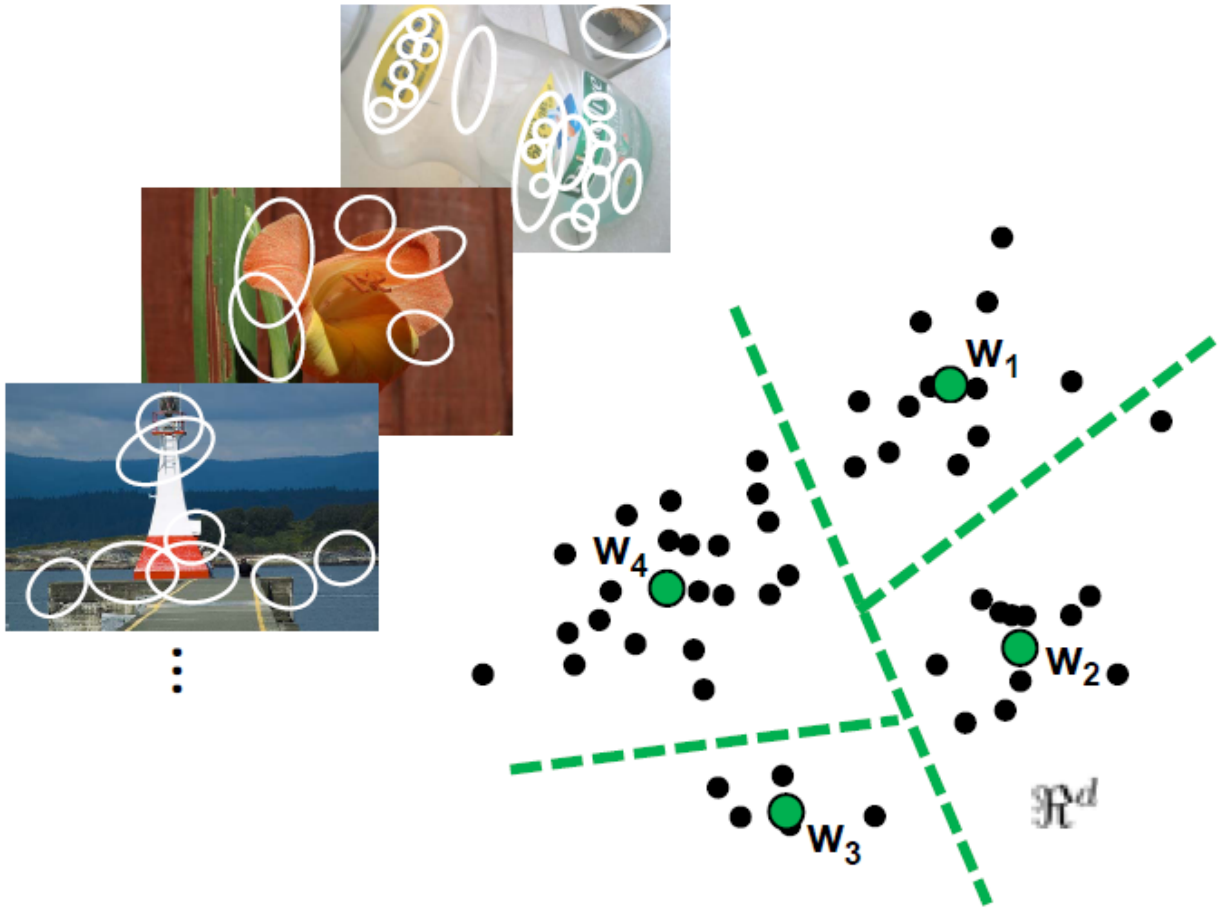
Idea: quantize SIFT descriptors of all training images
to extract representative visual words!

Bag of visual words (2)



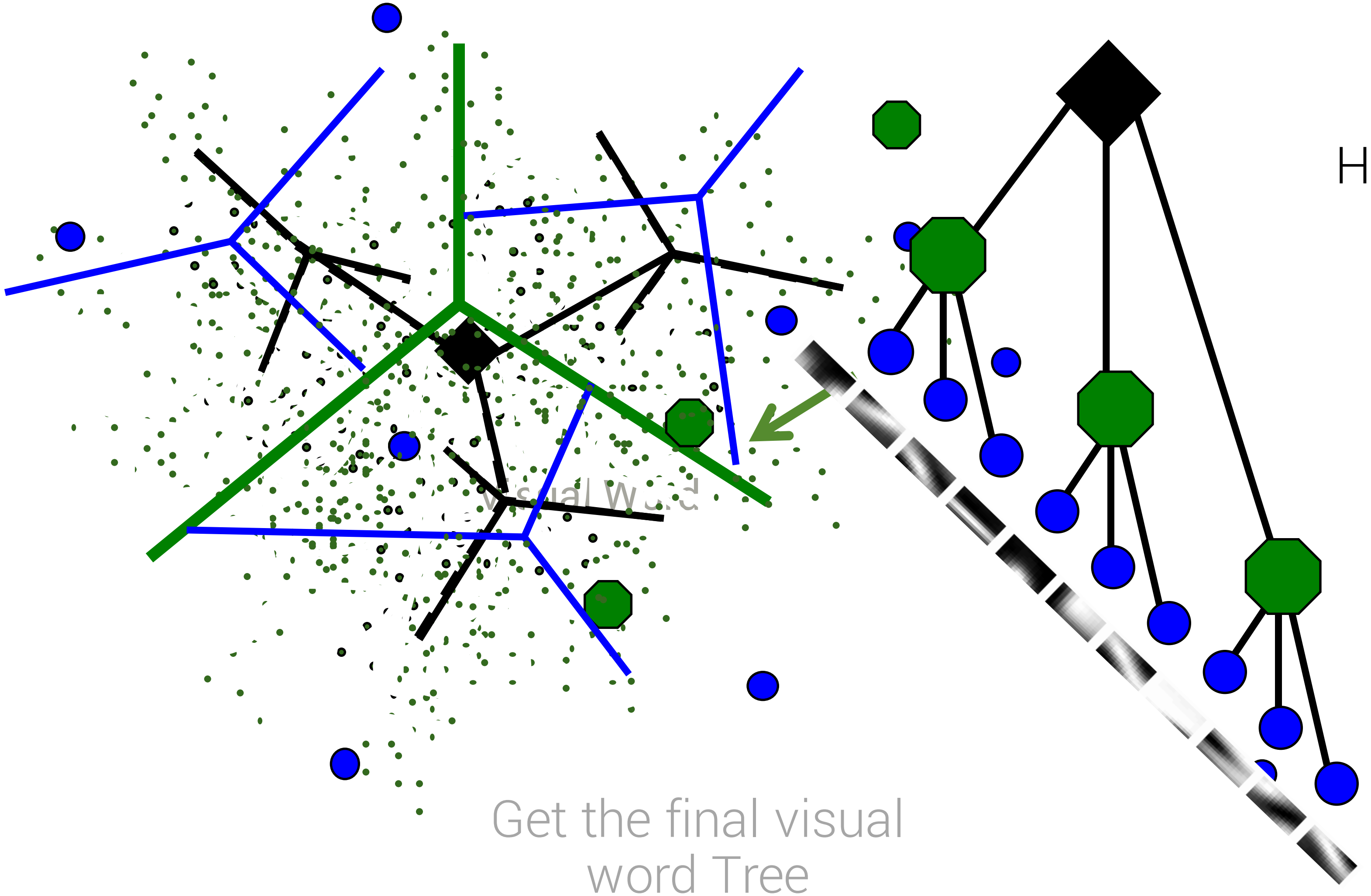
Step 1: Extract interest points of all training images

Bag of Visual Words (3)



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

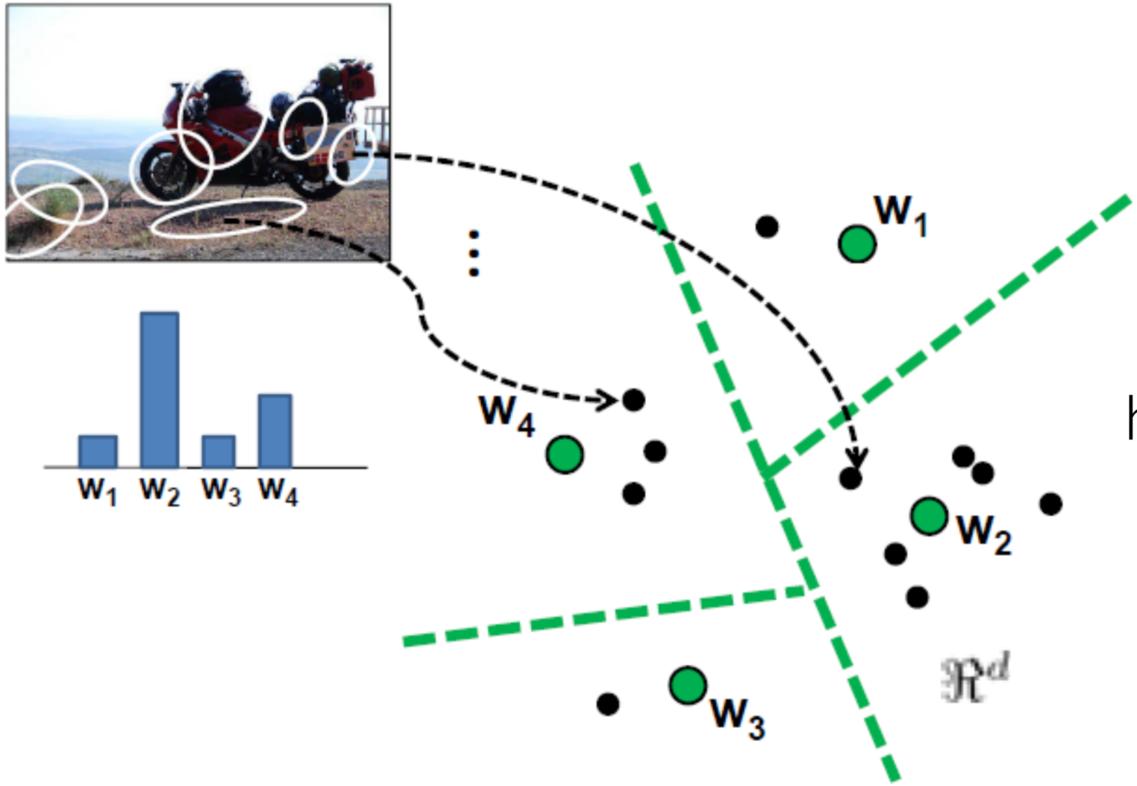
Bag of Visual Words (4)



Hierarchical k-means clustering

Get the final visual word Tree

Bag of Visual Words (5)



Step 3: Summarize (represent) each image as histogram of visual words and use as basis for matching and retrieval!

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Next Step: Image Concept Detection (1)

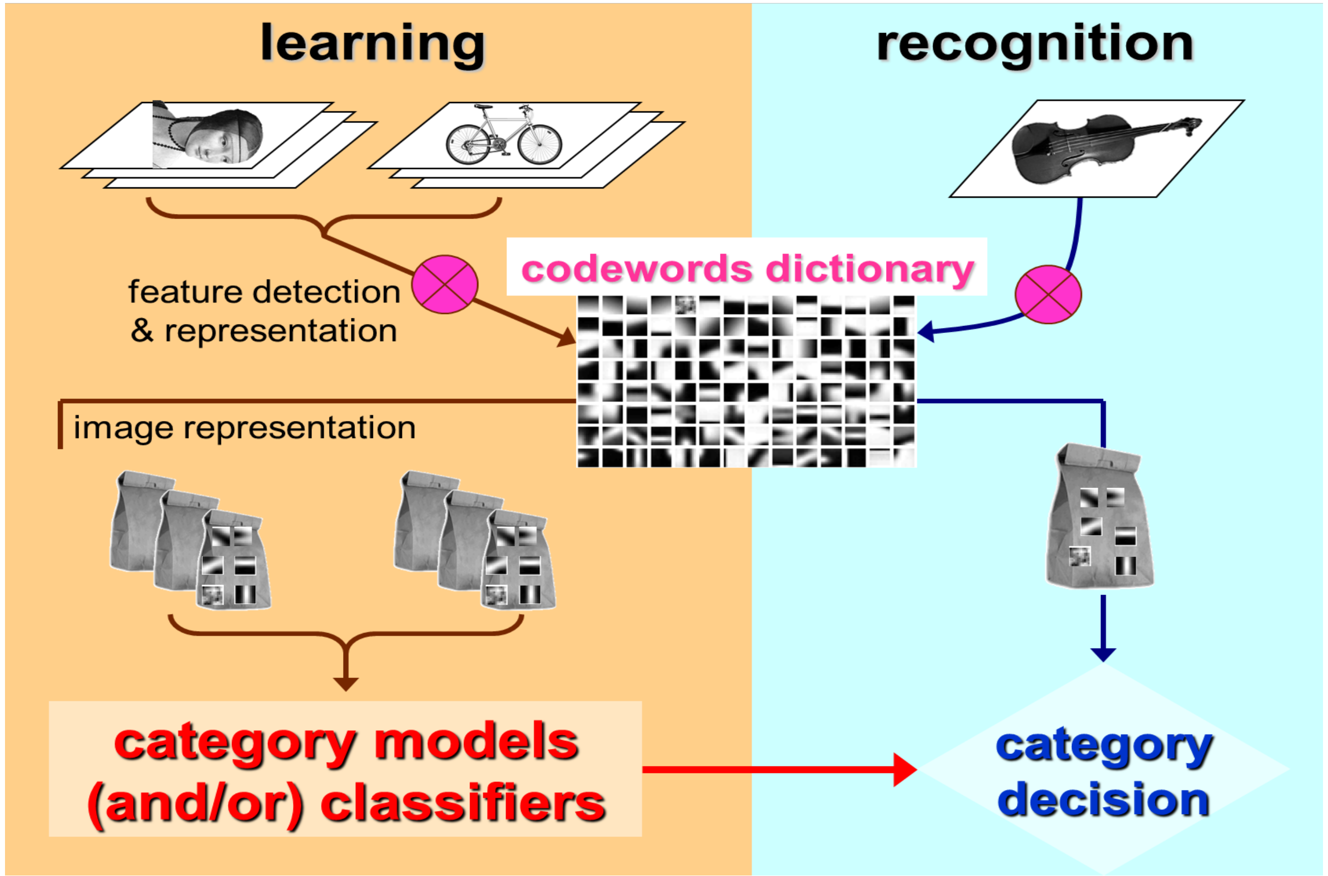
Tree concept detection answers the question: "Are there Trees on the picture?"

Next Step: Image Concept Detection (2)

It also answers the question:
"What objects are on the picture?"

E-Mail: farseev@u.nus.edu I Website: http://farseev.com

One way to detect image concepts

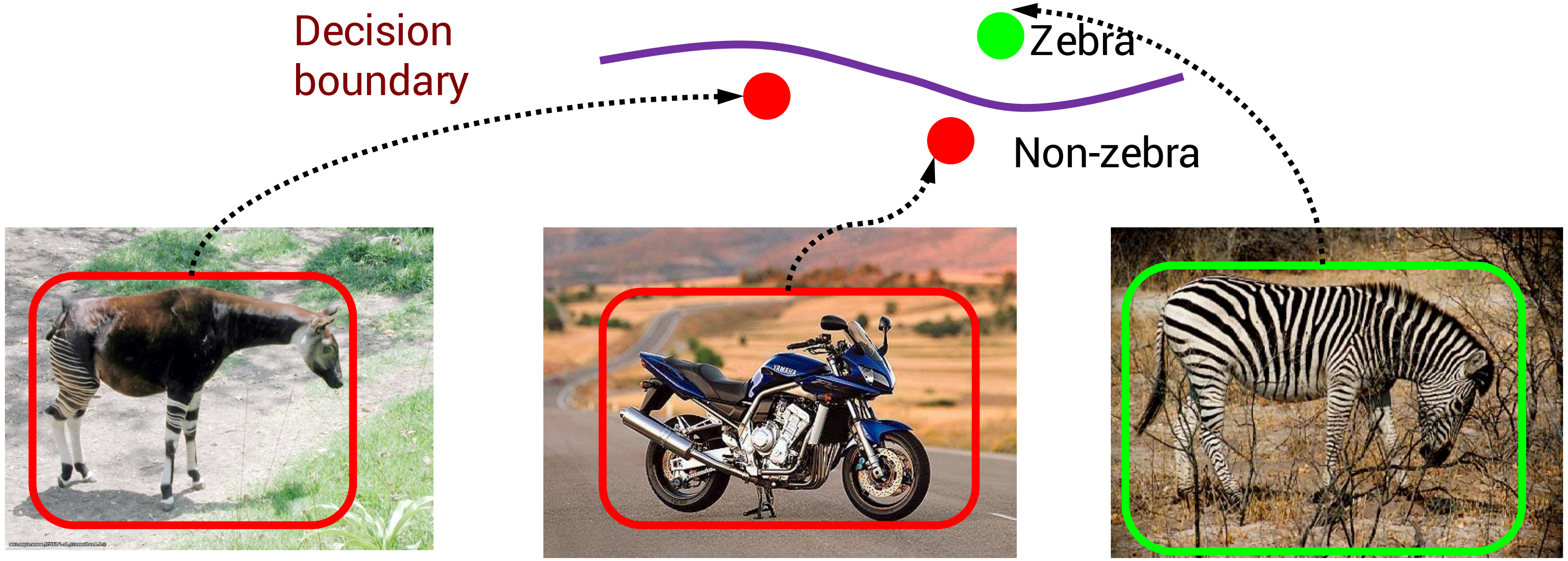


1. Representative set of images in each category is collected
2. An image is represented by a collection of "visual words"
3. Object categories are modeled by the distributions of these visual words

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Concept Detection: Discriminative Model

- Object detection and recognition is formulated as a classification problem. The image is partitioned into a set of overlapping windows, and a decision is taken at each window about if it contains a target object or not.
- Each window is represented by a large number of features that encode info such as boundaries, textures, color, spatial structure.
- The classification function, that maps an image window into a binary decision, is learnt using methods such as SVMs or neural networks



Location Data

Reference

*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.

Location-Based Social Networks

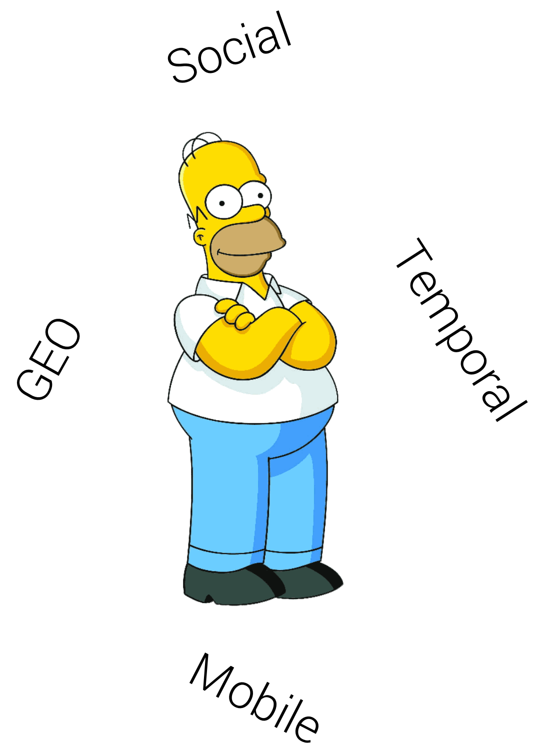
People want to share their geographic position with their friends.

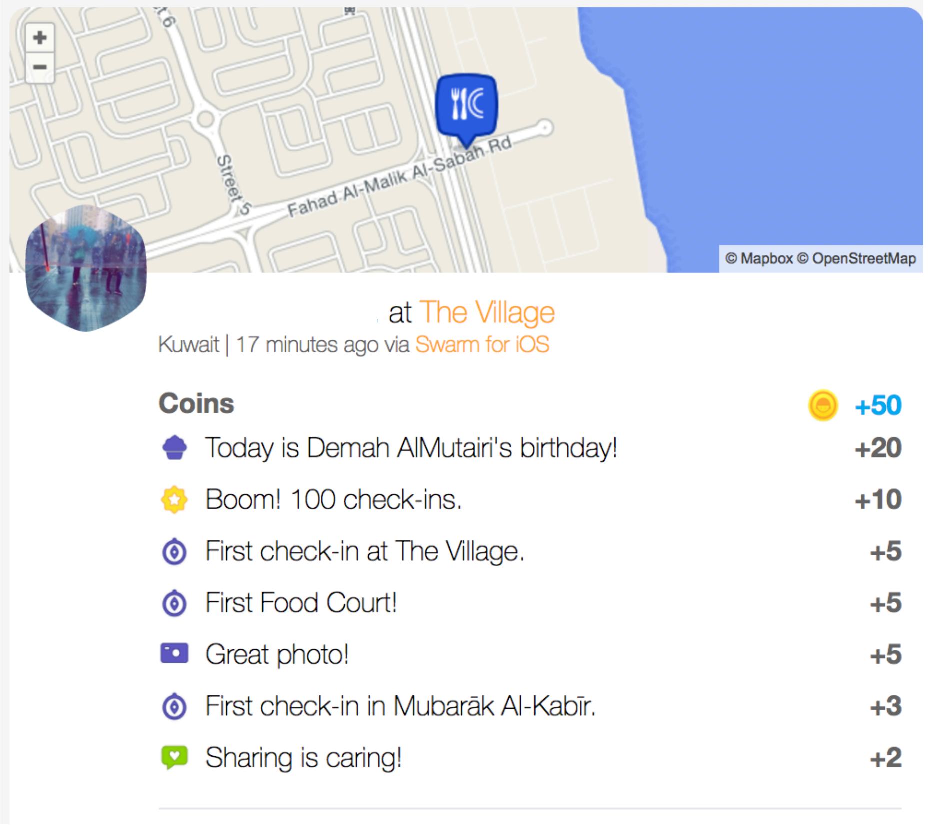
E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Foursquare and Facebook are main players

Foursquare and Facebook are the main players
with billions of check-in records.

The Multi-Dimensional Check-In

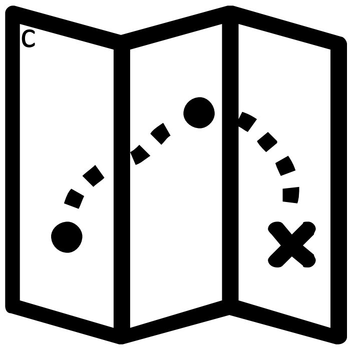


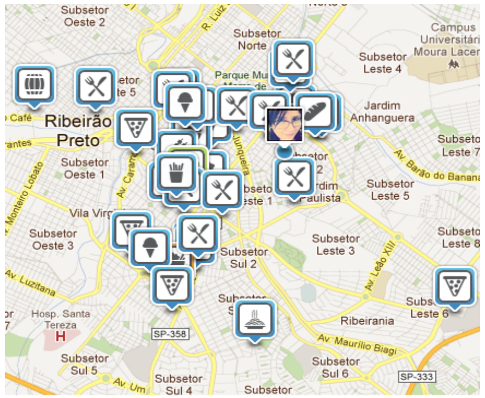


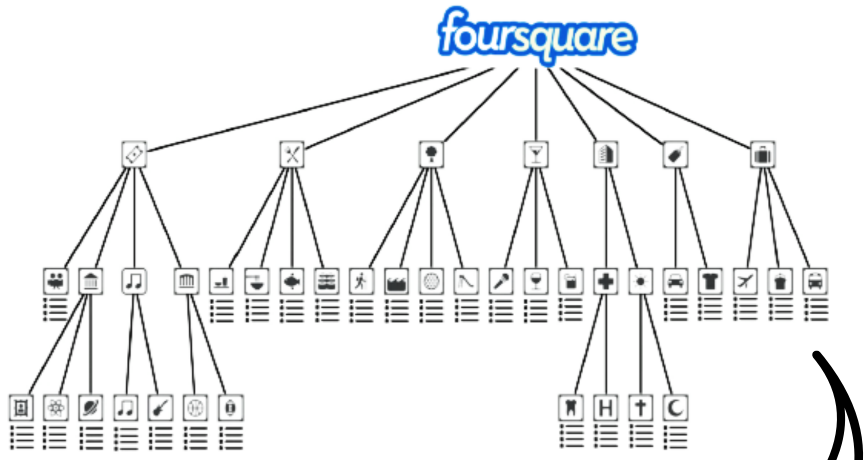
E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Location Data Representation: Utilizing Venue Semantics

Venues are not just GEO points, but multidimensional objects with rich semantics







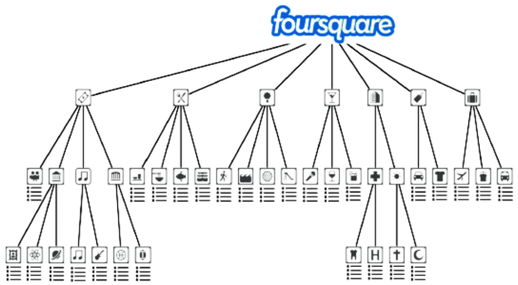
For case when user performed check-ins in two restaurants and airport but did not perform check-ins in other venues:

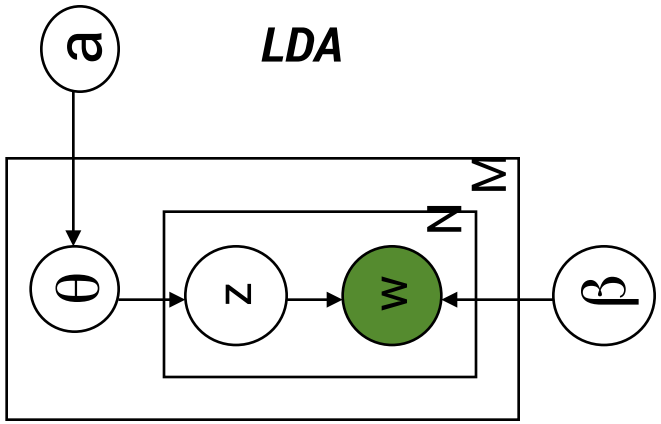
	Category 1	…	Category Restaurant	…	Category Airport	…	Category K
User 1	0	0	2	0	1	0	0
…	*	*	*	*	*	*	*
User N	*	*	*	*	*	*	*

Each user is represented as distribution among 764 venue categories

Location Data Representation: Location Topics

1. Map all Foursquare check – ins to Foursquare venue categories from category hierarchy.
2. Form user – related documents, containing venue categories of every check-in
3. Apply LDA on it represent as distribution among n latent topics, where Users – documents, words – Foursquare venue categories



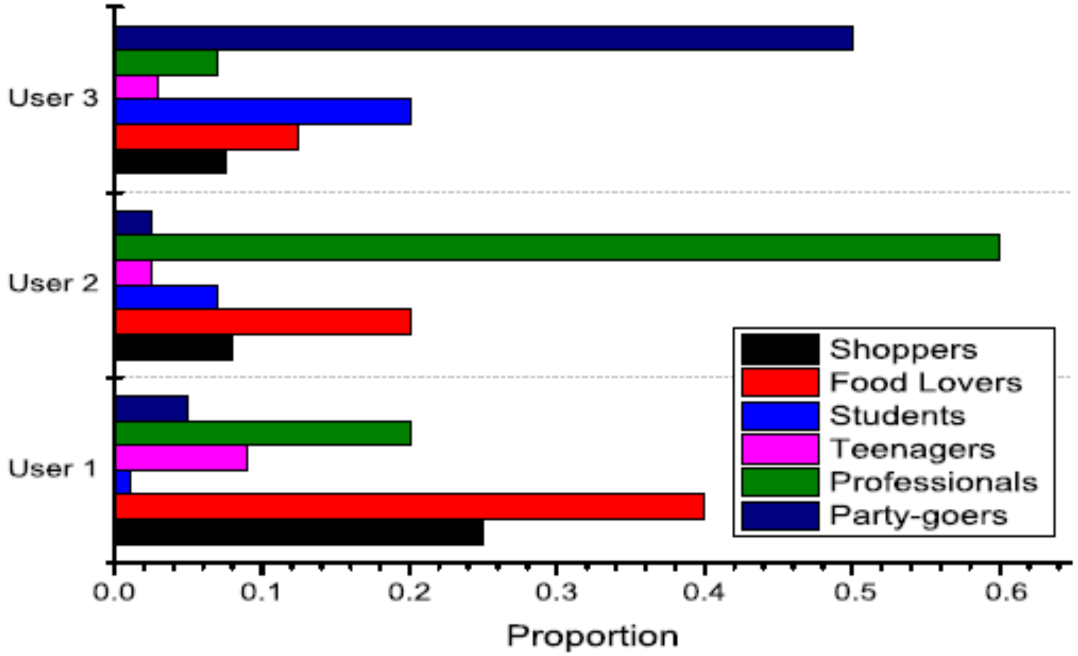


	Topic 1	Topic 2	Topic 3	...	Topic N
User 1	0.05	0.4	0.1	0.35	0.1
...	*	*	*	*	*
User 2	*	*	*	*	*

Location Data Representation: Location Topics (2)

Category distribution among LDA topics

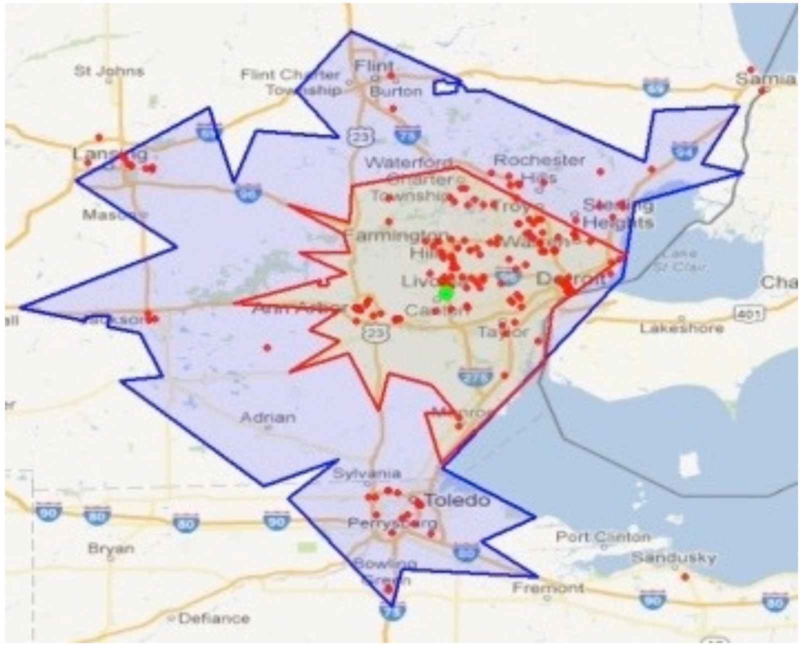
ID	Categories	LDA Topics
T1	Malay Res-t, Mall, University, Indian Res-t, Aisian Res-t	Food Lovers
T2	Cafe, Airport, Hotel, Coffee Shop, Chinese Res-t	Travelers (Business)
T3	Nightclub, Mall, Food Court, Trade School, Res-t, Coffee Shop	Party Goers
T4	Home, Office, Build., Neighbor-d, Gov. Build., Factory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



LDA word distribution over 6 topics for collected Foursquare check-ins.
Every venue category is considered as a word, each Foursquare user - as a document

Location Data Representation: GEO

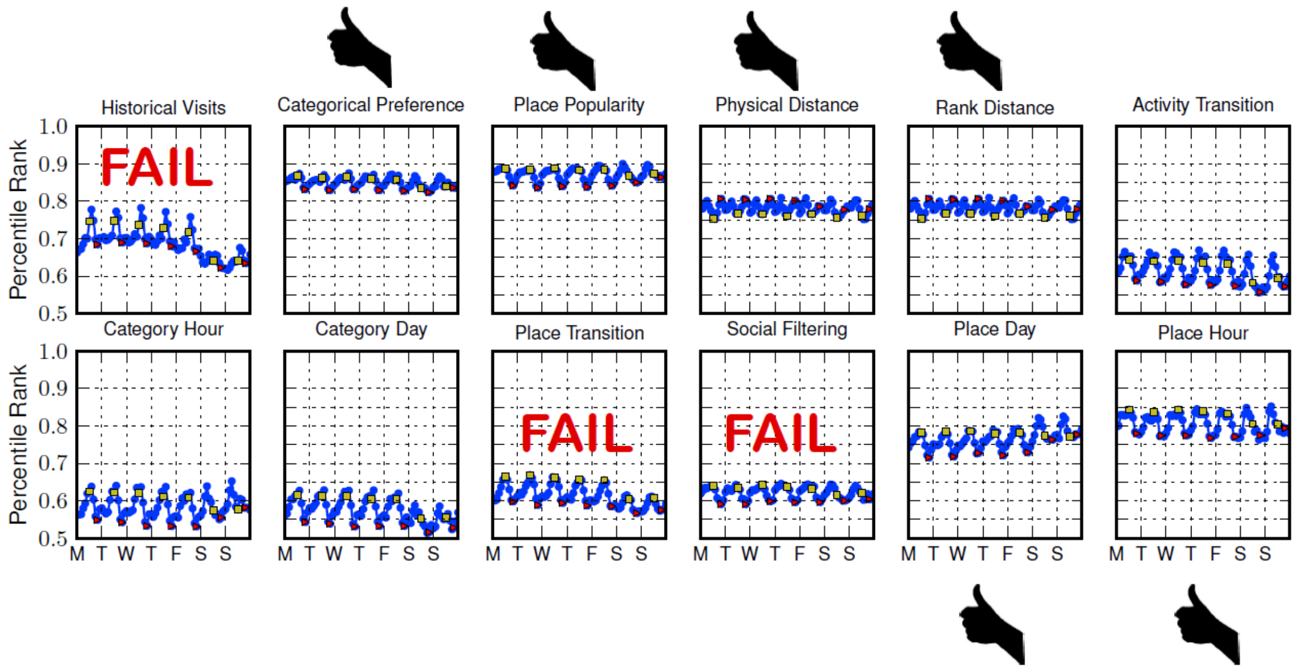
- # of posts in 8 daytime intervals
- # of AOIs
- median AOI size
- # of AOI outliers
- median distance between AOIs



User's Area of Interests (AoI) – area the most frequently visited by the user. *In fact, it is the convex hull over the dense user's check-in region (DB-Scan cluster).*

*Qu, Y., & Zhang, J. (2013, May). **Trade area analysis using user generated mobile location data**. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1053-1064). International World Wide Web Conferences Steering Committee

Location Data Representation: Temporal



Effectiveness of each feature over time changes.

- Predictions are more accurate at noon than in the evening
- Predictions for Physical & Rank distances reverse -- users cover shorter distance at night
- Predictions for Historical Visits & Place Transition drop significantly over weekends
- whereas Categorical Preference, Place Popularity & distance based features are more stable

*A Noulas S Scellato, N Lathia & C Mascolo (2012). Mining User Mobility Features for Next Place Prediction in Location-based Services. IEEE Int'l Conf. on Data Mining, 2012.

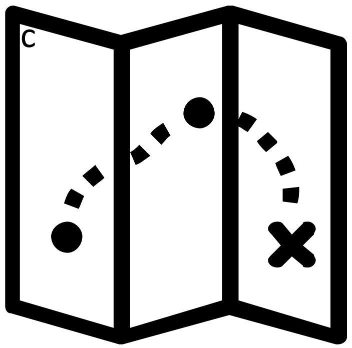
Sensor Data

Reference

*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.

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Sensor Data Representation: Utilizing Exercise Semantics

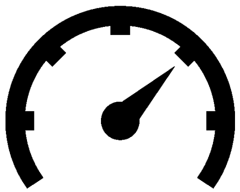


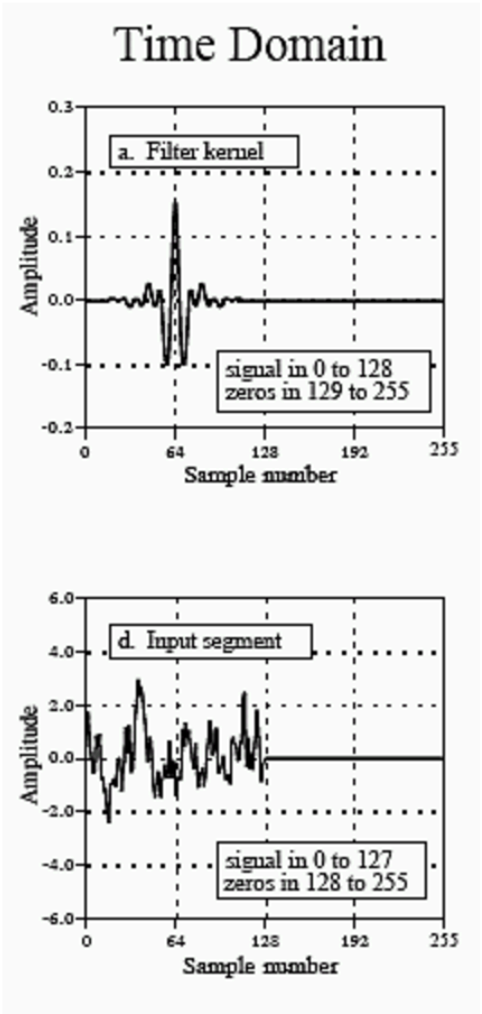
For case when user performed two swimming and one running workout

	$Category_1$	…	$Category_{swimming}$	…	$Category_{running}$	…	$Category_n$
U_1	0	0	2	0	1	0	0
…	*	*	*	*	*	*	*
U_n	*	*	*	*	*	*	*

Each user is represented as distribution among 95 workout categories

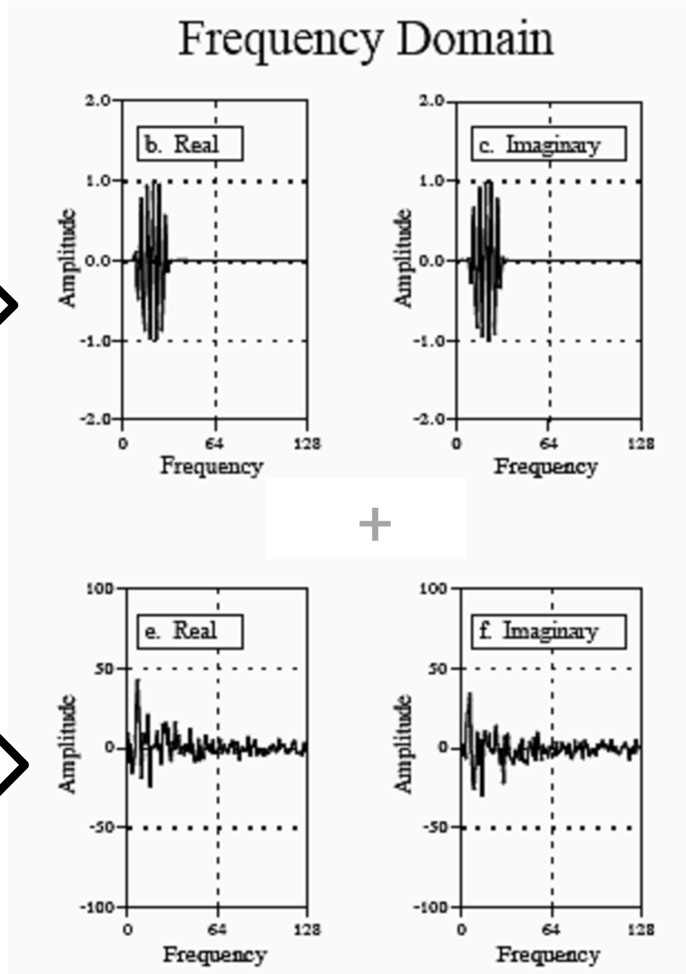
Sensor Data Representation: Frequency Spectra



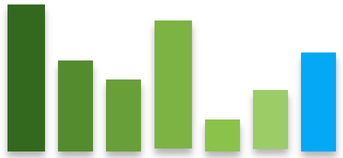


FFT

FFT

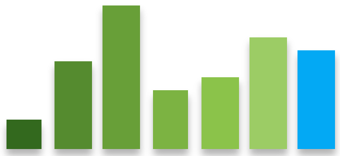


99 bins



Each user is represented as distribution among ~500 frequency bins

99 bins



Sensor Data Representation: Statistics

- # workouts
- Avg. heart rate (HR)
- Avg. ascent/descend
- HR
- Age / Gender
- Oxygen consumption

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Data Representation: Summary

All data types together

TEXT Features:

Linguistic features: LIWC; Latent Topics

Heuristic features: Writing behavior

Location Features:

Location Semantics: Venue Category Distribution

Mobility Features: Areas of Interest (AOI)

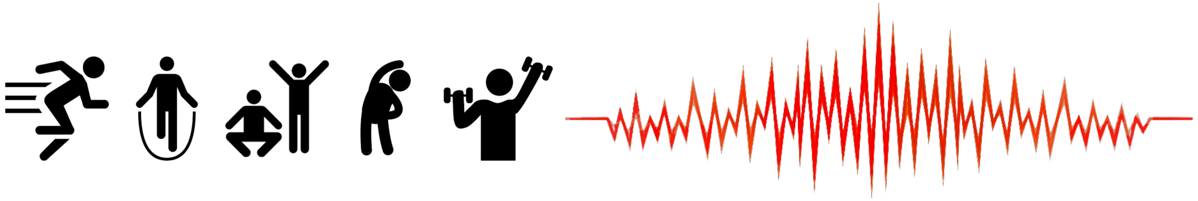
Image Features

Image Concept Distribution (Image Net)

⇨ Google Net ⇨ Image Concepts

Sensor Features

Exercise statistics + sport types + spectrum



On Data Gathering And Cross-Network Account Disambiguation

Reference

*Farseev, A., Akbari, M., Samborskii, I., & Chua, T. S. (2016). 360° user profiling: past, future, and applications by Aleksandr Farseev, Mohammad Akbari, Ivan Samborskii and Tat-Seng Chua with Martin Vesely as coordinator. ACM SIGWEB Newsletter, (Summer), 4.

E-Mail: farseev@u.nus.edu I Website: http://farseev.com

Data Gathering: Data sources

Which data sources and why?

FOURSQUARE

Largest LBSN

Instagram

Largest Personal Photo Sharing Service

twitter

Largest English-Speaking Microblog

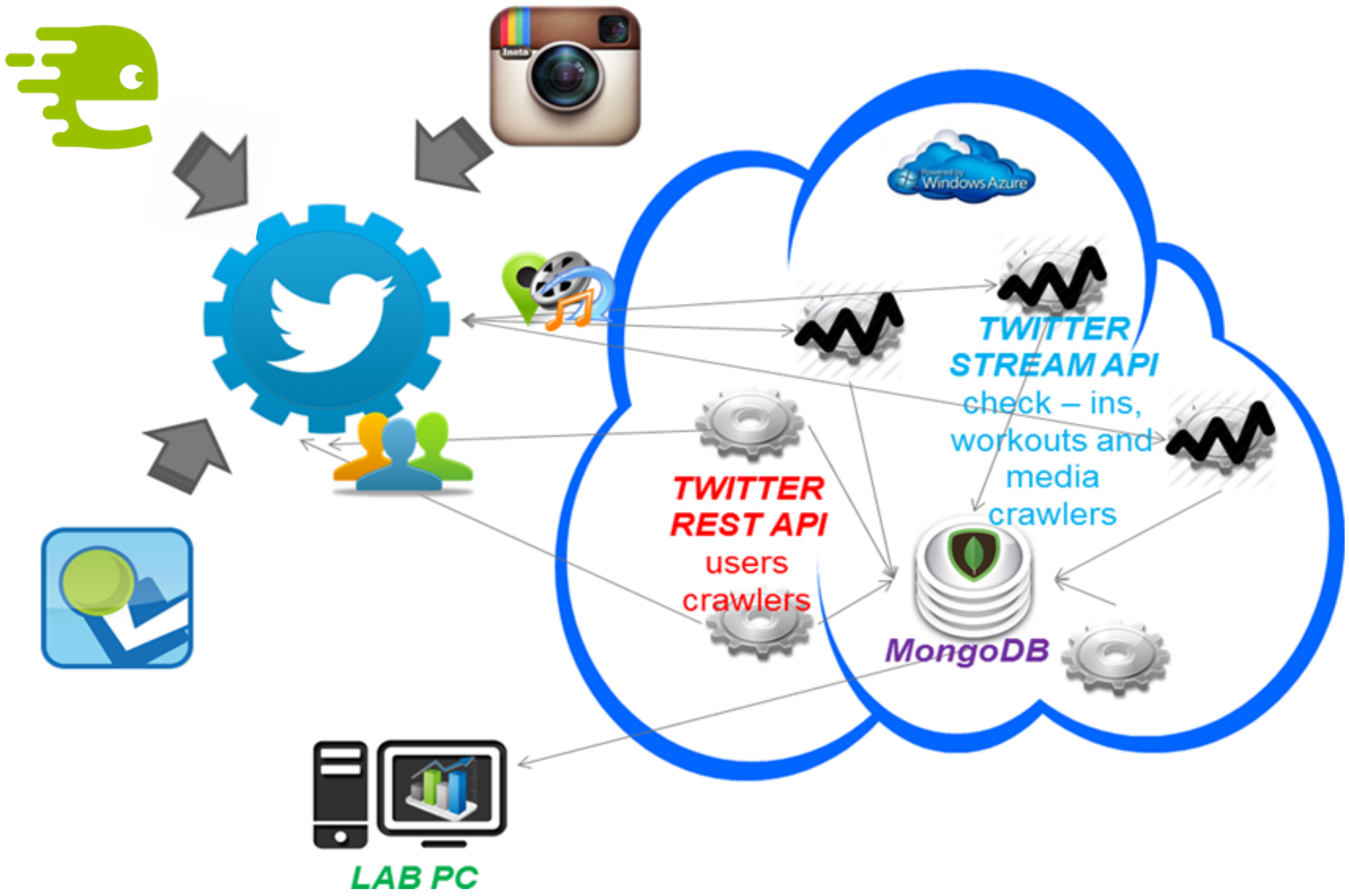
endomondo

One of the Largest Sports Tracking Network

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Data Gathering And Simultaneous Cross-Network Account Mapping

About finding the same users in different social networks…



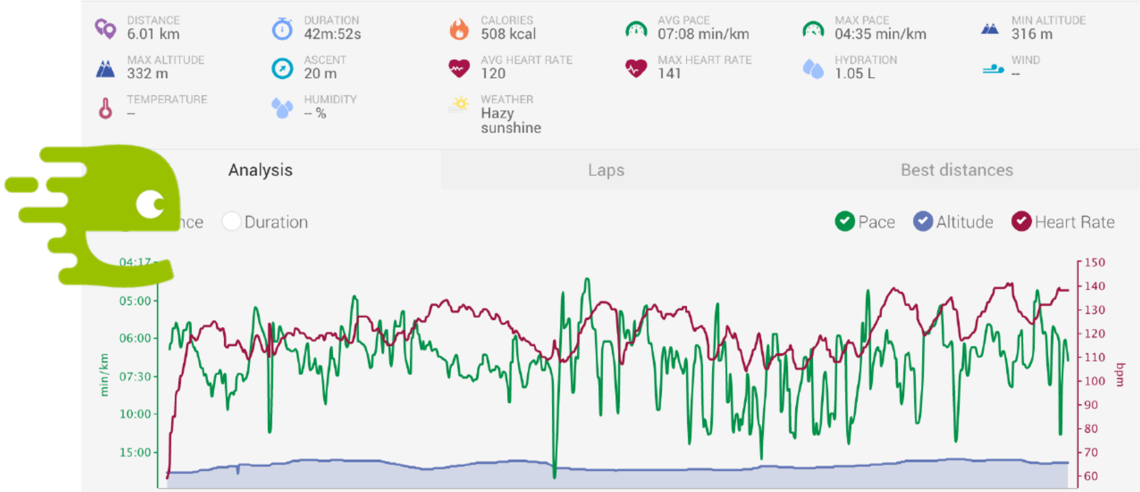
Twitter plays a role of a “sink” for multi-modal data from other social networks.

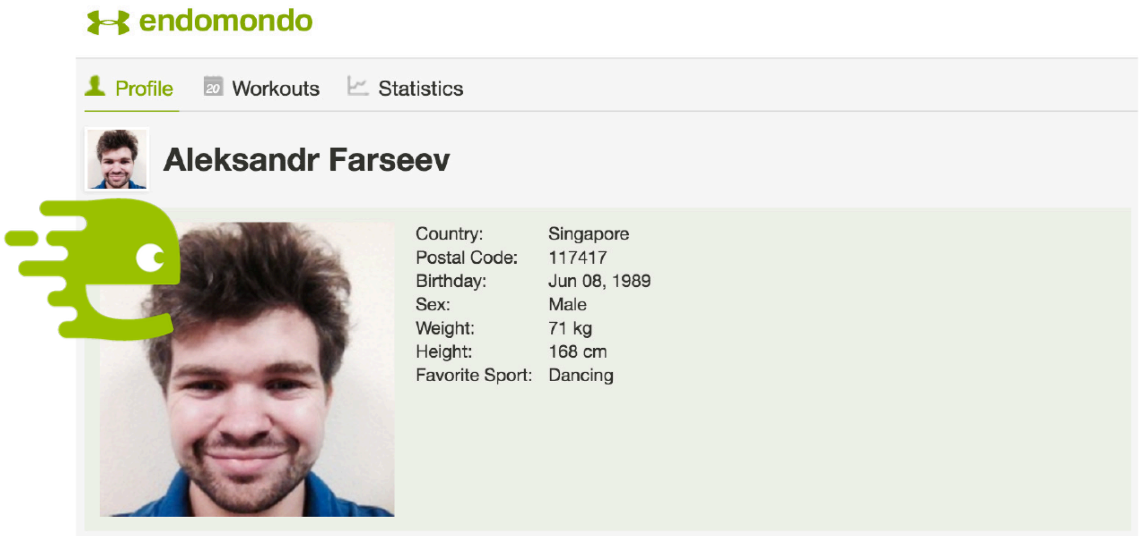
Cross-network ambiguity is resolved after first cross-network post.

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Cross-Network Account Mapping: Example

How to grab Alex's personal data…





Cross-network post

NExT

NUS National University of Singapore

360° User Profiling

How to learn from multi-source multi-modal data.

by Aleksandr Farseev

nus.academia.edu/farseev

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Agenda

Brief summary of the talk…

1 Motivation and Data

1. Definition and Motivation
2. NUS-MSS Dataset
3. NUS-SENSE Dataset
4. Framework Overview

2 Individual User Profiling

1. Multi-modal data representation
2. Multi-Source Ensemble Learning for Demographic Profiling
3. Learning From Temporal Data for Personality Profiling
4. Sensor data representation
5. Multi-Task Learning for Wellness Profiling

3 Group User Profiling

1. Multi-modal data representation
2. Clustering on Multi-Layer Graphs for Group Profiling
3. Cross-Domain Recommendation as Implicit Evaluation Approach
4. Real-World Application: bBridge Analytics Platform

4 Conclusion

1. Conclusion
2. Future Work

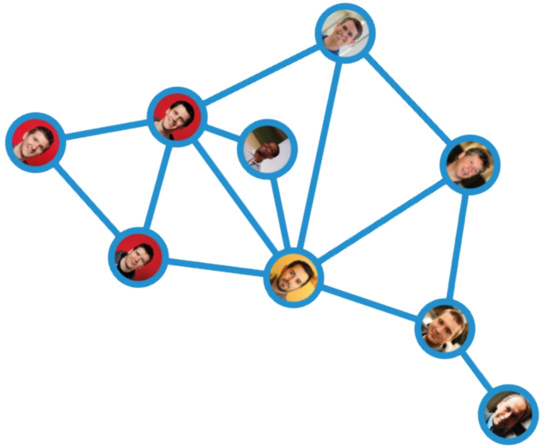
Let's … Trump

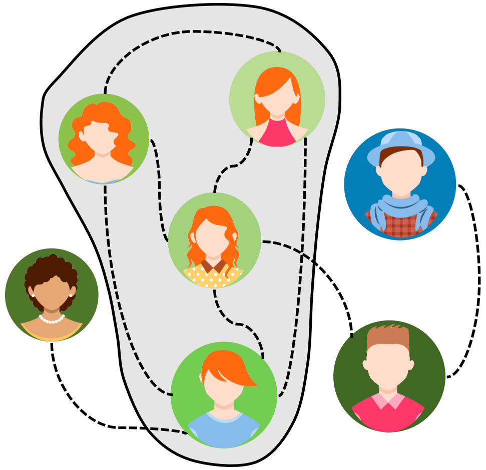
How …

> 60 Articles from 16 Countries

What is user profile?

Well, it may mean many things…





E-Mail: farseev@u.nus.edu I Website: http://farseev.com

Multiple social networks describe users from multiple views

Some facts about social networks…

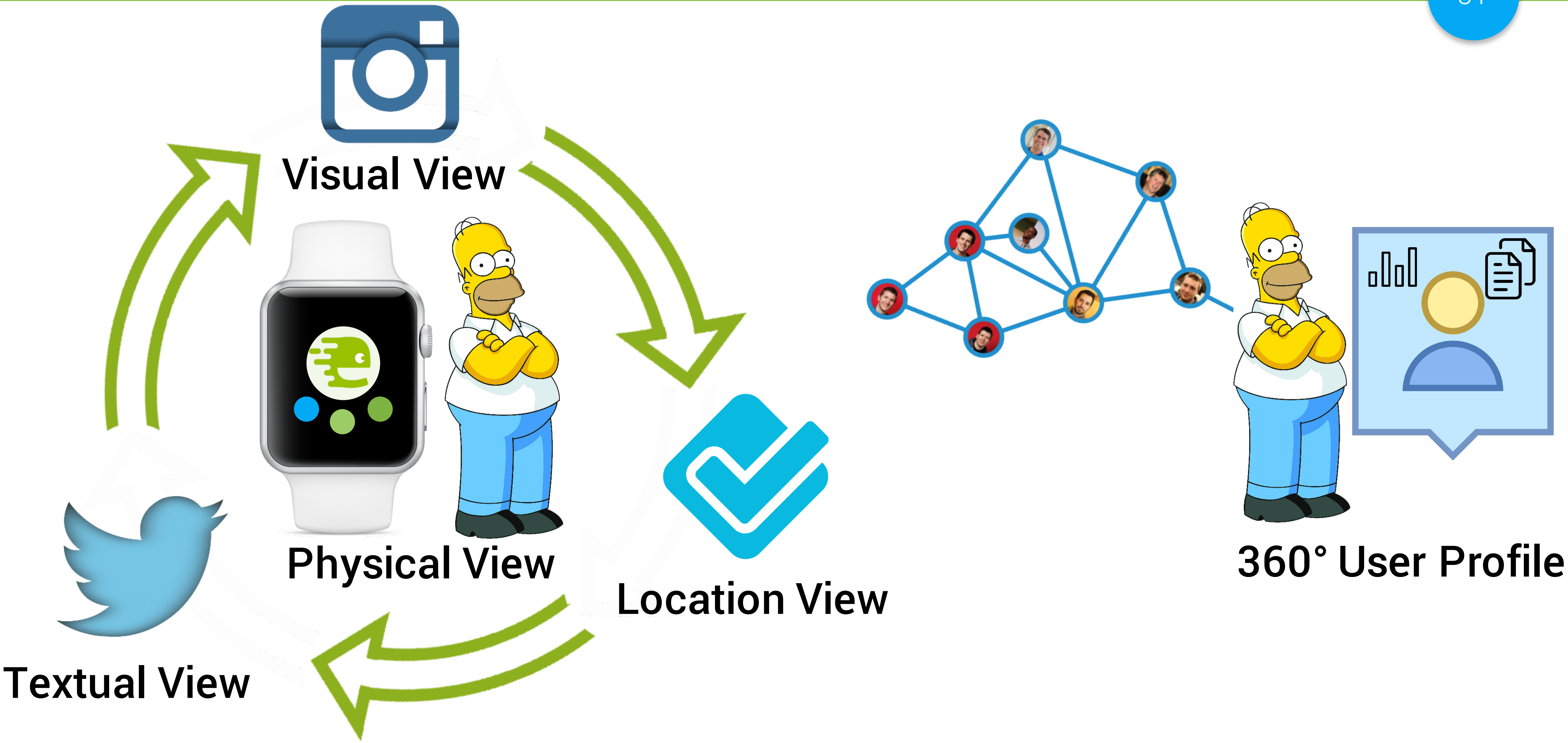
More than 50% of online-active adults use more than three social networks in their daily life*

*According Paw Research Internet Project's Social Media Update 2014 (www.pewinternet.org/fact-sheets/social-networking-fact-sheet/)

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

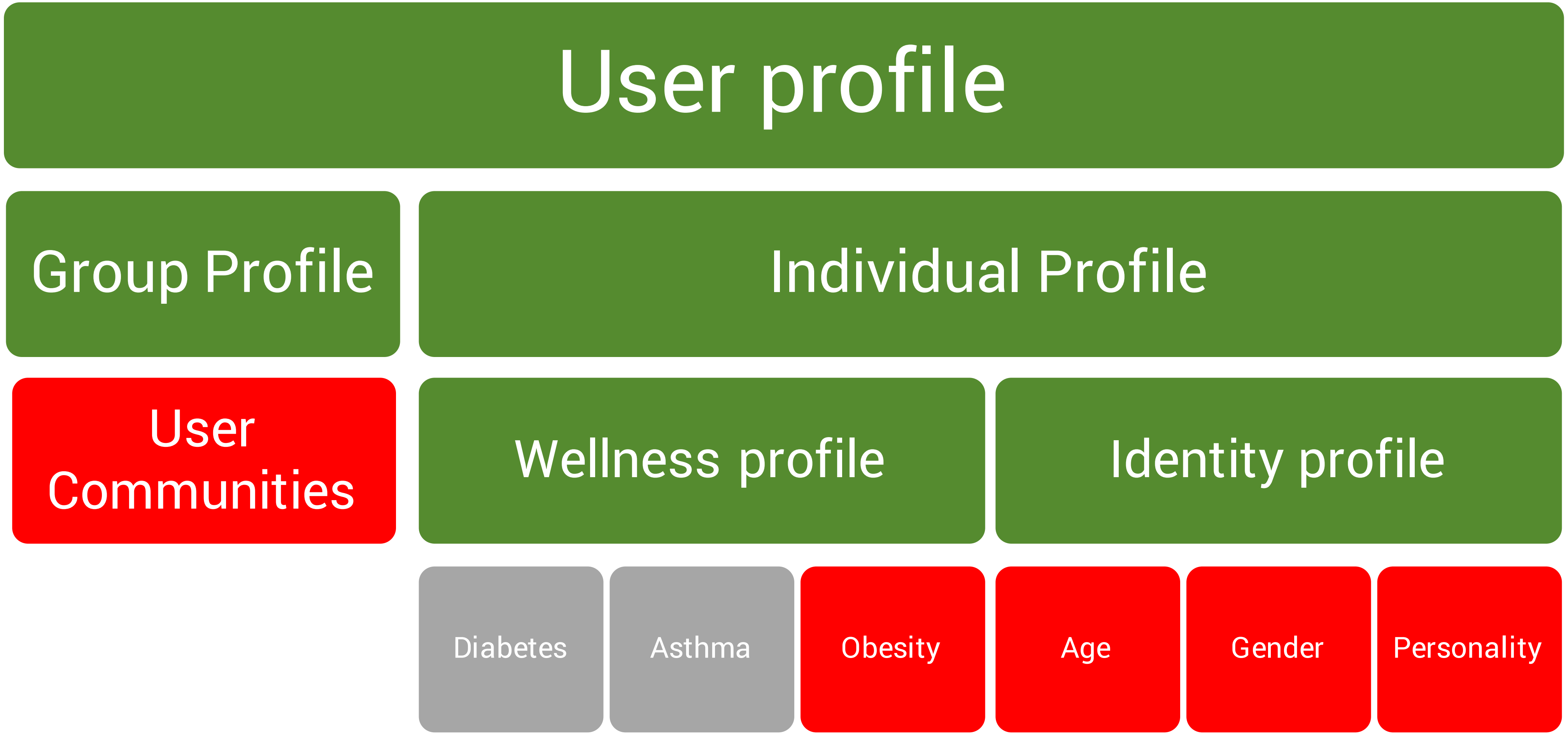
Different data modalities describe users from multiple views

Indeed, they are:



User profiling in our works

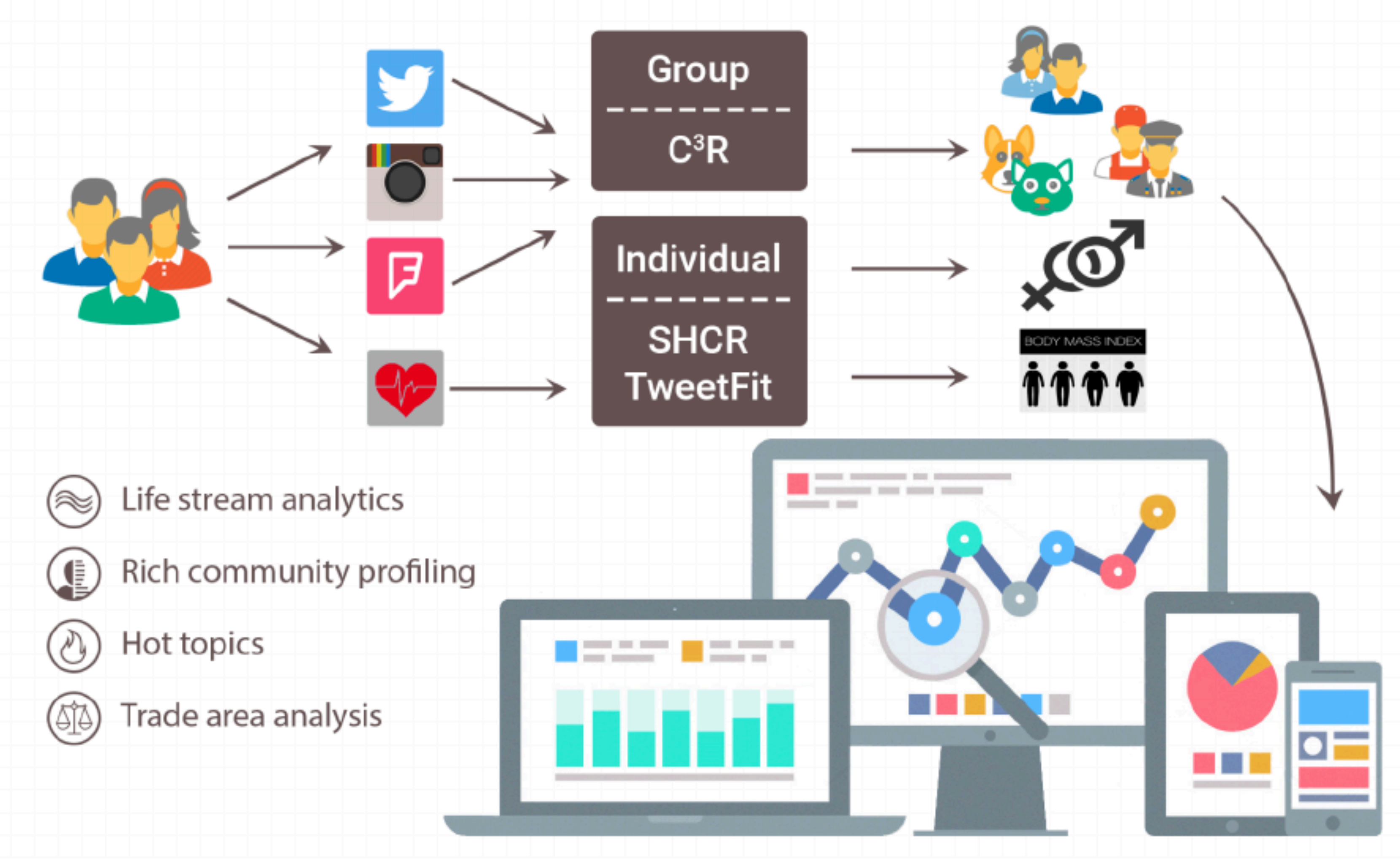
Those attributes that we infer



E-Mail: farseev@u.nus.edu | Website: http://farseev.com

360° User Profile Learning Framework Overview

Multi-Source User Profiling And Its Applications in One Framework....



1) Individual And Group User Profiling

2) Applications

Data for User Profiling

*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.

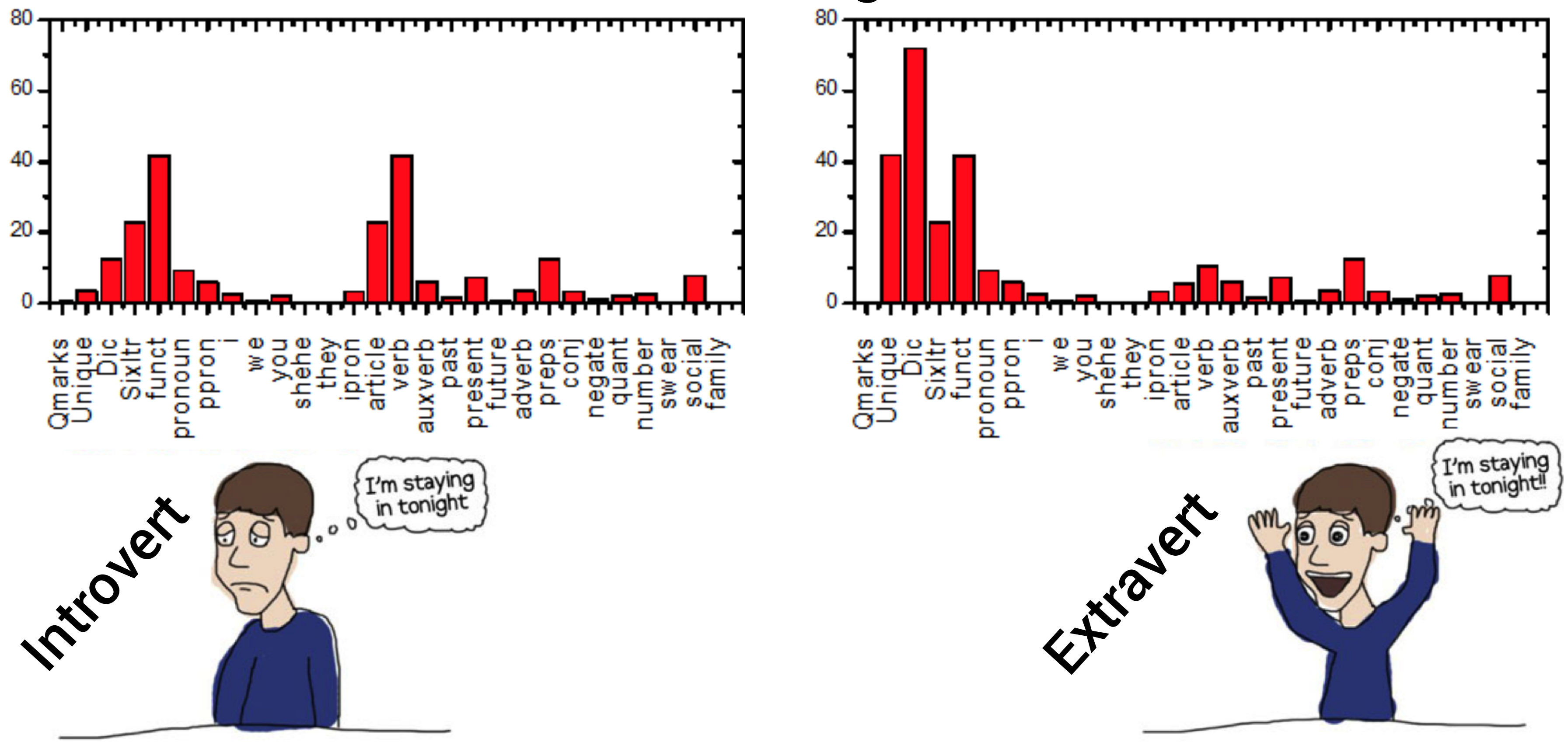
The following is relevant to:

- NUS-MSS: http://nusmultisource.azurewebsites.net
- NUS-SENSE: http://nussense.azurewebsites.net
- NUS-PERSONALITY: Dropbox

Textual data representation (1): LIWC

Pennebaker's LIWC Descriptors

69 LIWC Categories



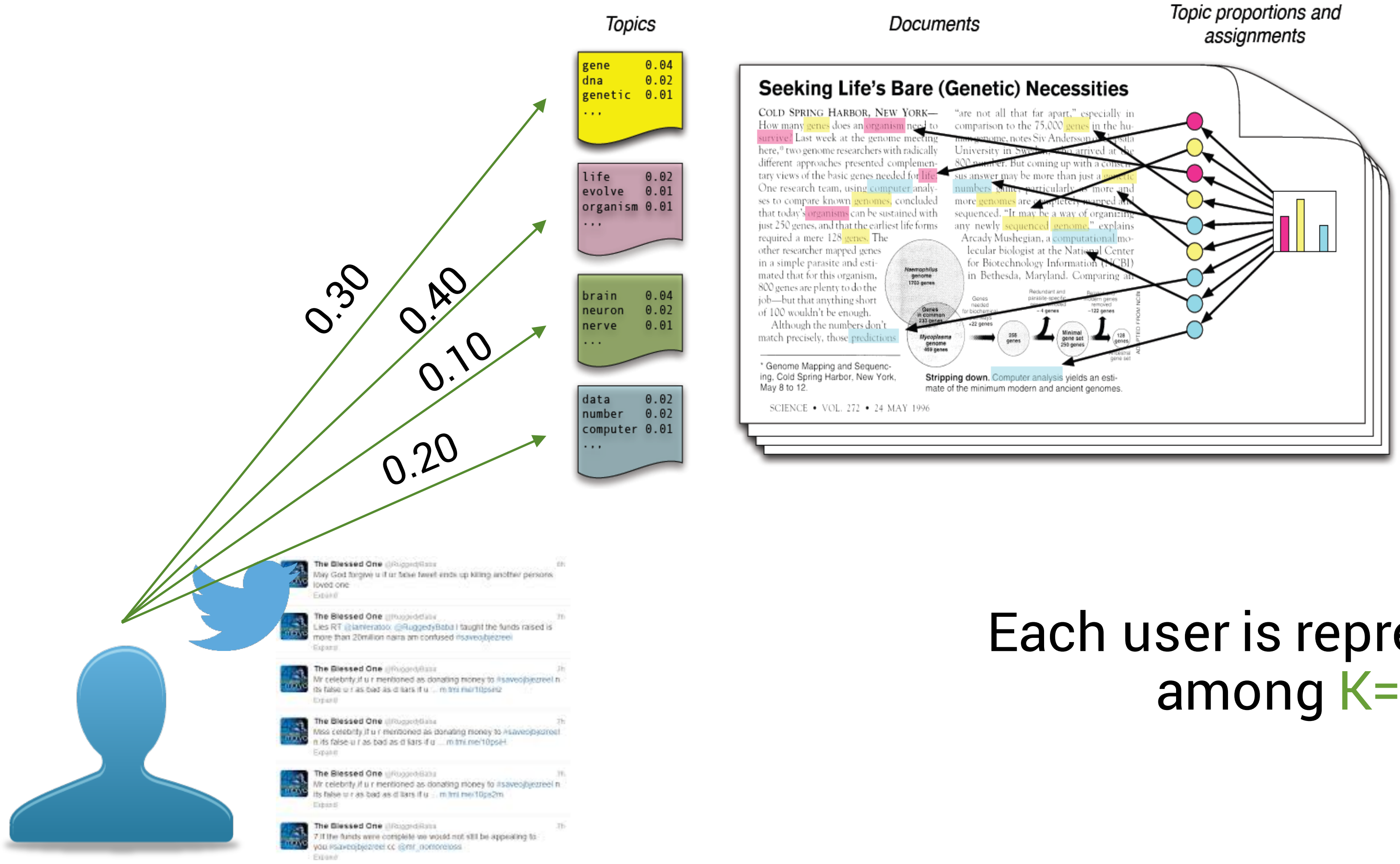
Each user is represented as distribution among K=69 LIWC categories

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

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Textual data representation (2): Latent Topic Modeling

Latent Dirichlet Allocation



Each user is represented as distribution among K=50 latent topics

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Textual data representation (3): Writing style features

Heuristically inferred and found to be useful…

- # hashtags
- # slang words
- # URLs
- # User mentions
- # repeated chars
- # emotion words
- # emoticons
- Avg. sentiment level
- Avg. sentiment score
- # Misspellings
- # Unknown words
- Avg. # terms

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Image Data Representation: Image Concept Detection

Going deeper with convolutions with Google Net.

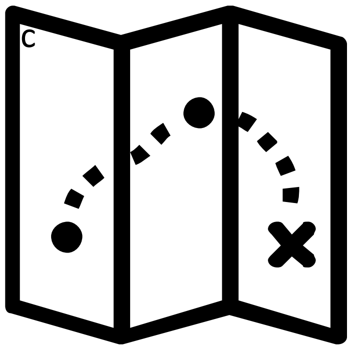
Each user is represented as distribution among K=1000 image concepts

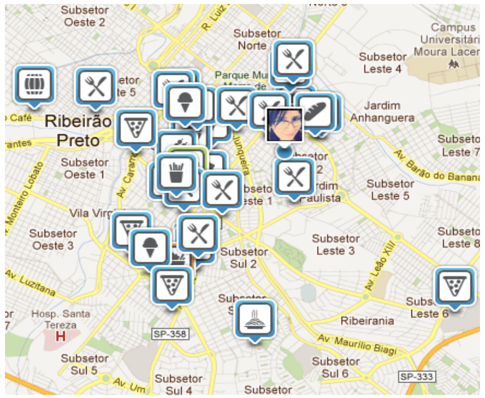
*Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-9).

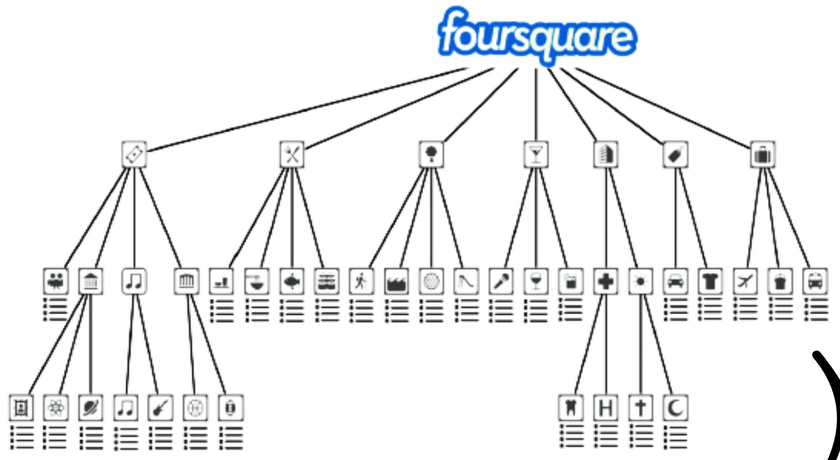
E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Location Data Representation (1): Utilizing Venue Semantics

Venues are not just GEO points, but multidimensional objects with rich semantics







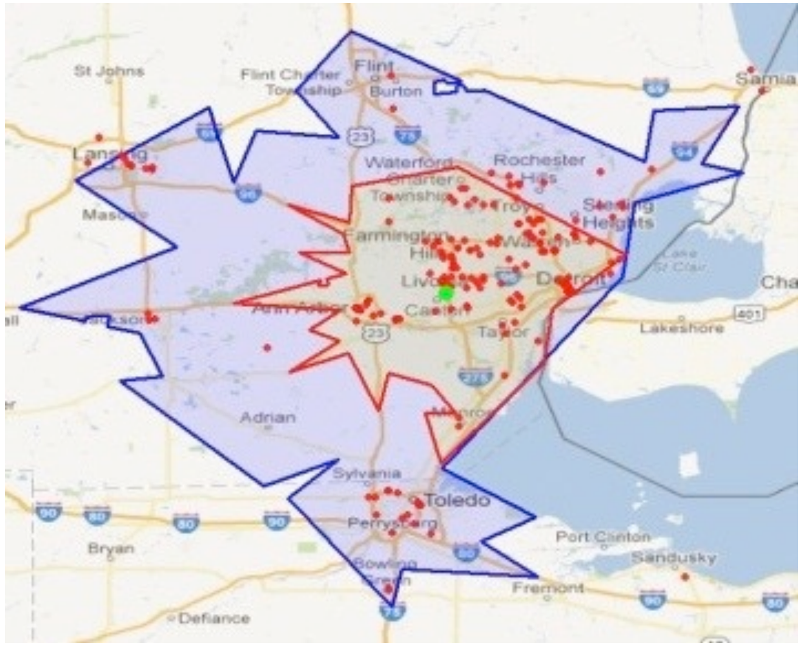
Each user is represented as distribution among 764 venue categories

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 K
User 1	0	0	2	0	1	0	0
…	*	*	*	*	*	*	*
User N	*	*	*	*	*	*	*

Location Data Representation (2): Mobility Features

- # of posts in 8 daytime intervals
- # of AOIs
- median AOI size
- # of AOI outliers
- median distance between AOIs



User's Area of Interests (AoI) – area the most frequently visited by the user. *In fact, it is the convex hull over the dense user's check-in region (DB-Scan cluster).*

*Qu, Y., & Zhang, J. (2013, May). **Trade area analysis using user generated mobile location data**. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1053-1064). International World Wide Web Conferences Steering Committee

Sensor Data Representation

Heuristically inferred and found to be useful...

- # workouts
- Avg. heart rate (HR)
- Avg. ascent/descend
- Age / Gender
- HR
- Oxygen consumption

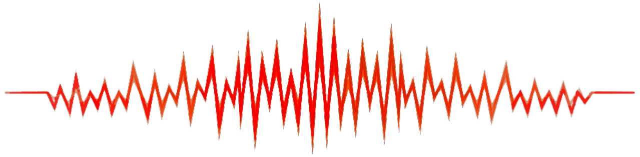
Each user is represented as distribution among 500 + 95 + 6 = 601 sensor features

Distribution over 95 sport categories

Furrier's Spectrum

500 frequency bins

$$F_{Max} = 0.5\ dHz$$



E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Data Representation: Summary

All data types together

TEXT Features:

Linguistic features: LIWC; Latent Topics

Heuristic features: Writing behavior

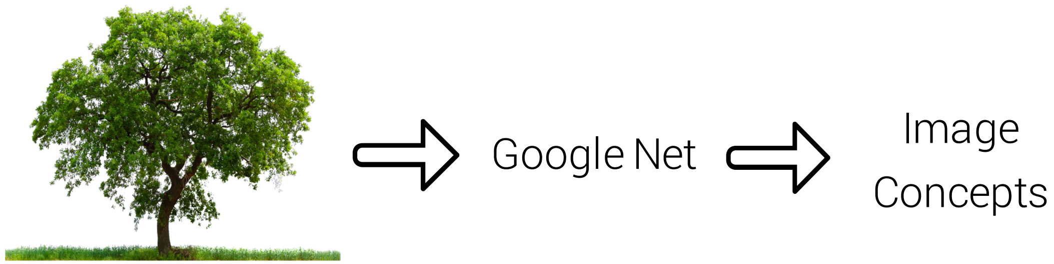
Location Features:

Location Semantics: Venue Category Distribution

Mobility Features: Areas of Interest (AOI)

Image Features

Image Concept Distribution (Image Net)



Sensor Features

Exercise statistics + sport types + spectrum

Lunch Break

60 Minutes

“

The demographics of who’s on what social network are shifting — older social networks are reaching maturity, while newer social messaging apps are gaining younger users fast…

- Thiago Guimaraes

Individual User Profiling

Part I: Demographic Profiling (NUS-MSS)

*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.

On importance of basic demographic attributes

What we can do if we know Homer's age?

Age: 40
Gender: 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.

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Related Research Directions: Individual User Profiling

Conventional Data Sources

Manually-collected small-scale datasets: voice records, hand-written texts, etc.

Demographic Profiling

PAN 13', 14', 15', 16', 17'
Mostly mono-modal data processing (Text from Microblogs, Blogs, etc.)

Other Profiling Domains

Mostly mono-modal mono-source data processing.

Multi-source Profiling

Utilize multiple sources, but usually require all sources to be complete, which is unrealistic.

Sensor data utilization

Mostly mono-source and solve tasks as human activity recognition, which is not directly related.

Physical Attributes Inference

Mostly single-modal. Limited number of contributions to the field..

Usually, this problem of individual user profiling is treated as a supervised learning task and performed based on mono-source mono-modal datasets. Research on multi-source data learning is relatively sparse and limited by unrealistic assumptions. Wellness attributes inference was not yet comprehensively studied.

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Idea I

From what it all started…

Input One

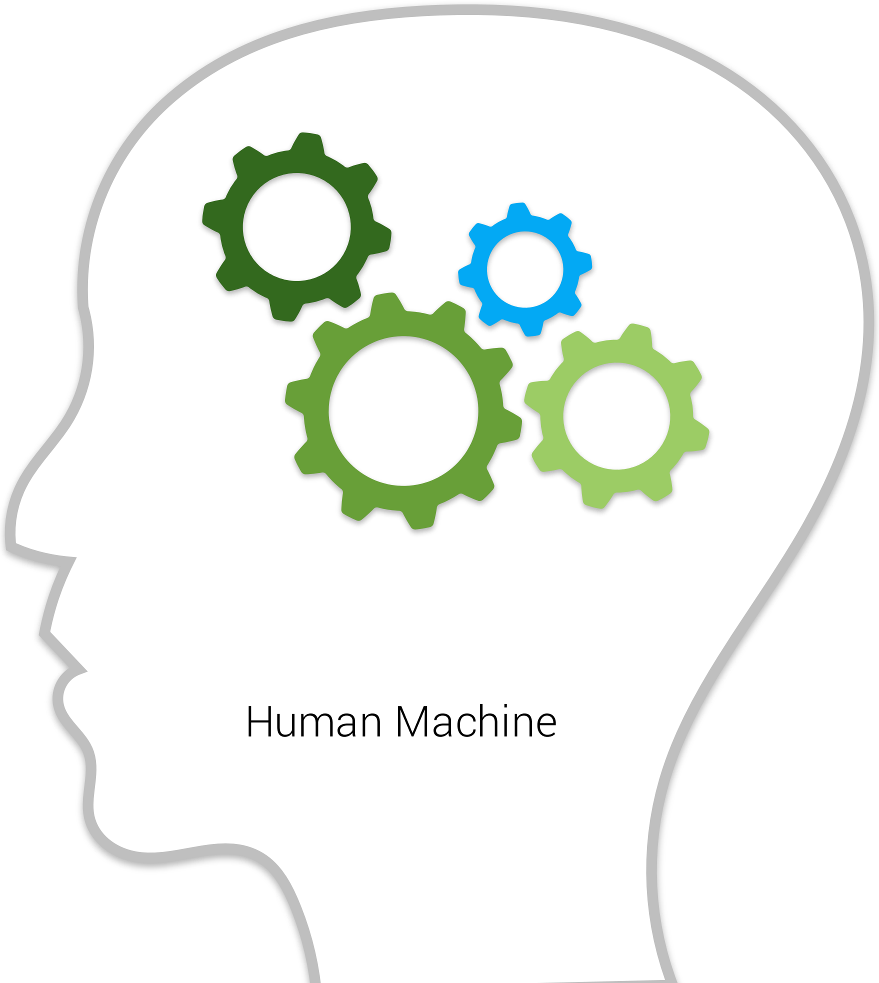
1

Multi-Source Multi-Modal Data Describes Users From Multiple Views

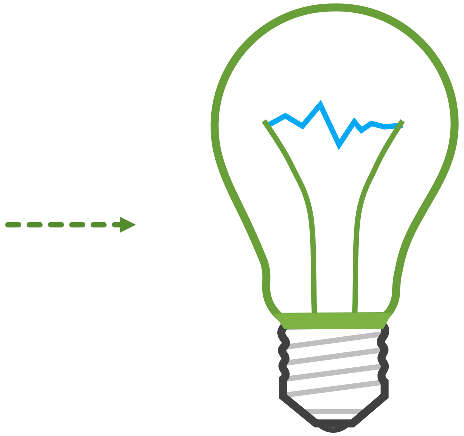
Second Input

2

Demographic Profiling is possible Based On Textual Data



Human Machine

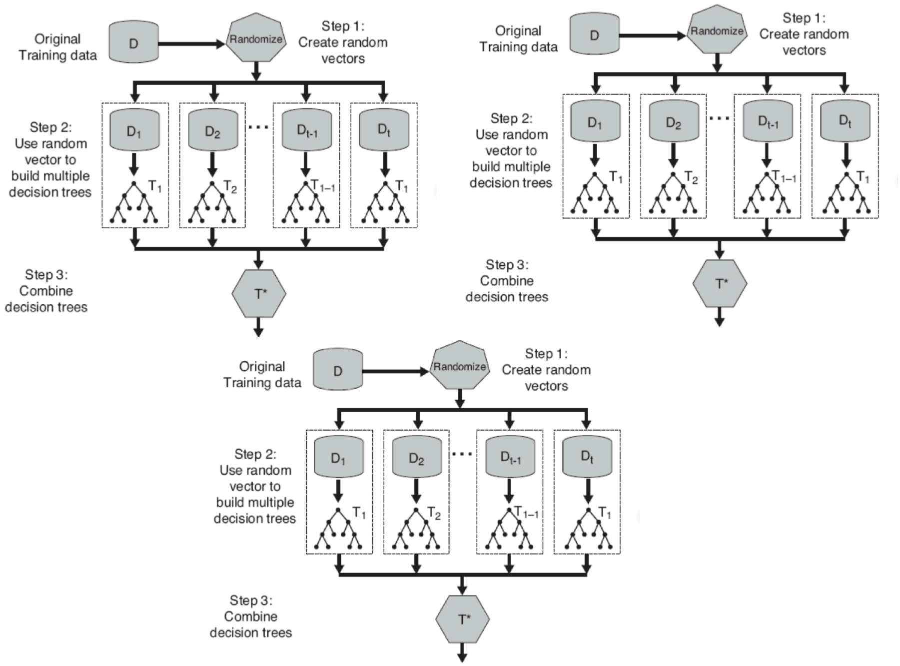


Idea

Perform Demographic Profiling Based On Multiple Social Networks

Age and Gender Prediction

Running Random Forests With Random Restart



$d_i \times l_i$

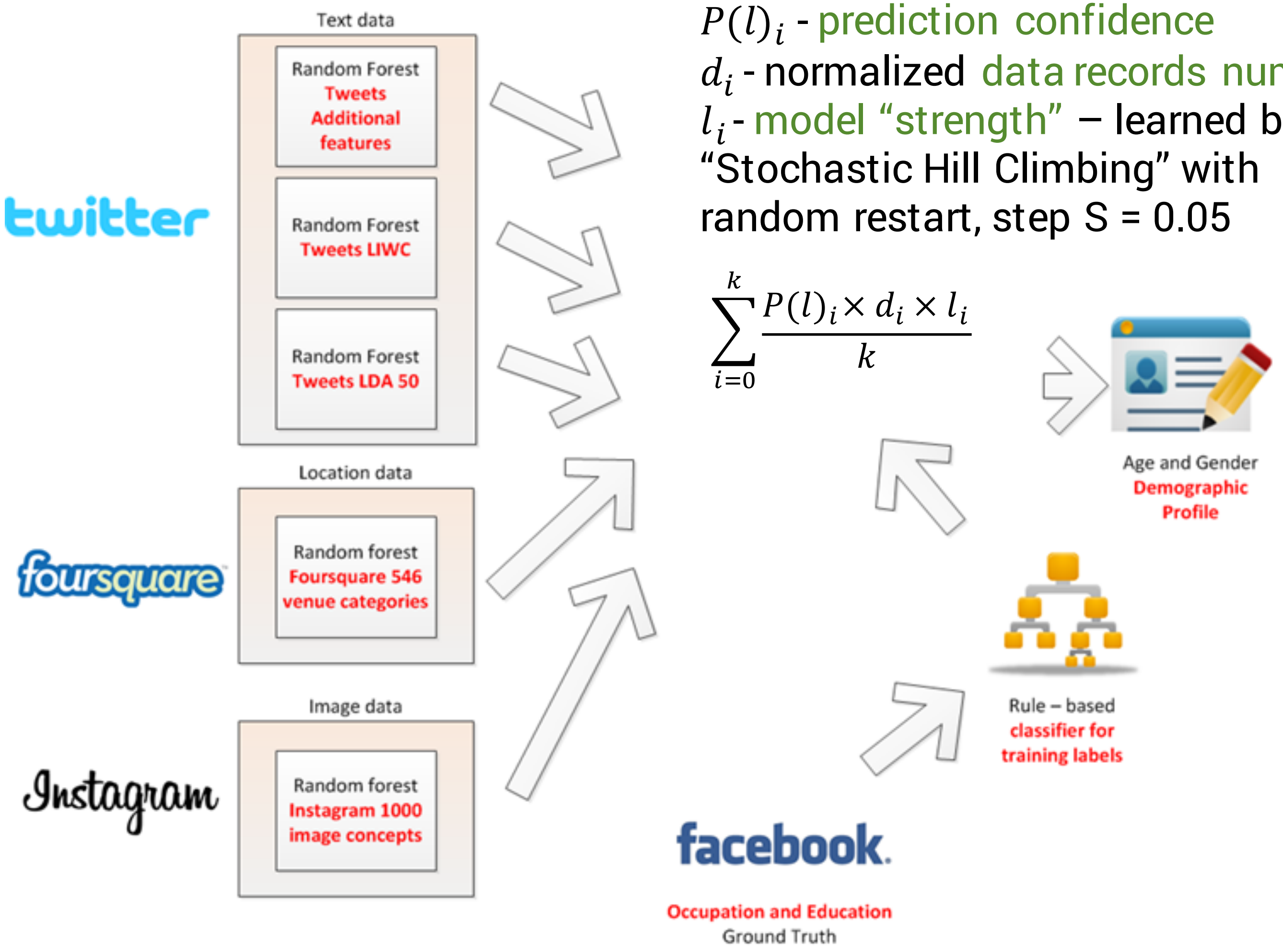
Age and Gender Confidence Score $P(l)_i$

$P(l)_i$ - prediction confidence

d_i - normalized data records number

l_i - model "strength" – learned by "Stochastic Hill Climbing" with random restart, step S = 0.05

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



Age and Gender Prediction: Details

Some details that you may find in the paper

- 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 and 1000 random restarts.

*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.

Age and Gender Prediction: Results

About The Power Of Multiple Sources...

Data Source Combinations

Method	Gender	Age
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**

Other Baselines

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

4 Age Groups: <20; 20-30; 30-40; >40
2 Genders: Male; Female

*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.

Contributions…

One of The First Works
On Multi-Source
Multi-Modal
Individual User Profiling

01

First Large-Scale Multi-Source
Cross-Modal Cross-Region Dataset
NUS-MSS

02

*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.

The most important kind of freedom is to be what you really are.

- Jim Morrisson

Individual User Profiling

Part II: Personality Profiling (NUS-PERSONALITY)

Idea II

From what it all started…

1 Input One

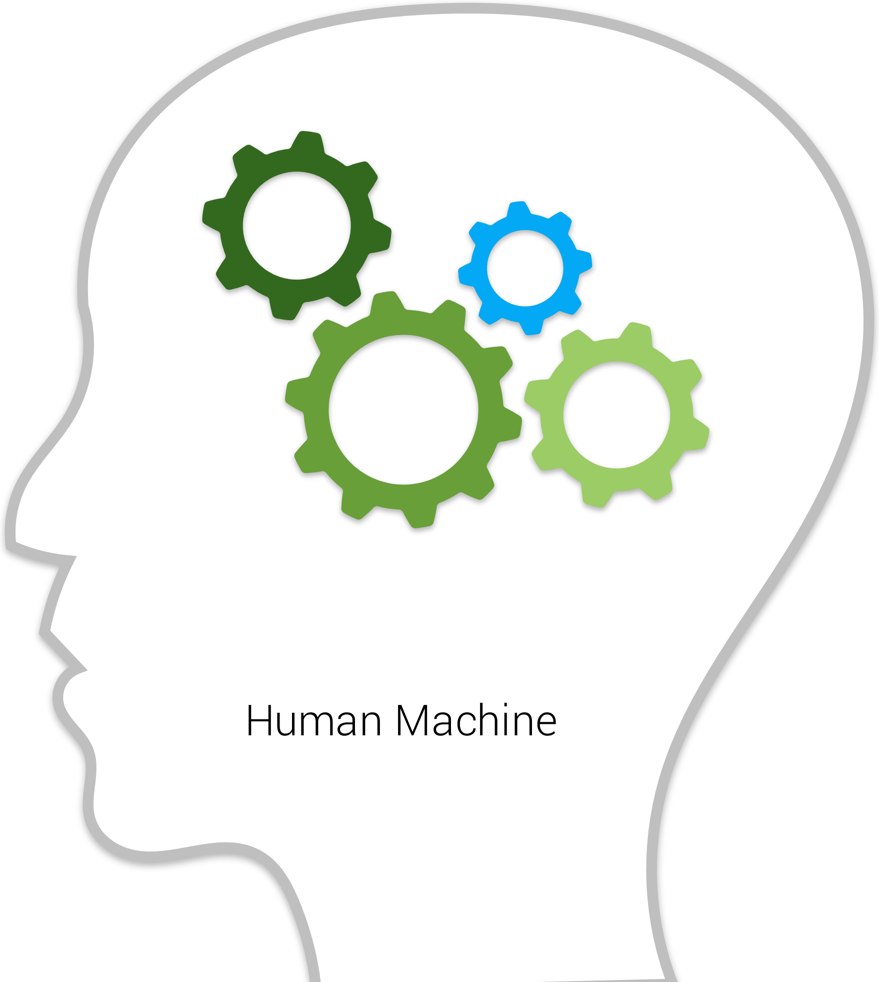
Demographic Profiling is possible Based On Multi-Source Data

2 Second Input

Demographic Profiling is possible Based On Textual Data

3 Third Input

People of Different Personality Exhibit Different Online Behavior



Idea

Perform Personality Profiling Based On Multiple Social Networks Considering Data Temporality

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

What is user personality?

Relationships

Life challenges

Personality

The powerful description an individual that affects:

Emotions

Behavior

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Mayers-Briggs Type Indicator (MBTI)

MBTI – the typology, which is designed to exhibit psychological preferences in how people perceive the world around them and distinguishes 16 personality types.

Experiments

- #1 Experiment. Test the models trained based on multisource data for different size of temporal window. Use NMF to complete missing data points.
- #2 Experiment. Test the best temporal model from Experiment 1 on different data modalities and their combinations

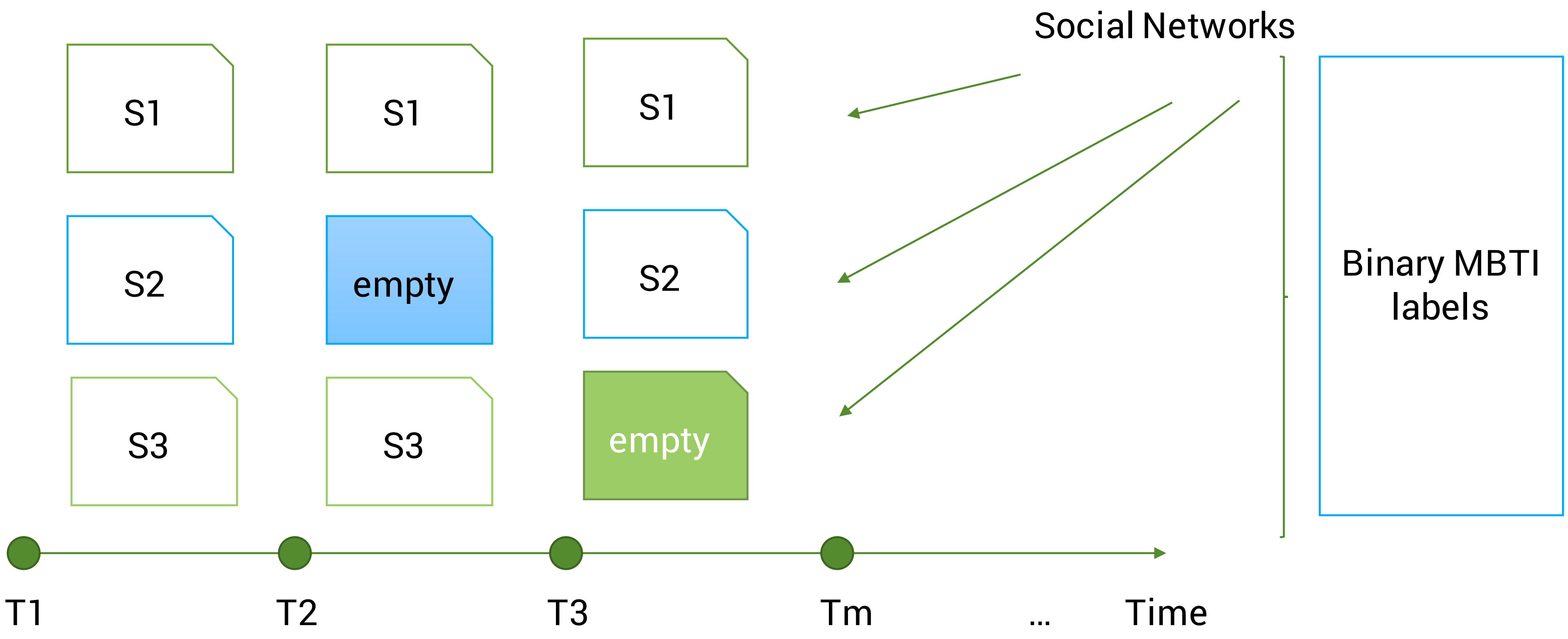
Test/Train sets:

- Divide ALL users into test(20%) and train(80%) based on the distribution of their MBTI type
- For #1 Experiment use all test data
- For #2 Experiment use only complete (with 3 social networks) users.

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

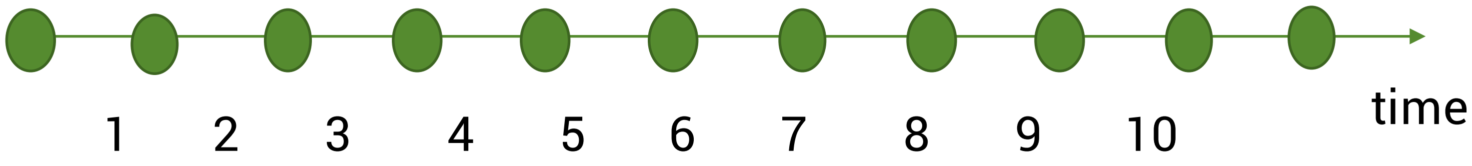
Incorporation of temporal aspect

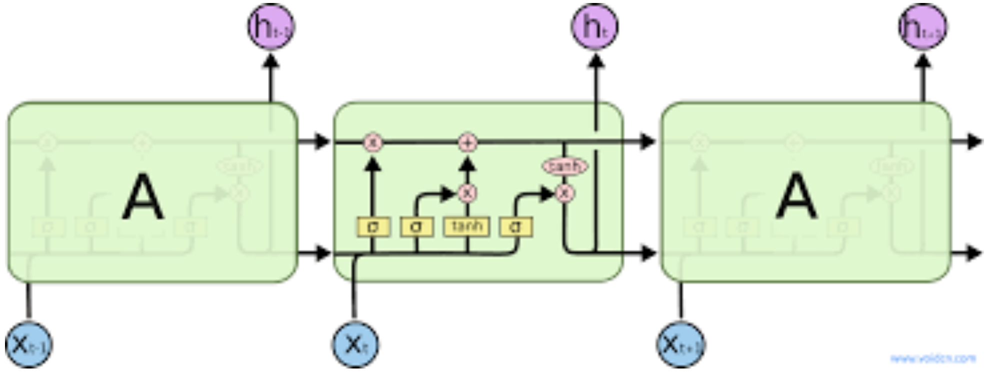
Divide user activity into 10 time periods



Dealing with missing data

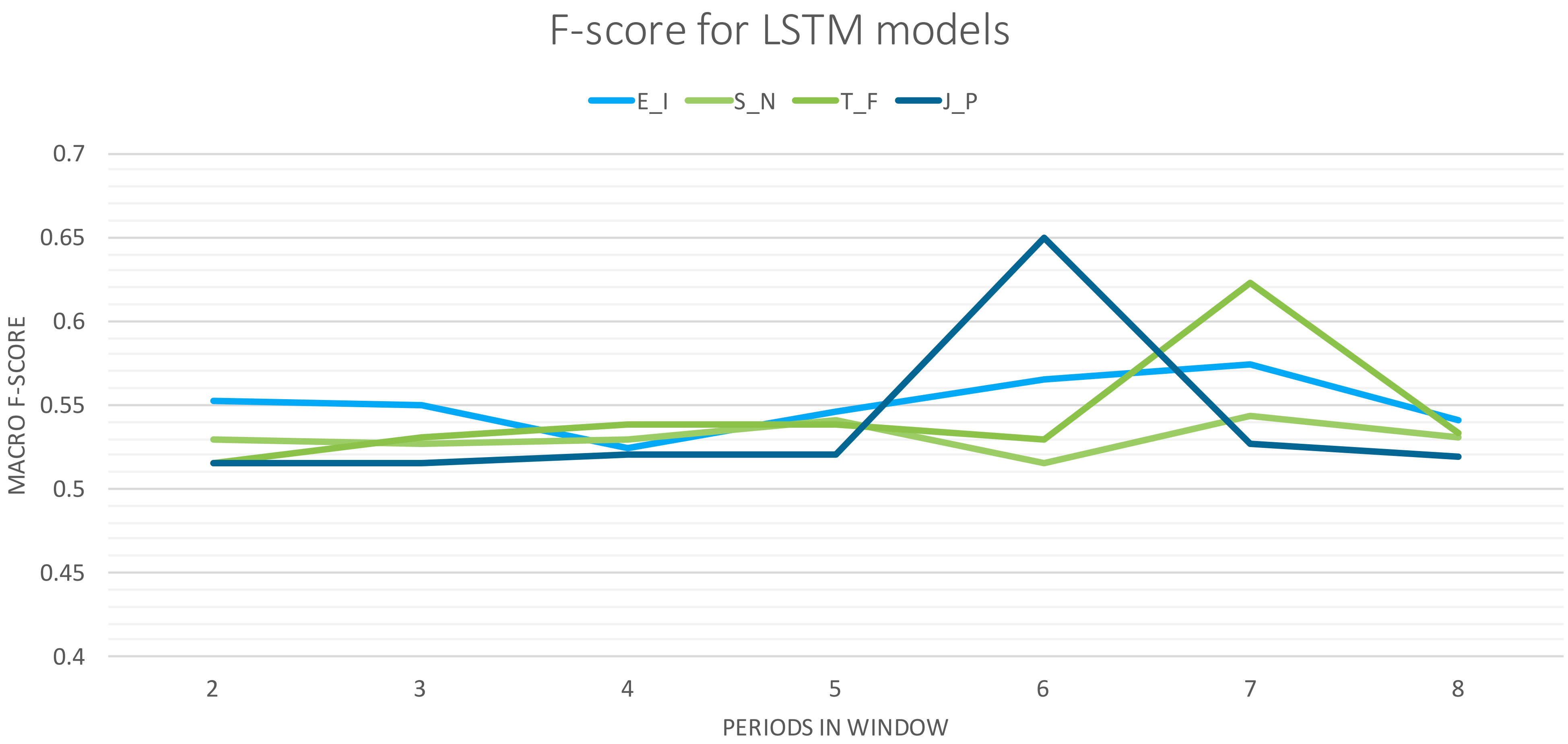
- Model: LSTM for time periods. Neural network that will take into account not only the multimodality in data, but also the temporal aspect.
- Problems: very few users with 3 social networks in each n-period window(~120)
- Solution:
  - Use non-negative matrix factorization for data completion for every k-period window





E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Results: Different Time Windows



Results: Different Data Source Combinations

	Precision	Recall	F-measure	Macro F-measure	Model
E	0,60	0,61	0,61		Gradient Boosting
I	0,64	0,62	0,62	0,62	
S	0,388	0,373	0,38		LSTM (7 intervals window)
N	0,7	0,713	0,706	0,543	
T	0,462	0,451	0,623		LSTM (7 intervals window)
F	0,613	0,634	0,623	0,623	
J	0,531	0,697	0,65		LSTM (6 intervals window)
P	0,609	0,697	0,65	0,65	

Contributions…

The First Works On Multi-Source Temporal Personality Profiling

01

On of the first works on temporal multi-source learning

02

“

Health is a state of body. Wellness is a state of being…

- James Stanford

Individual User Profiling

Part III: Wellness Profiling (NUS-SENSE)

*A. Farseev, A., & Chua, T. S. (2017). Tweetfit: Fusing multiple social media and sensor data for wellness profile learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI.

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

People are often not aware of their lifestyle problems

Why Homer often end up in Hospital

Weight Problems Consequences

It is not just about looking not fit…

Weight Problems Consequences

- All-causes of death (mortality)
- **High blood pressure (Hypertension)**
- High / Low HDL cholesterol
- **Type 2 diabetes**
- **Coronary heart disease**
- Stroke
- Gallbladder disease
- **Osteoarthritis**
- **Some cancers**
- **Mental illness such as clinical depression**
- Body pain

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

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Idea III

Going further towards realizing the ideal of 360° User Profile Learning

Input One

1

Multi-Source Individual User Profiling is possible and allows for significant demographic prediction performance boosting

Second Input

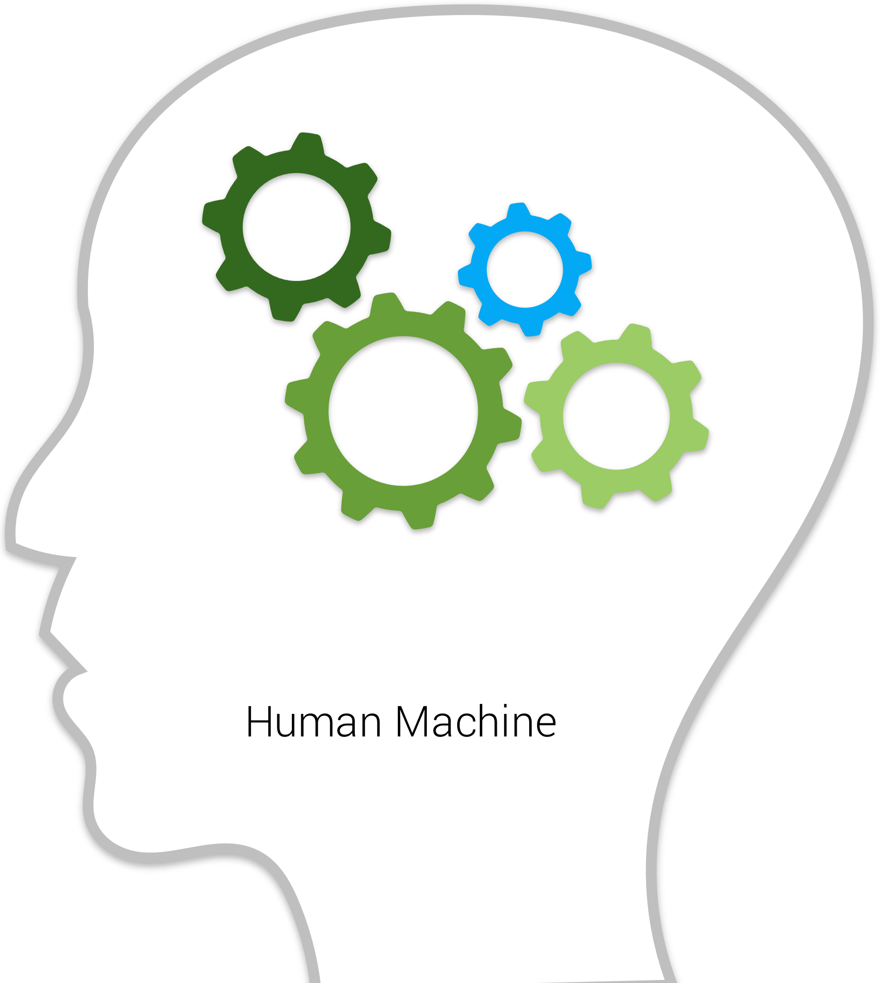
2

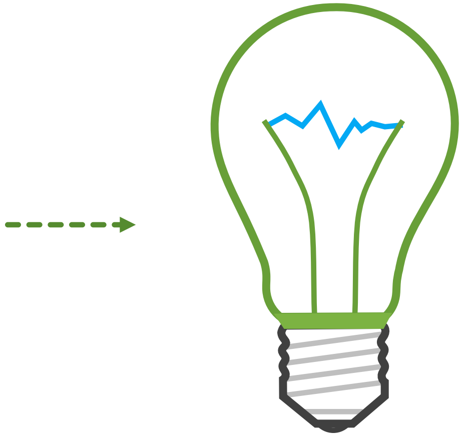
Wearable devices are widely popular. Sensor Data is available.

Third Input

3

Sensor data is tightly related and correlated with human overall health.





Idea

Fuse Sensor Data And Data From Multiple Social Networks For Wellness Profile Learning

Unite Social Media And Wearable Sensors For Physical Attributes Inference

Just tweet to be fit….

FOURSQUARE

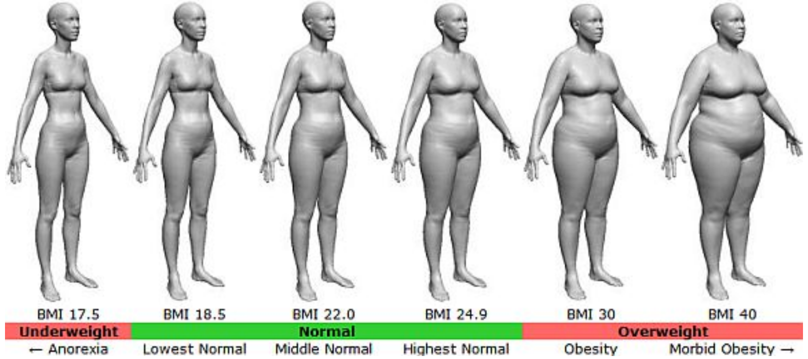
Instagram

twitter

+

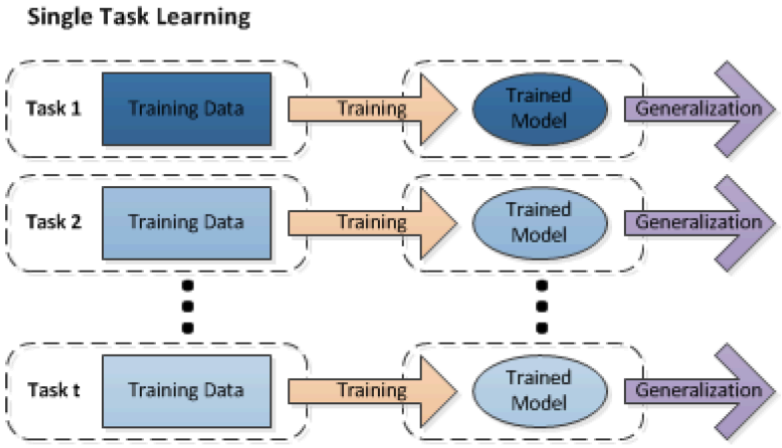
Weight Fluctuation Trend (BMI Trend)

+

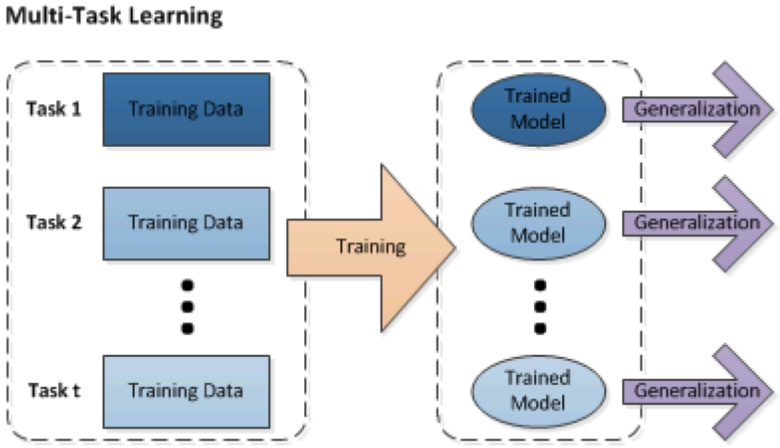


$$\mathrm{BMI} = \frac{height}{weight^2}$$

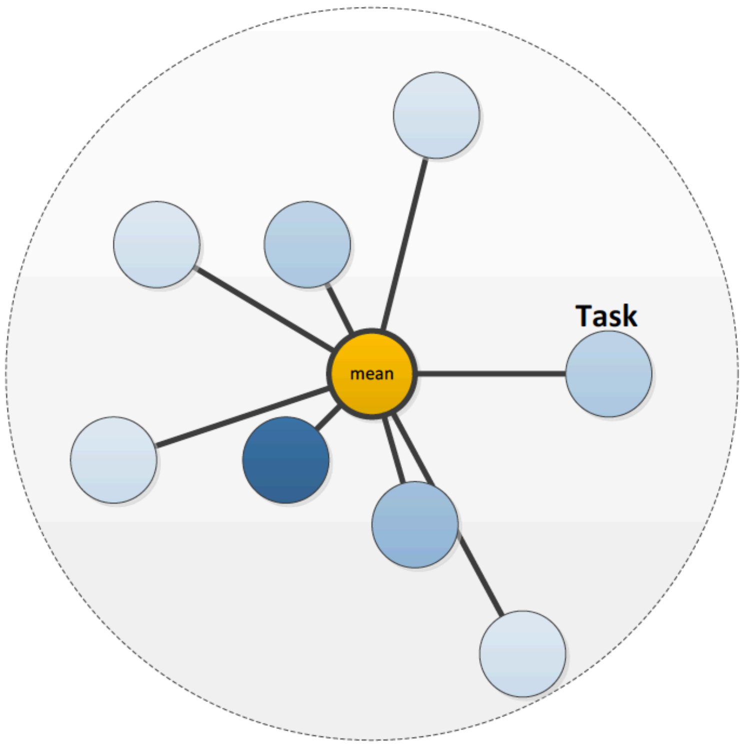
Background: Multi-Task Learning (MTL)



Multi-task Learning is different from single task learning in the training (induction) process.



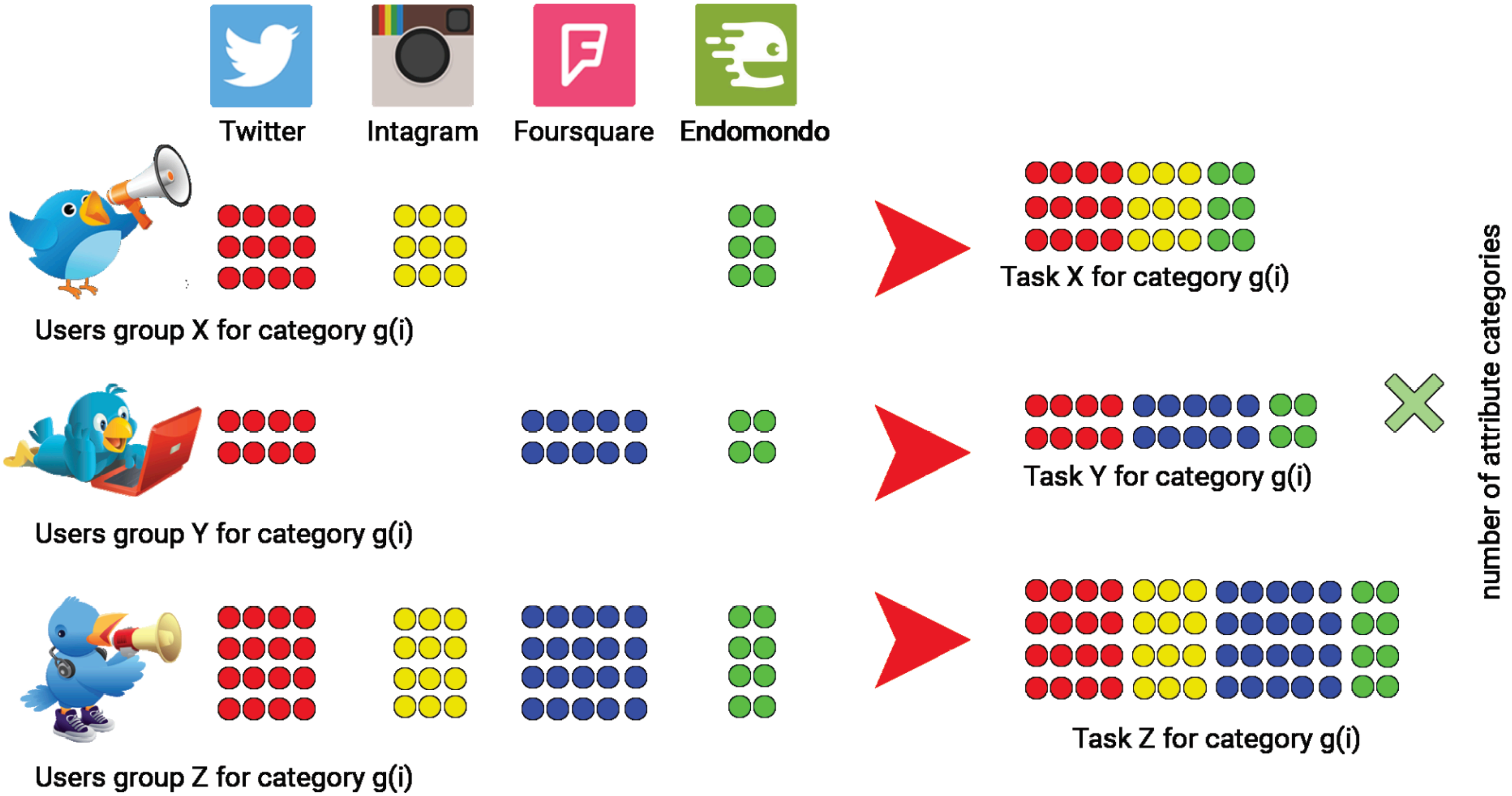
Inductions of multiple tasks are performed simultaneously to capture intrinsic relatedness.



Example of Regularized MTL: Assumption: task parameter vectors of all tasks are close to each other.

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Multi-Source Multi-Task Learning



E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Doing Predictions via Multi-Source Multi-Task Learning

Notations

Notation	Description
N	Number of exclusively labeled data samples
$S(\geq 2)$	Number of data sources (data modalities)
$G(\geq 1)$	Number of inference attribute categories (for BMI category, $G = 8$; for "BMI Trend", $G = 1$)
g	Inference attribute category (class). For example, "Obese" or "Normal" in case of BMI category attribute.
T	Number of multi-task learning Tasks
t	A multi-task learning Task
D_t	Dimensionality (feature vector dimension) of the task t
D_{max}	Maximum possible dimensionality of a task
N_t	Number of data samples of the task t
$\hat{T}$	Number of different existing combinations of sources
$f_t(\mathbf{x}_j^t; \mathbf{w}^t)$	Linear prediction model for the jth data sample of task t
$\mathbf{w}^t \in \mathbb{R}^{D_t}$	Model parameter vector of task t
$\mathbf{W}$	All model parameters, denoted as linear mapping block matrix
$\Gamma(\mathbf{W})$	Objective function
$\Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y})$	Loss function
$\Upsilon(\mathbf{W})$	Sparsity regularizer
$\Omega(\mathbf{W})$	Inter-category smoothness regularizer
$\rho(s, f)$	Index function that denotes all the model parameters of the fth feature from the sth source
$\xi(t, g)$	Index function that picks up the model parameter ($\mathbf{w}_{g+1}^t$), which corresponds to the attribute category $g + 1$ (adjacent to g)

$$\Gamma(\mathbf{W}) = \arg\min_{\mathbf{W}} \Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \lambda \Upsilon(\mathbf{W}) + \mu \Omega(\mathbf{W}),$$

$$\Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y}) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{N_t} \sum_{i=1}^{N_t} \log(1 + e^{-y_i^t f_t(\mathbf{x}_i^t; \mathbf{w}^t)})$$

$$\Upsilon(\mathbf{W}) = \sum_{s=1}^{S} \sum_{f=1}^{F_s} \left\| \mathbf{w}_{\rho(s,f)} \right\|$$

$$\Omega(\mathbf{W}) = \sum_{t=1}^{\hat{T}} \sum_{g \in C_{D_t}} \kappa_{g,\xi(t,g)} \left\| \mathbf{w}_g^t - \mathbf{w}_{\xi(t,g)}^t \right\|^2$$

BMI Category and BMI Trend Prediction: Results (1)

Data Source Combinations

Data Source Combination	BMI category prediction R_{Mac}/P_{Mac}	BMI category prediction $F_{1,Mac}$
Visual	0.049/0.188	0.077
Venue Semantics & Mobility	0.194/0.107	0.137
Sensors	0.153/0.158	0.155
Textual	0.229/0.146	**0.178**
Visual + Sensors	0.174/0.201	0.186
Visual + Text	0.126/0.245	0.166
Visual + Venue Semantics & Mobility	0.161/0.154	0.157
Text + Venue Semantics & Mobility	0.160/0.204	0.179
Sensors + Venue Semantics & Mobility	0.163/0.233	**0.191**
Sensors + Text	0.148/0.270	**0.191**
Visual + Text + Venue Semantics & Mobility	0.126/0.233	0.163
Sensors + Text + Visual	0.137/0.207	0.164
Sensors + Text + Venue Semantics & Mobility	0.182/0.236	0.205
Sensors + Venue Semantics & Mobility + Visual	0.180/0.283	**0.221**
All Data Sources	0.214/0.292	**0.246**

Other Baselines

Method	BMI category R_{Mac}/P_{Mac}	BMI category $F_{1,Mac}$	"BMI Trend" R_{Mac}/P_{Mac}	"BMI Trend" $F_{1,Mac}$
MSESHC [47]	0.141/0.145	0.142	0.634/0.655	0.644
Random Forest	0.135/0.226	0.169	0.333/0.863	0.480
iMSF [160]	0.171/0.174	0.172	0.649/0.649	0.649
$aMTFL_2$ [85]	0.162/0.215	0.184	0.700/0.722	0.710
$TweetFit$	0.222/0.202	0.211	0.705/0.732	**0.718**
M^2WP	0.221/0.229	**0.225**	Ω is not applicable	

8 BMI Categories:
Thinness I, II, III; Normal; Obese I, II, III, IV
2 BMI Trends: Increase; Decrease

Contributions

First Work on Combining Sensor And Social Media Data For Wellness Attribute Inference — 01

Novel Generic Model For Supervised Learning From Multi-Source Multi-Modal Incomplete Data — 02

First Large-Scale Multi-Source Social-Sensor Dataset NUS-SENSE — 03

*Farseev, A., & Chua, T. S. (2017). Tweet can be Fit: Integrating Data from Wearable Sensors and Multiple Social Networks for Wellness Profile Learning. ACM Transactions on Information Systems (TOIS).

Every successful individual knows that his or her achievement depends on a community of persons working together.

- Paul Ryan

Group User Profiling

And Its Application In Urban Mobility Domain

(NUS-MSS)

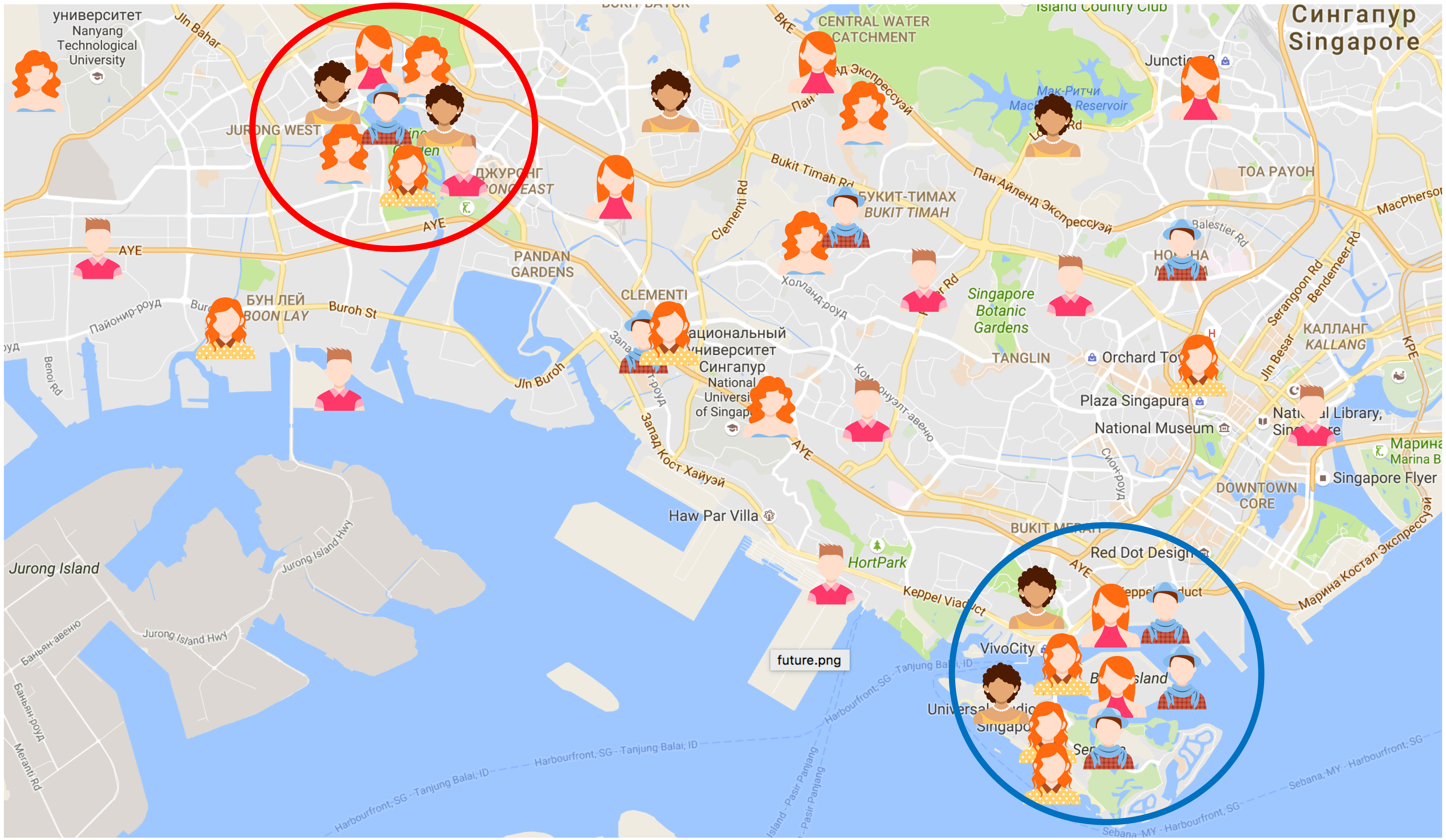
E-Mail: farseev@u.nus.edu I Website: http://farseev.com

User Community Definition

Communities are distinct groups of people, divided by a certain property.

Why do we need to detect user communities (1)?

For facility planning.



More Restaurants And Shops In Jurong West

More Beach Facilities And Pubs in Sentosa

Why do we need to detect user communities (2)?

For personalized advertisement and marketing

E-Mail: farseev@u.nus.edu I Website: http://farseev.com

Idea IV

Going further towards realizing the ideal of 360° User Profile Learning

Input One

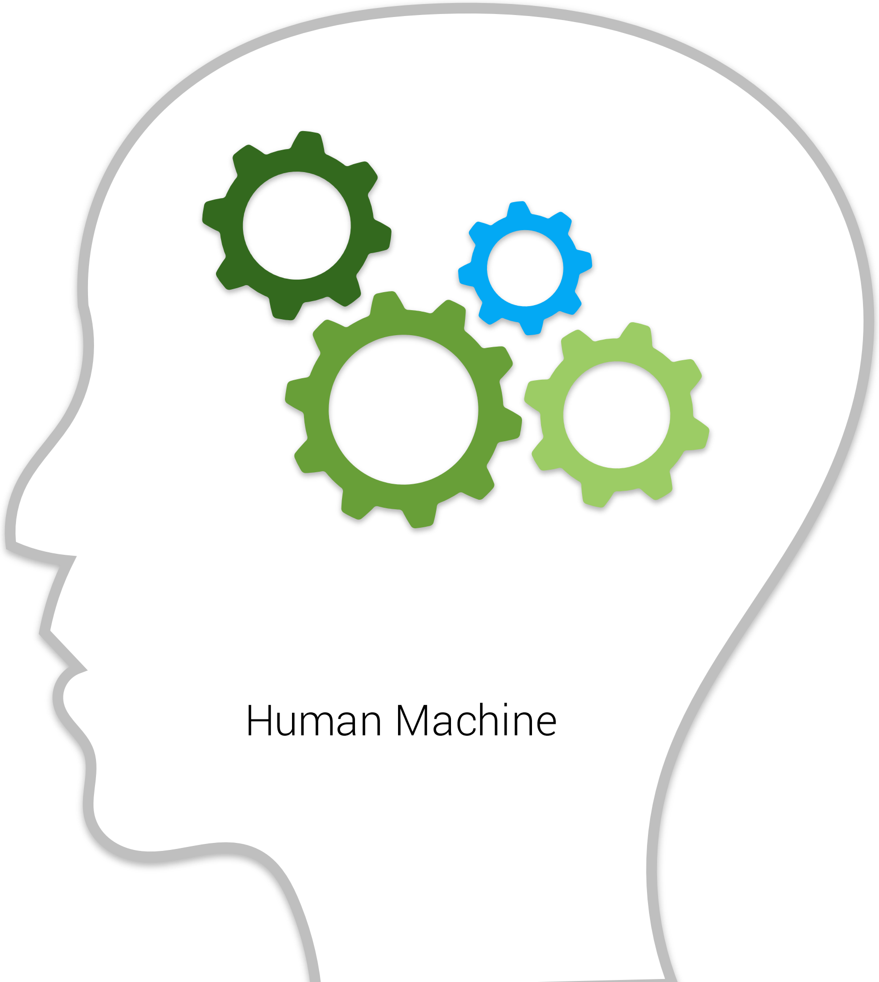
1

Integration of Data from multiple social networks helps in improving Individual Profiling Performance.

Second Input

2

It is still unclear whether multi-source data allows for group profiling performance improvement.



Idea

Perform Group User Profiling (User Community Detection) Based Multiple Social Networks

Related Research Directions: Group User Profiling

Single-Source Clustering

Mostly mono-modal mono-source data processing:

- Graph Partitioning
- Hierarchical Clustering
- Partition Clustering
- Spectral clustering
- Quality Optimization

Multi-Source Clustering

Limited number of contributions to the field. Graphs are often combined in late fusion manner, but not learned jointly. Inter-source relationship is not considered.

Urban Mobility Applications

Mostly mono-modal mono-source data processing in: recommendation, prediction, qualitative analysis.

Cross-Domain Recommendation

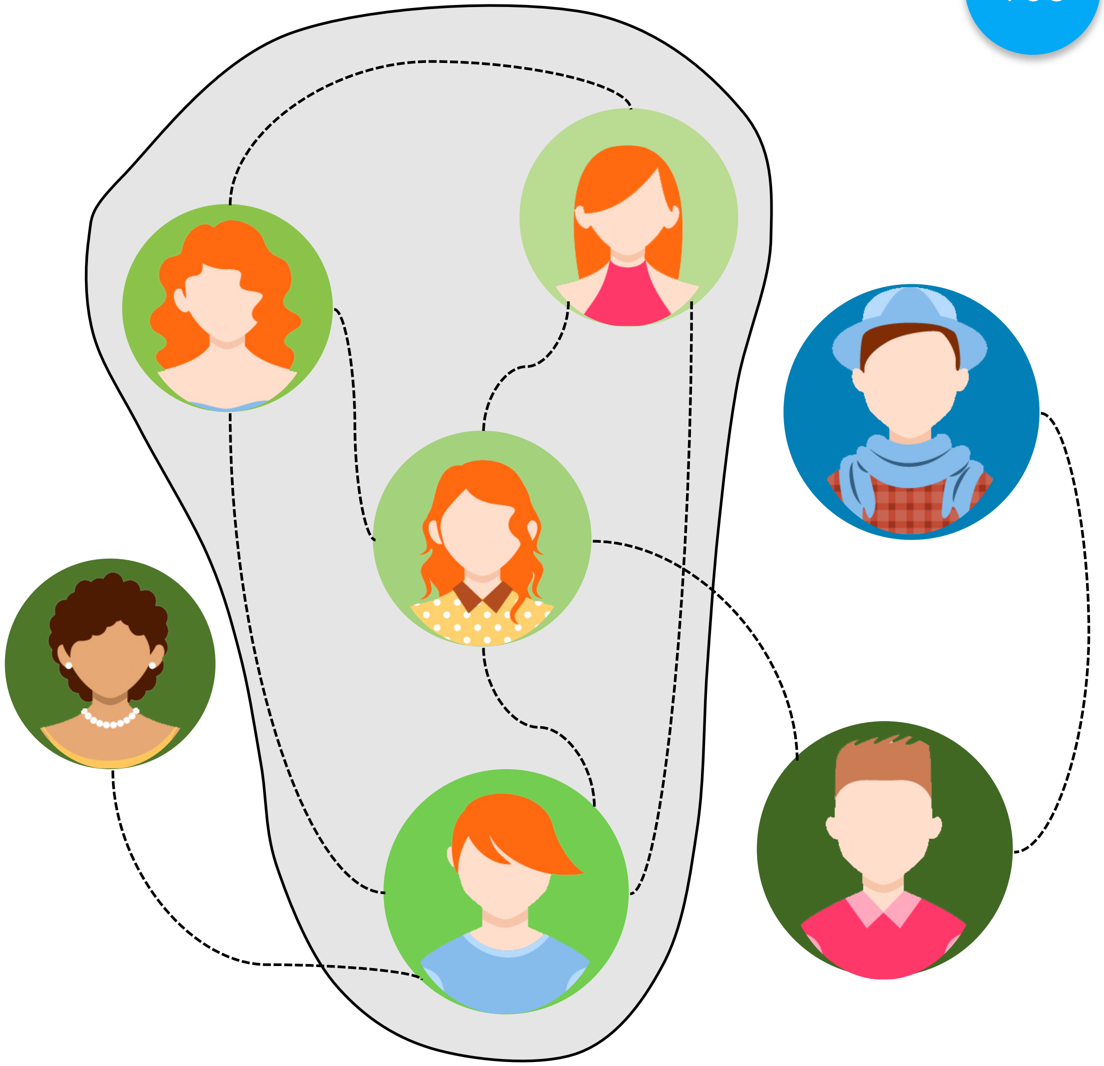
Inter-source relationship is not considered in most of the studies.

Usually, this problem of group user profiling is treated as a unsupervised learning task and performed based on mono-source mono-modal datasets. Research on multi-source group profiling is relatively sparse and limited by lack of inter-source relationship modeling. Implicit evaluation is adopted in most of the related studies.

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

User Relations and Community Representation

One way to find representative user groups is to model users' relationships in the form of a graph so that dense subgraphs are considered to be user communities.



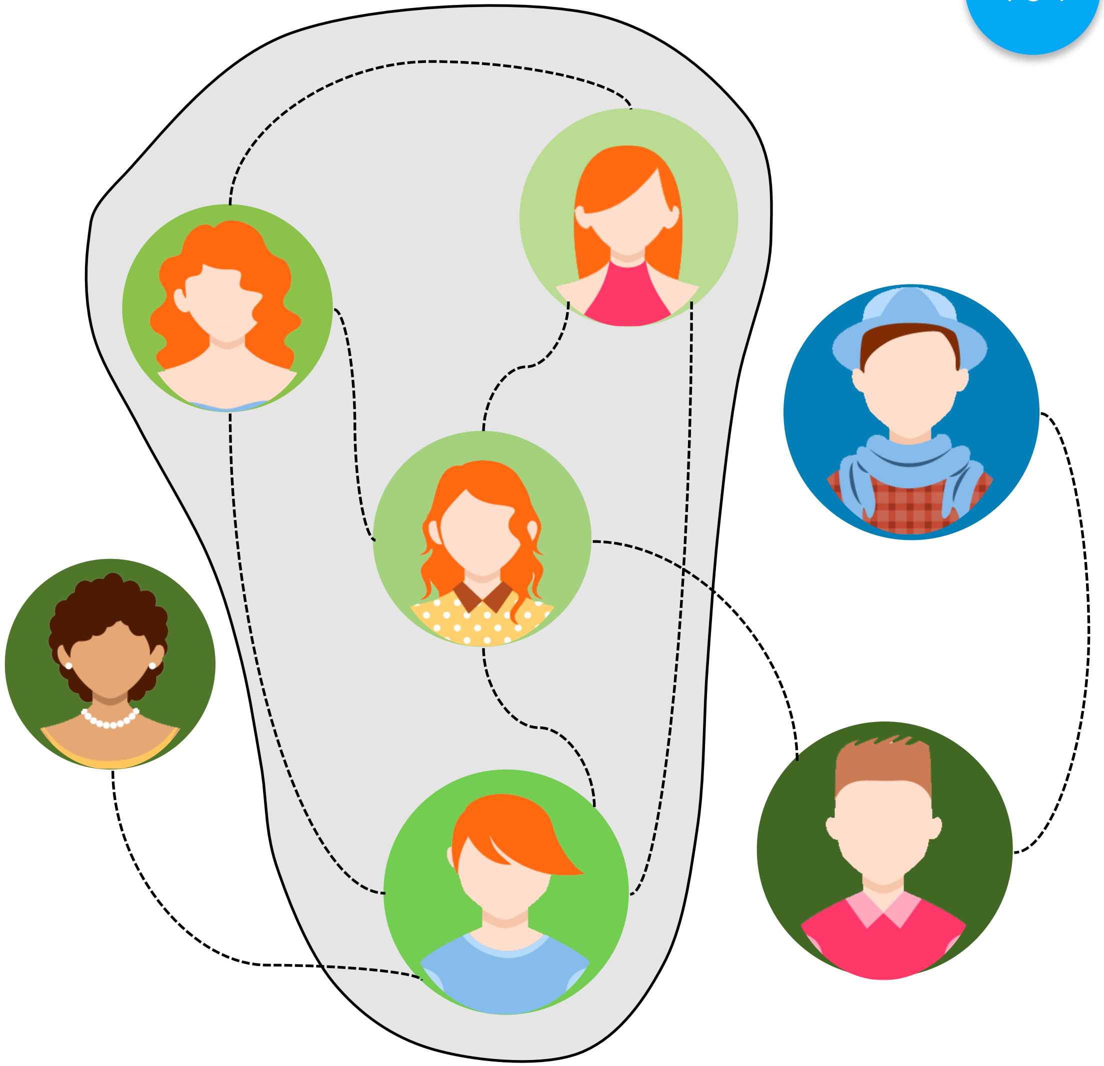
Finding Communities in single-source data

One of the commonly used formulations of the community detection problem is its representation in a form of $MinCut$ problem.

For a given number k of subsets, the $MinCut$ involves choosing a partition $C_1, \dots, C_k$ such that it minimizes the expression:

$$cut(C_1, \dots, C_k) = \sum_{i=1}^{k} W(C_i, \bar{C}_i)$$

*W is the sum of weights of edges attached to vertices in C_i



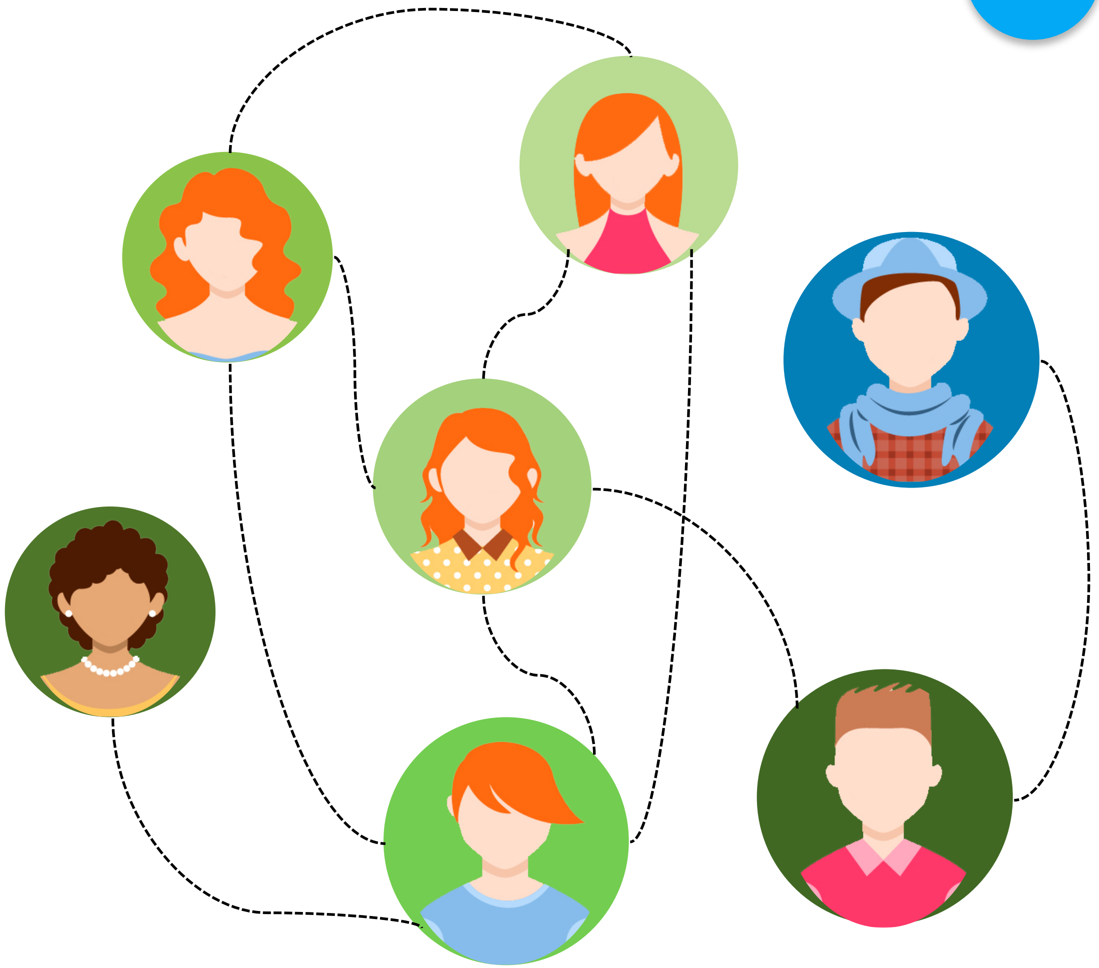
Approximating and solving $MinCut$ problem?

The state-of-the-art approximation of $MinCut$ is formulated in a form of **standard trace minimization problem**

$$\min_{U \in R^{n \times k}} \operatorname{tr}(U^T L U), \text{s.t. } U^T U = I$$

which can be solved by **Spectral Clustering:**

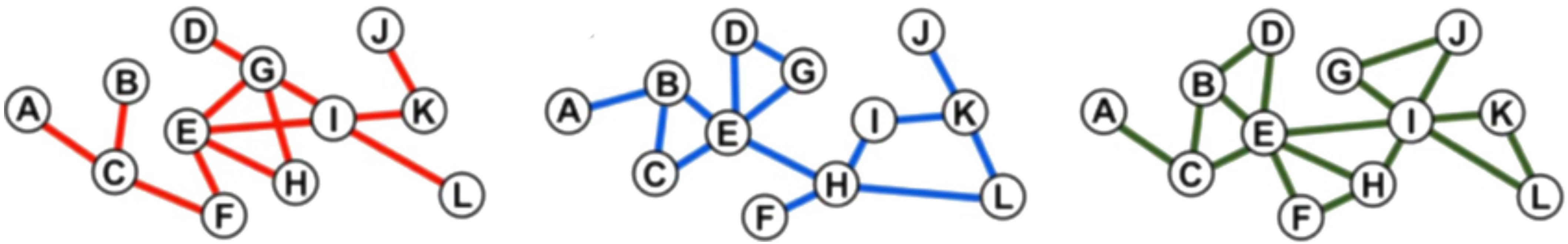
1. Calculates Laplacian matrix $L \in R^{n \times n}$
2. Builds matrix of the first k eigenvectors $U \in R^{n \times k}$ (correspond to the smallest eigenvalues of L)
3. Clusters data in a new space U using i.e. k-means algorithm



E-Mail: farseev@u.nus.edu l Website: http://farseev.com

How to represent multi-source data?

Multi-layer graph – graph G, where $G = \{G_i\}, G_i = (V, E_i)$:

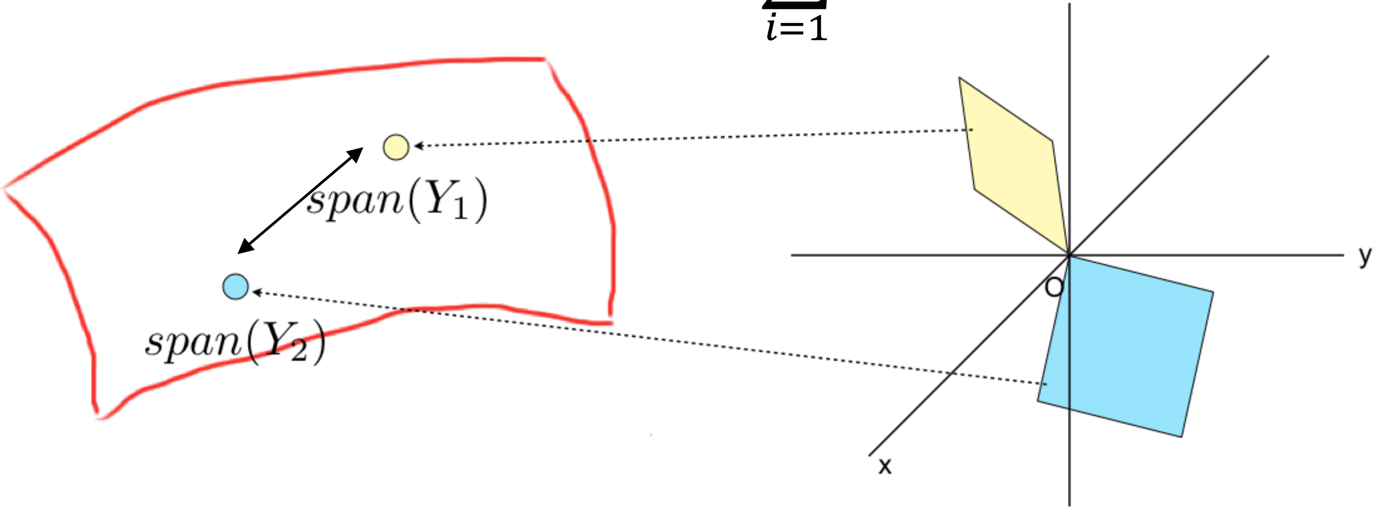


Regularized Clustering on Multi-layer Graph (1)

Utilize **Grassman Manifolds** to keep final latent representation "close" to all layers of multi-layer graph*. Where projected distance between two spaces Y_1 and Y_2:

$$d^2_{Proj}(Y_1, Y_2) = \frac{1}{2}\left\|Y_1Y_1^T - Y_2Y_2^T\right\|_F^2, \text{where } \|A\|_F \text{ is the Frobenius norm}$$

$$d^2_{Proj}(S, \{S_i\}_{i=1}^M) = kM - \sum_{i=1}^{M} \mathrm{tr}(SS^T - S_iS_i^T)$$



* X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov. Clustering on multi-layer graphs via subspace analysis on grassmann manifolds. IEEE Transactions on Signal Processing, 2014.

E-Mail: farseev@u.nus.edu l Website: http://farseev.com

Regularized Clustering on Multi-layer Graph (2)

Extends the objective function to introduce the subspace analysis regularization

$$\min_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^{M} \operatorname{tr}\left(U^T L_i U\right) + \alpha \left(kM - \sum_{i=1}^{M} \operatorname{tr}\left(U U^T U_i U_i^T\right) \right), \text{s.t. } U^T U = I$$

$$\min_{U \in \mathbb{R}^{n \times k}} \operatorname{tr}\left(U^T L_{mod} U\right)$$

$$L_{mod} = \sum_{i=1}^{M} \left(L_i - \alpha U_i U_i^T\right)$$

Incorporating inter-layer relationship (1)

By using previously defined distance on Grassman Manifold, we present the objective function for the i^{th} layer as follows:

$$\min_{\hat{U}_i \in \mathbb{R}^{n \times k}} \sum_{i=1}^{M} \operatorname{tr}\left(\hat{U}_i^T L_i \hat{U}_i\right) + \beta_i \left(kM - \sum_{j=1, j \neq i}^{M} w_{i,j} \operatorname{tr}\left(\hat{U}_i \hat{U}_i^T U_j U_j^T\right) \right), \text{s.t.} \hat{U}_i^T \hat{U}_i = I$$

$$\min_{U \in \mathbb{R}^{n \times k}} \operatorname{tr}\left(\hat{U}_i{}^T \hat{L}_i \hat{U}_i\right)$$

$$\hat{L}_i = L_i - \beta_i \sum_{j=1, j \neq i}^{M} w_{i,j} \operatorname{tr}\left(U_j U_j^T\right)$$

Final Objective Function

Let's combine equations from previous slides to define the final objective function:

$$\min_{\hat{U}_i \in \mathbb{R}^{n\times k}} \sum_{i=1}^{M} \mathrm{tr}\left(\hat{U}_i^T L_i \hat{U}_i\right) + a\left(kM - \sum_{j=1, j\neq i}^{M} \boxed{w_{i,j}}\, \mathrm{tr}\left(\hat{U}_i \hat{U}_i^T U_j U_j^T\right)\right), \mathrm{s.t.}\, \hat{U}_i^T \hat{U}_i = I$$

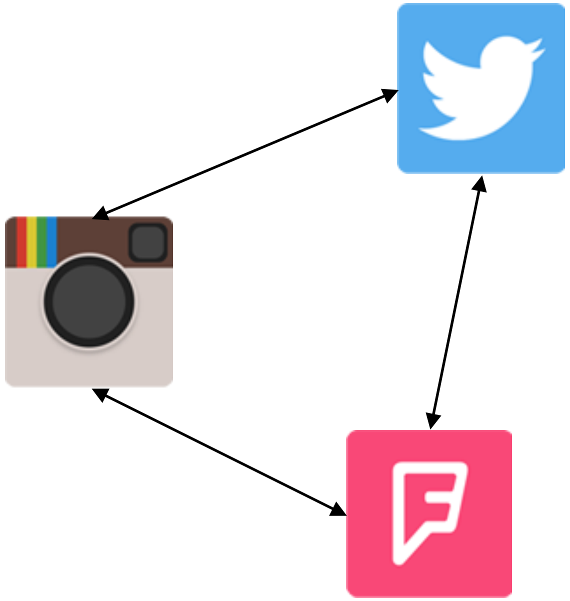
$$\hat{L}_i = L_i - \beta_i \sum_{j=1, j\neq i}^{M} w_{i,j}\, \mathrm{tr}\left(U_j U_j^T\right)$$

Inter-layer relationship graph construction

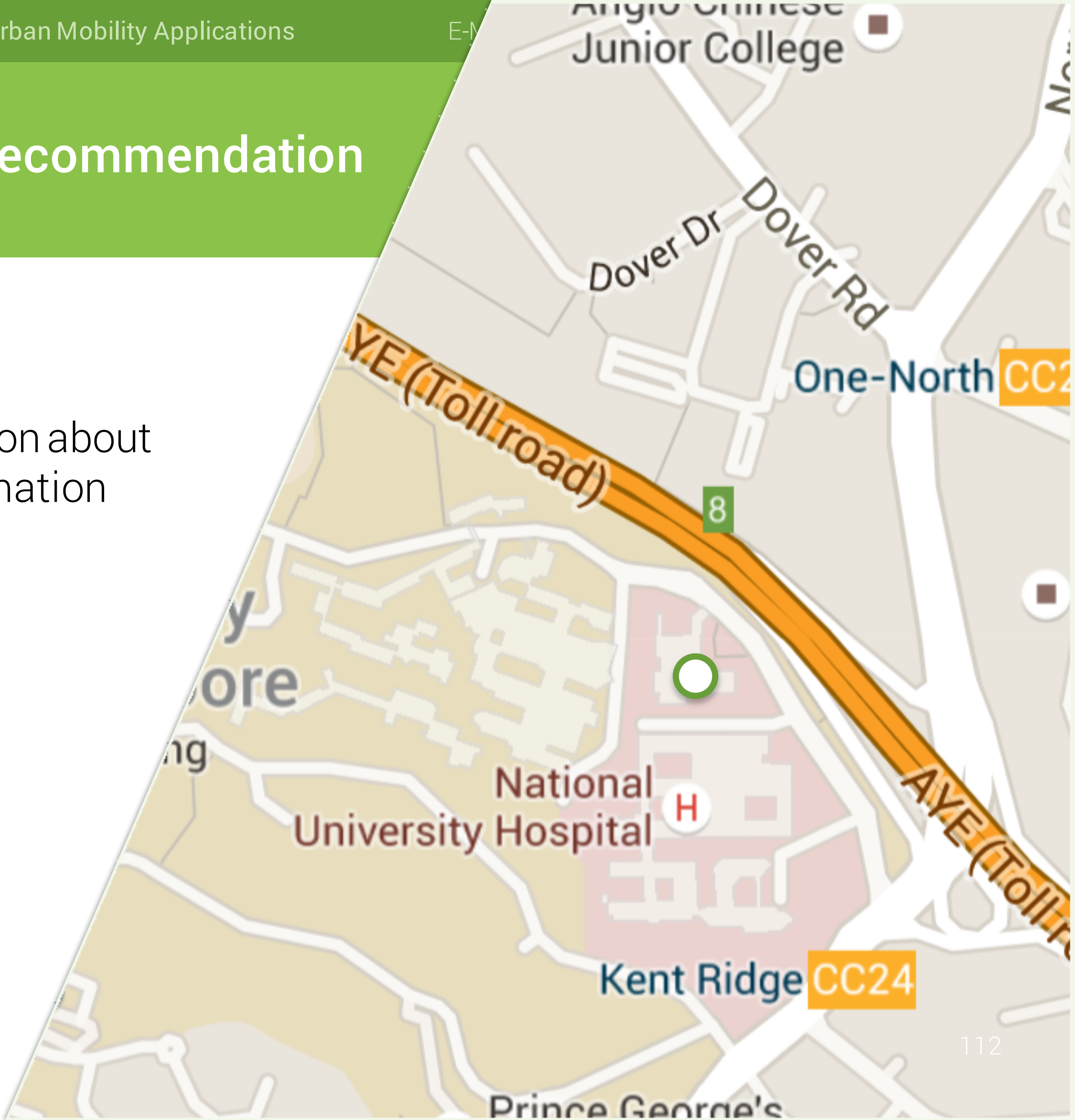
Inter-layer relationship graph $\boldsymbol{R(V,E)}$ – weighted graph which represents the similarity between layers.

$$\forall(i,j) \in E, w_{i,j} = \frac{\sum_{k=2}^{K}\left(1 - \frac{\left\|M_{i,k} - M_{j,k}\right\|}{\sqrt{N(N-1)}}\right)}{K-1}$$

where $M_{i,k}$ is clustering co-occurrence matrix of layer i, $m_{a,b} = 1$, if users a and b assigned to the same cluster, and 0 otherwise.



Implicit Evaluation: Venue Category Recommendation



Venue category recommendation – recommendation of venue categories (i.e. restaurant, cinema) to user using information about his/her profile (i.e. past visits) and/or information about users from the same domain.

Venue category:
- Clothing Store
- Hotel
- Ice Cream Shop

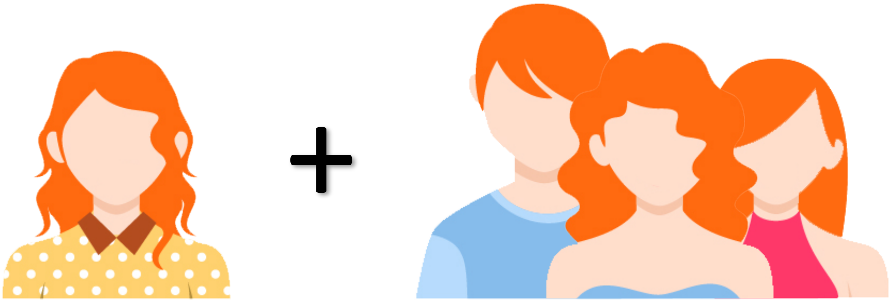
Total 764 different categories

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Cross-Source Community-Based Collaborative Recommendation

We perform venue category recommendation based on both Individual and Group Knowledge, which naturally models the impact of society on an individual's behavior during the selection of a new place to go:

$$rec(u) = sort\left(\gamma \cdot vec_u + \theta \frac{\sum_{v \in C_u} vec_v}{|C_u|}\right)$$



Venue Recommendation Performance in Three Cities (1): Baselines

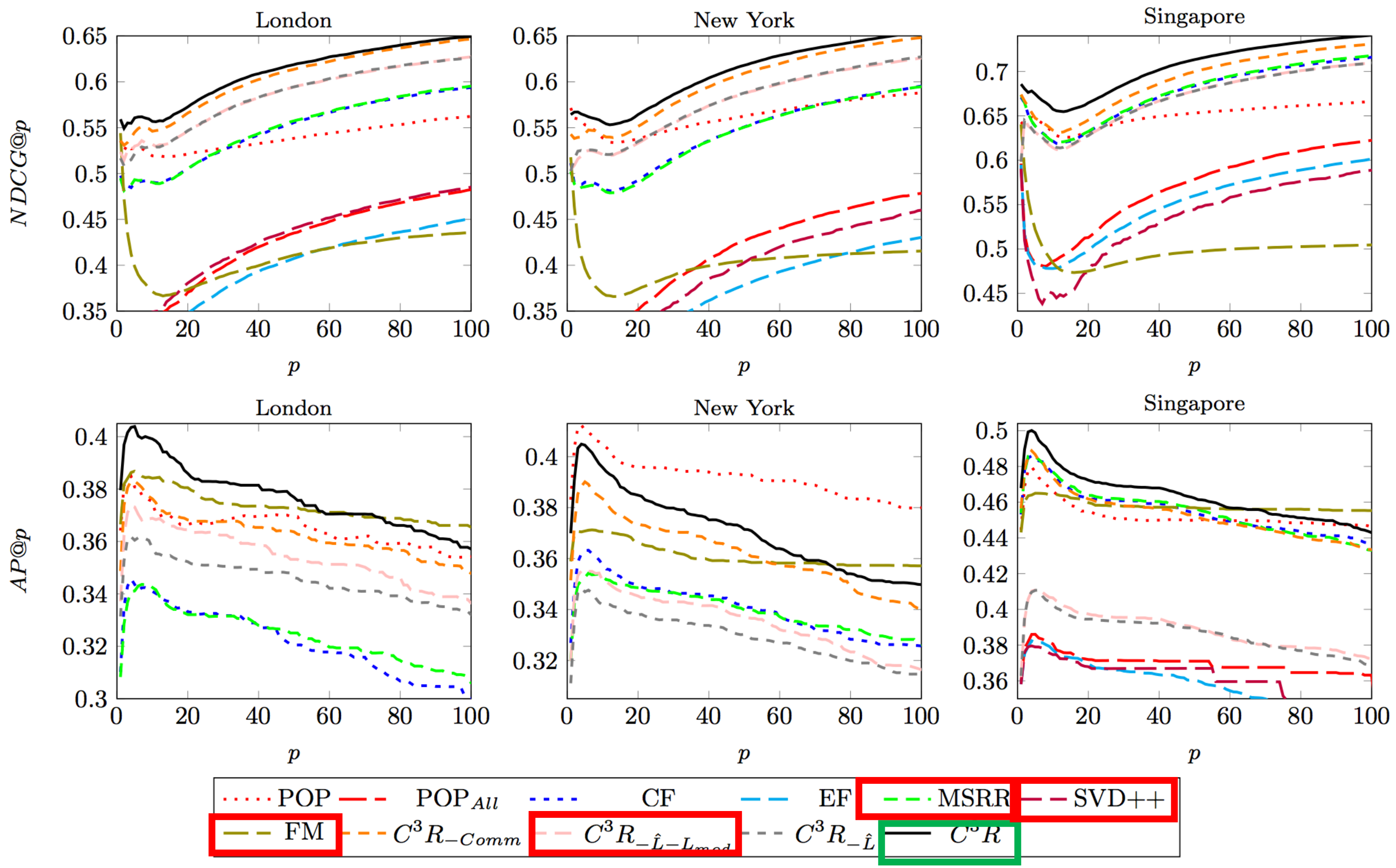


Figure 1: Evaluation of C^3R recommendation framework against baselines ($NDCG@p$, $AP@p$)

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Venue Recommendation Performance in Three Cities (2): Source Combinations

SN combination	London	New York	Singapore
tw	0.547	0.535	0.627
fsq	**0.549**	**0.536**	**0.628**
inst	0.548	0.535	**0.628**
tw + fsq	**0.557**	**0.541**	0.636
tw + inst	0.550	0.537	0.632
fsq + inst	0.551	0.539	**0.646**
tw + fsq + inst	**0.559**	**0.548**	**0.653**

*NDCG@60

+ Different Data Sources Differently Affect Community Detection And Recommendation
+ In Different Geo Regions, Different Data Sources Play Different Roles

$$W_R = \left(\begin{array}{c|ccccc} & \textbf{tw} & \textbf{4sq} & \textbf{inst} & \textbf{tmp} & \textbf{mob} \\ \hline \textbf{tw} & 1 & 0.632 & 0.621 & 0.643 & 0.561 \\ \textbf{4sq} & 0.632 & 1 & 0.614 & 0.631 & 0.570 \\ \textbf{inst} & 0.621 & 0.614 & 1 & 0.621 & 0.551 \\ \textbf{tmp} & 0.643 & 0.631 & 0.621 & 1 & 0.560 \\ \textbf{mob} & 0.561 & 0.570 & 0.551 & 0.560 & 1 \end{array}\right)$$

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Contributions

New Approach For Clustering Multi-Source Multi-Modal Data that considers inter-network relationship — 01

Novel Schema For Automatic Inter-Layer Relationship Inference Directly From Data — 02

*Farseev, A., Samborskii, I., Filchenkov, A., & Chua, T. S. (2017 August). Cross-Domain Recommendation via Clustering on Multi-Layer Graphs. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.

“

Big Data is not about the data

- Gary King

Business Application

"bBridge" Social Multimedia Analytics Platform

*Farseev, A., Samborskii, I., & Chua, T. S. (2016, October). bBridge: A Big Data Platform for Social Multimedia Analytics. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 759-761). ACM.

Video: http://video.bbridge.net

Demo: http://demo.bbridge.net

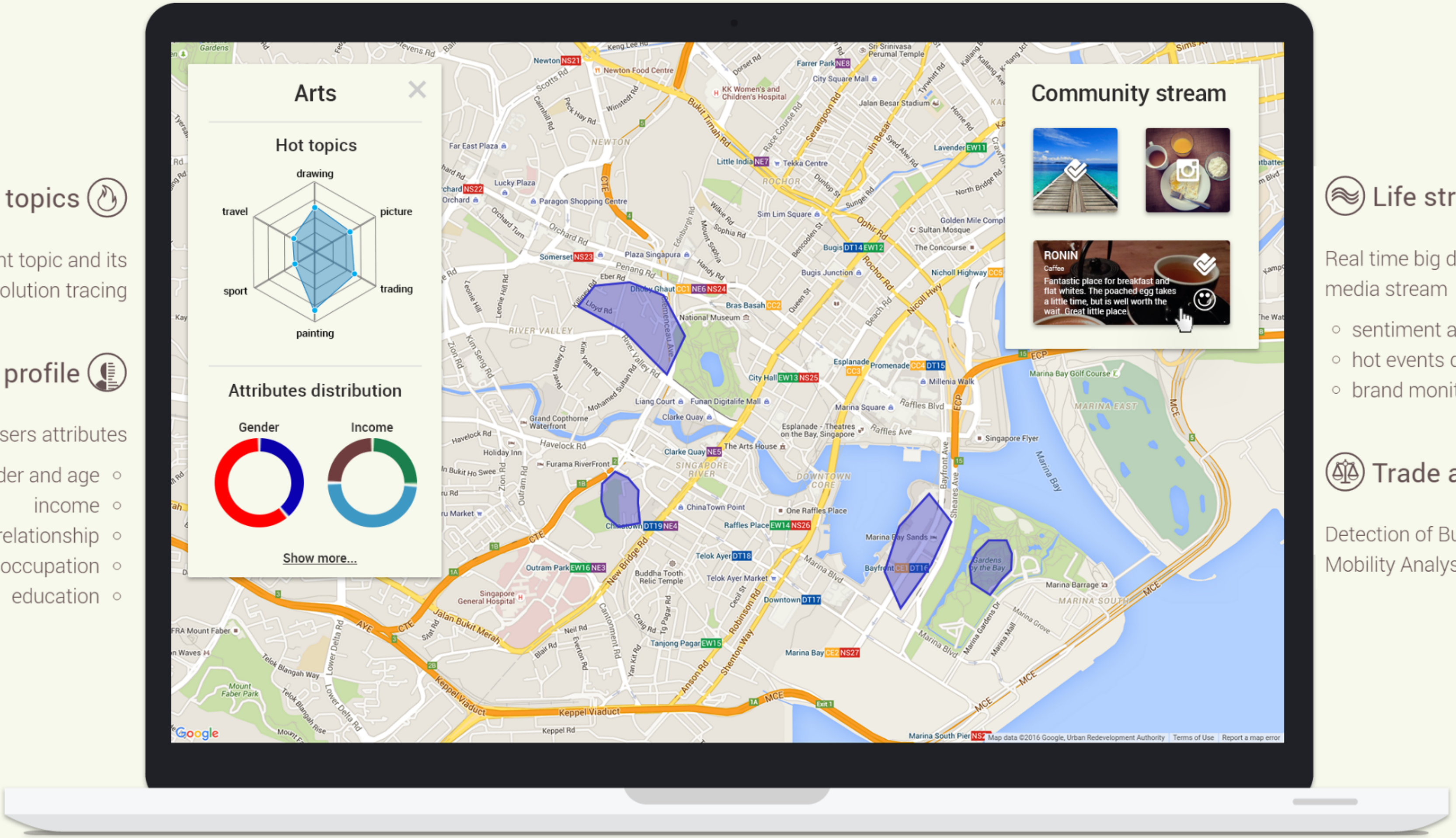
Hot topics

Extraction of popular multi-modal latent topic and its evolution tracing

Rich community profile

Automatic detection of community users attributes

- gender and age
- income
- relationship
- occupation
- education



Life stream analytics

Real time big data stream analytics on community social media stream

- sentiment analysis
- hot events detection
- brand monitoring and many more

Trade area analysis

Detection of Business Trading Areas based upon Urban Mobility Analysis

*Farseev, A., Samborskii, I., & Chua, T. S. (2016, October). bBridge: A Big Data Platform for Social Multimedia Analytics. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 759-761). ACM.

E-Mail: farseev@u.nus.edu l Website: http://farseev.com

Tech. Behind bBridge

The Analytics Pipeline

User Communities Detection

via clustering

on multi-layer graph

User Communities Profiling

via users' attributes inference

Hot Topic Detection

via multi-modal topic modelling

Sentiment Analysis and Hot Events Detection

User Communities Monitoring

bBridge API

Something specially for WSSS’17

Demo: http://video3.bbridge.net

User Profiling
Age, Gender, Occupation, Education, Relationship inference

Image Object Detection
Detecting Objects and Concepts on Images

Multilingual Natural Language Processing
Named Entity Recognition, Sentiment Analysis, Part of Speech Detection

API Access: http://api.bbridge.net

E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Contributions

One of the First Social Multimedia Analytics Platforms that Make Use of Text, Images, and Locations Simultaneously

01

“

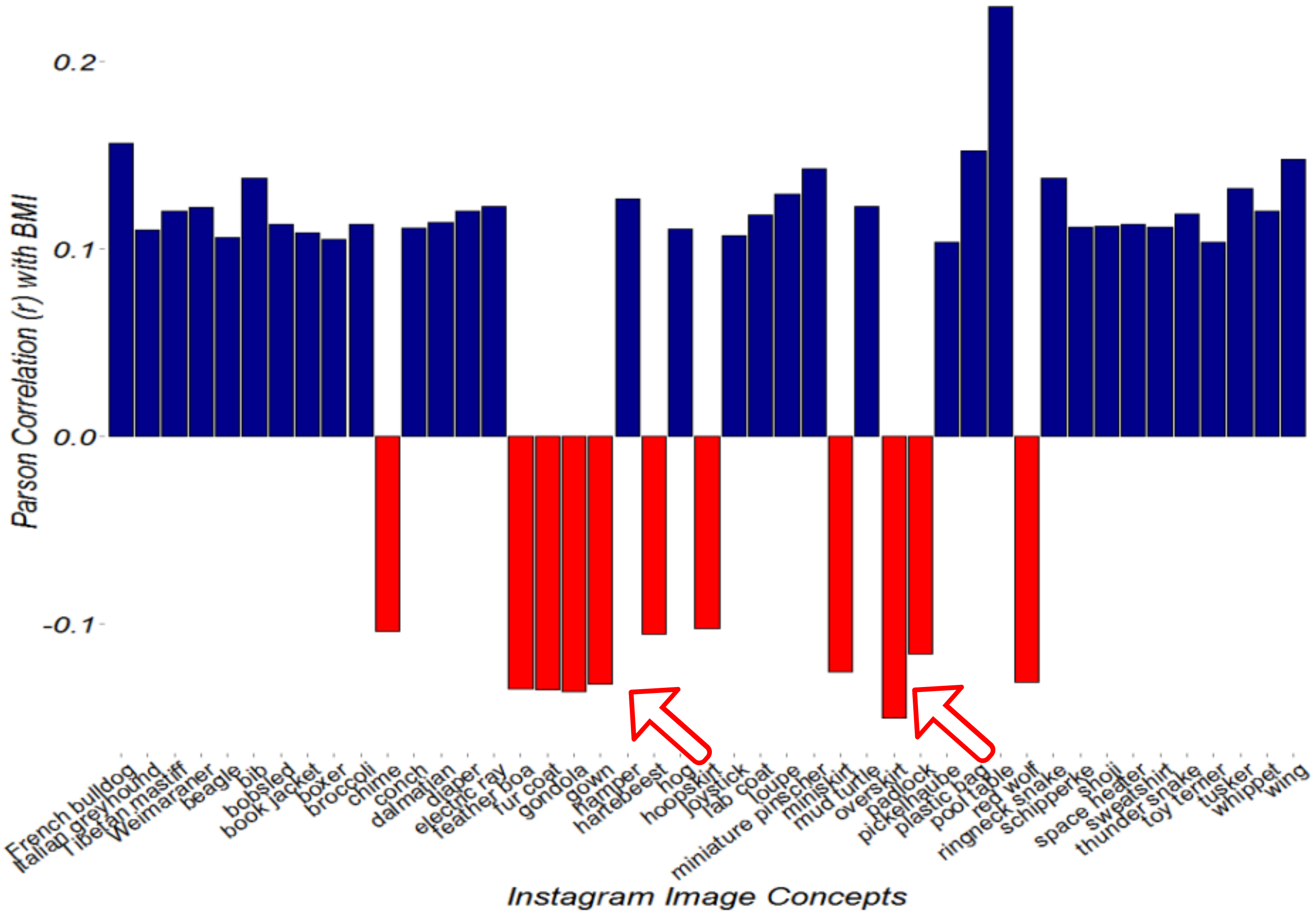
It is a mistake to try to look too far ahead. The chain of destiny can only be grasped one link at a time.

- Winston S. Churchill

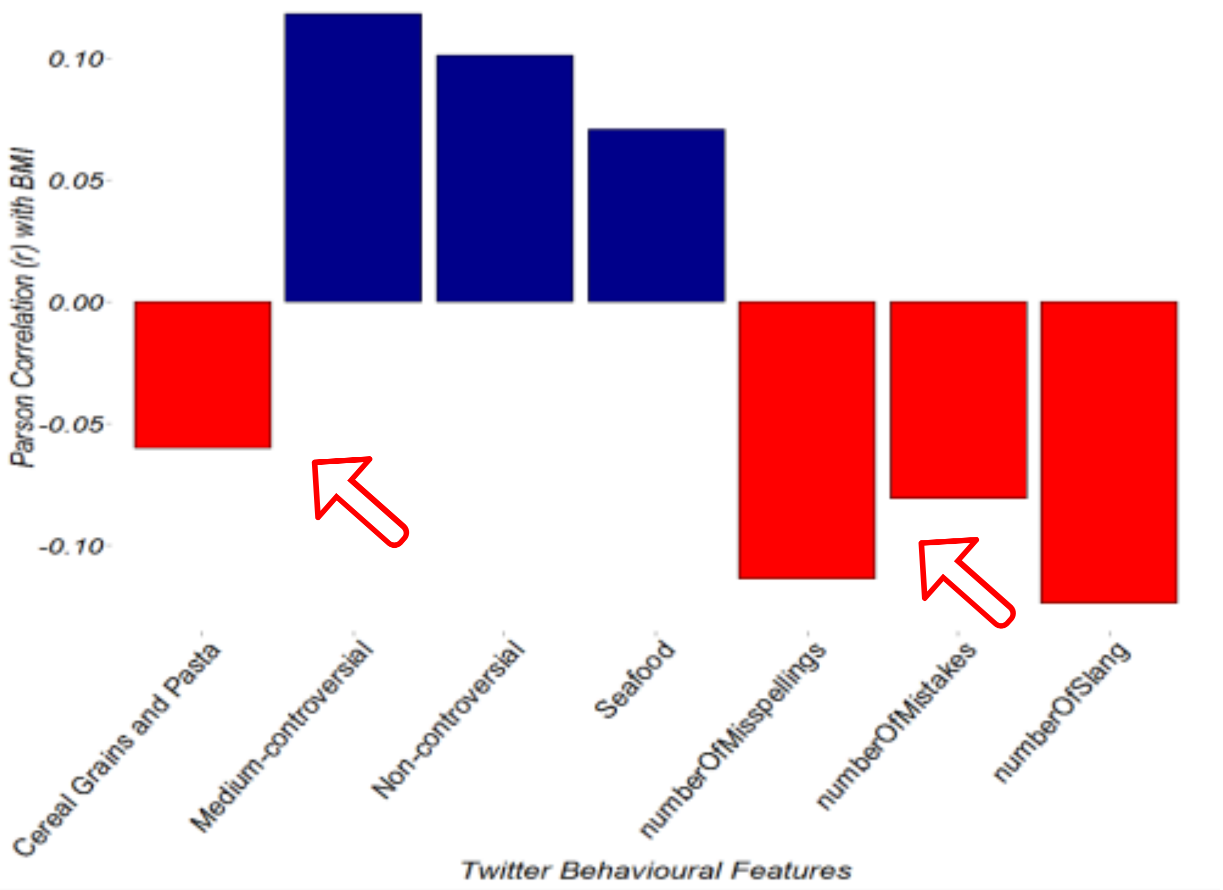
Some additional inspiration…

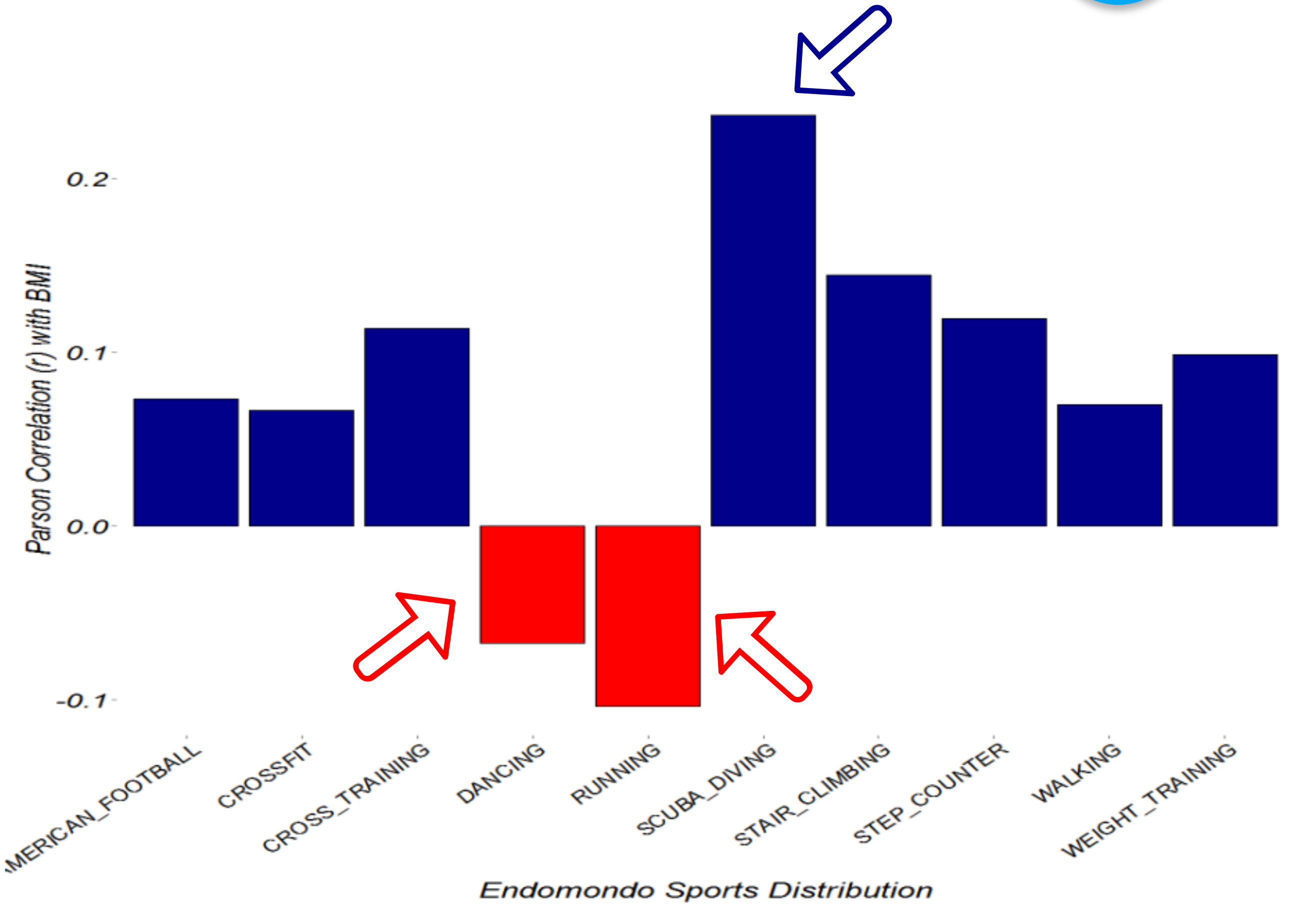
E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Examples of Wellness Correlations (1): Individual Profiling



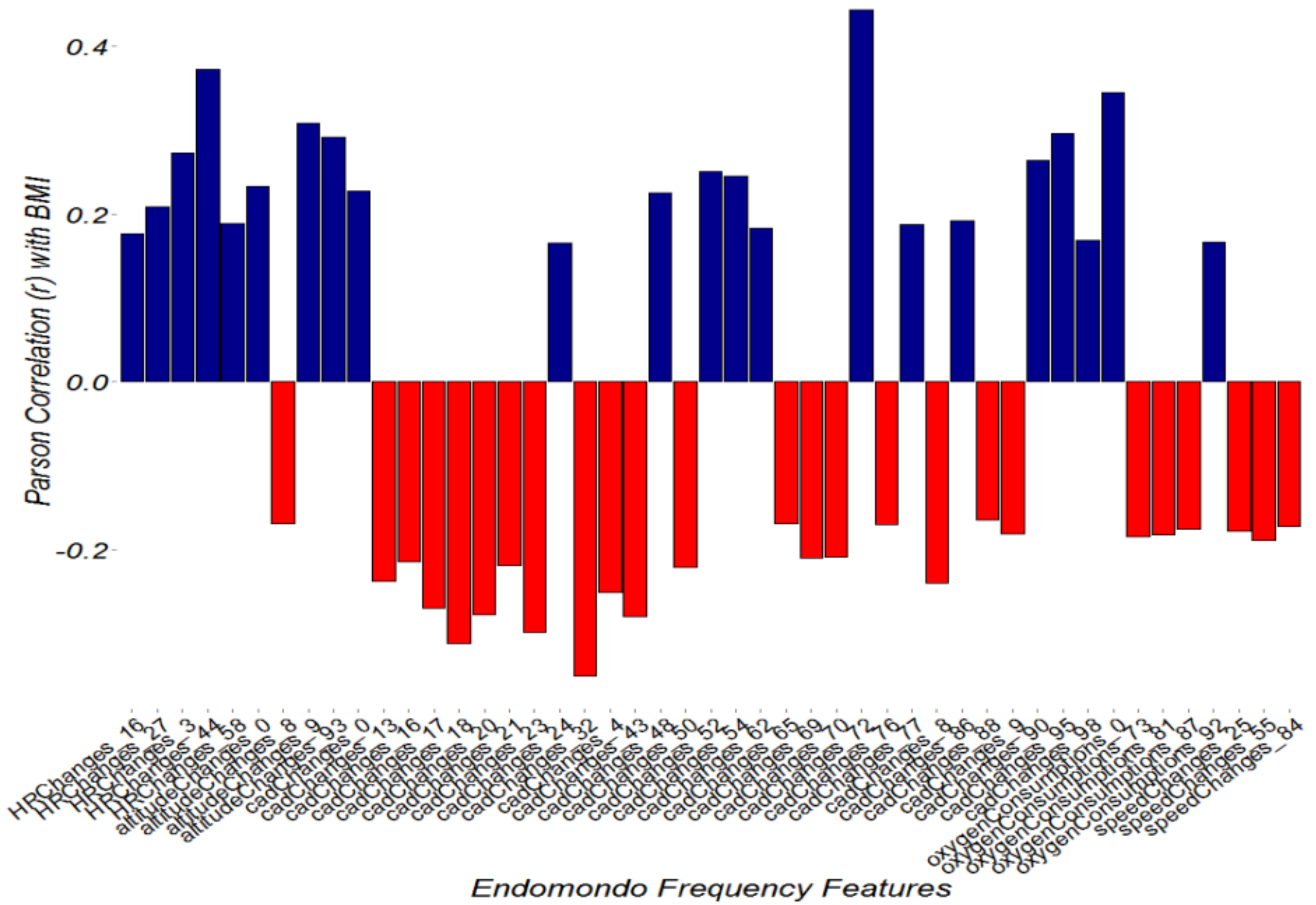
Examples of Wellness Correlations (2): Individual Profiling



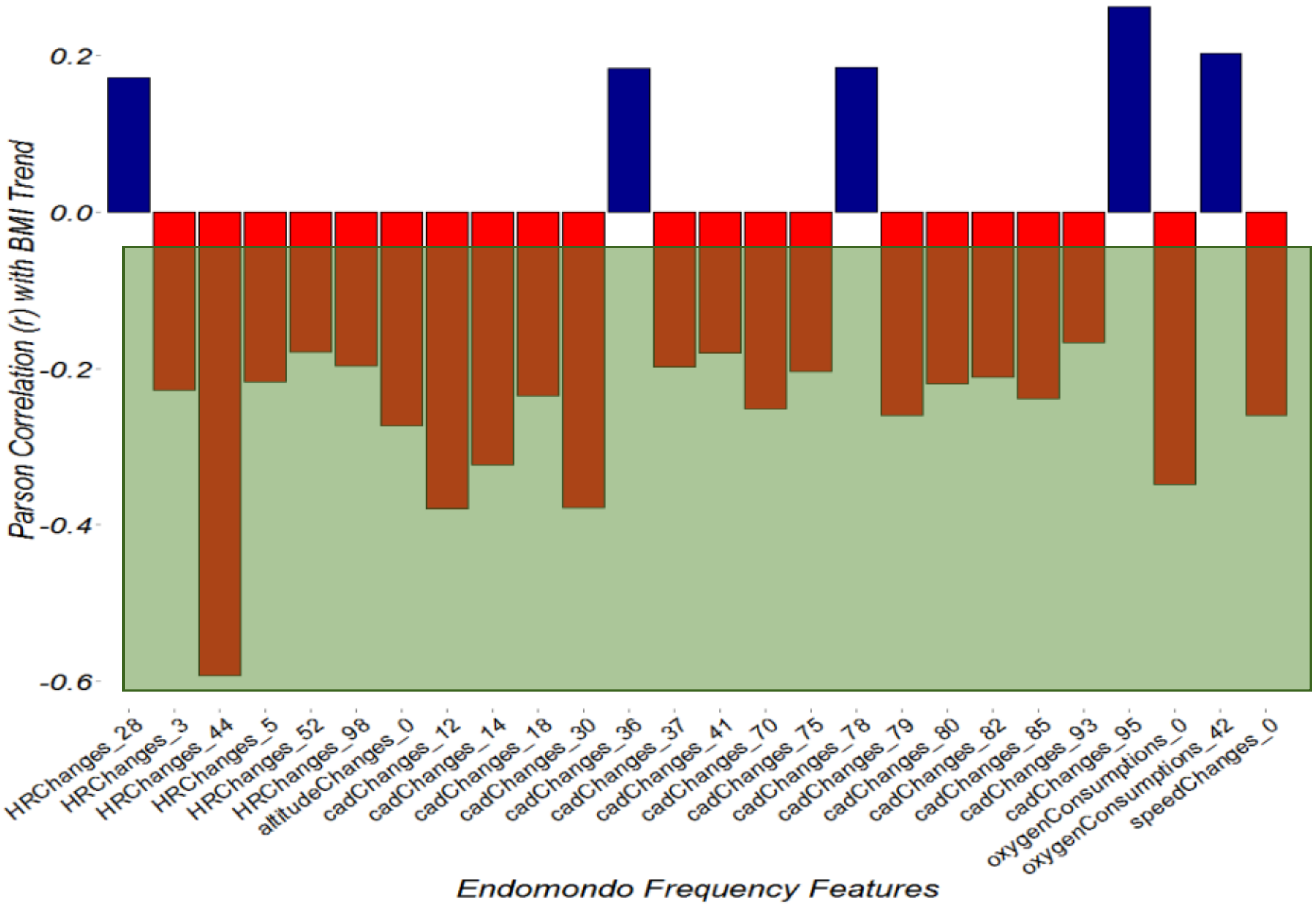


E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Examples of Wellness Correlations (3): Individual Profiling

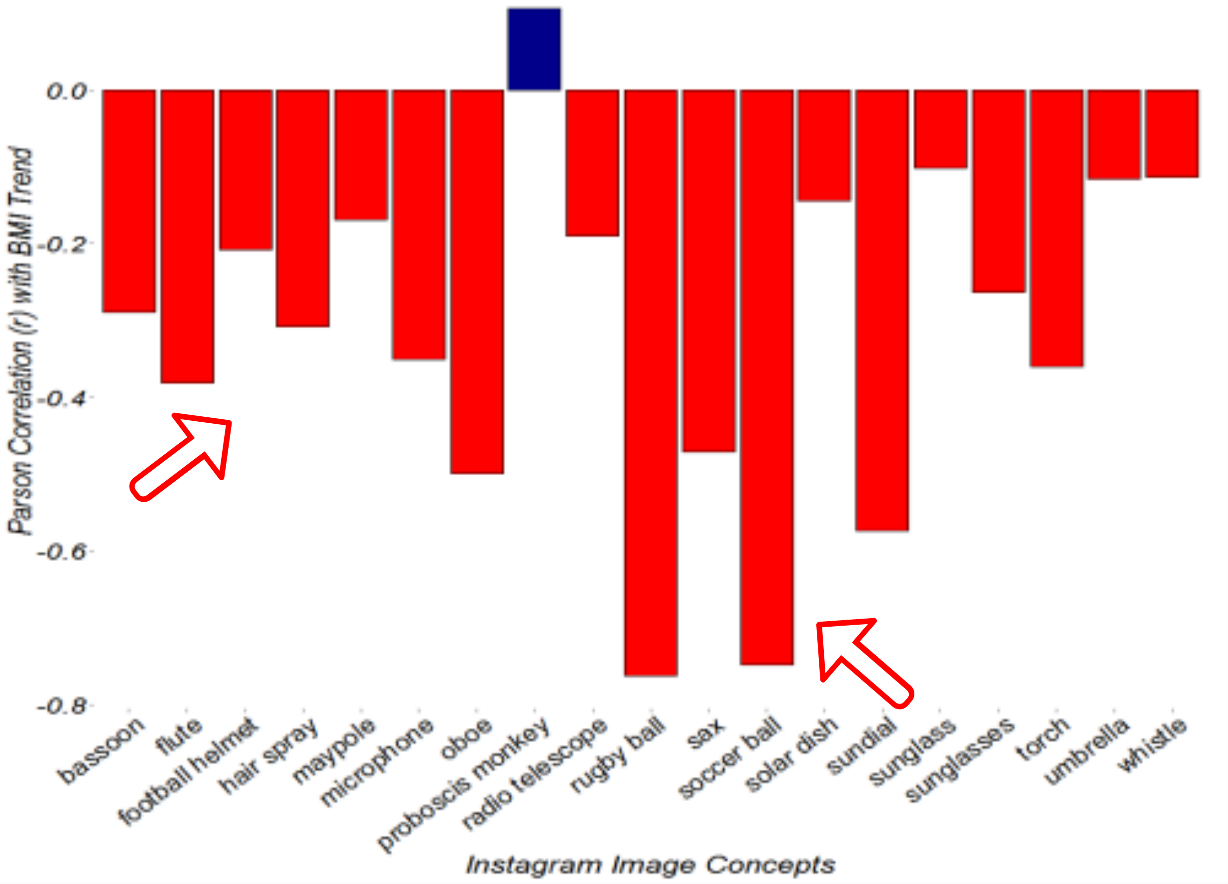


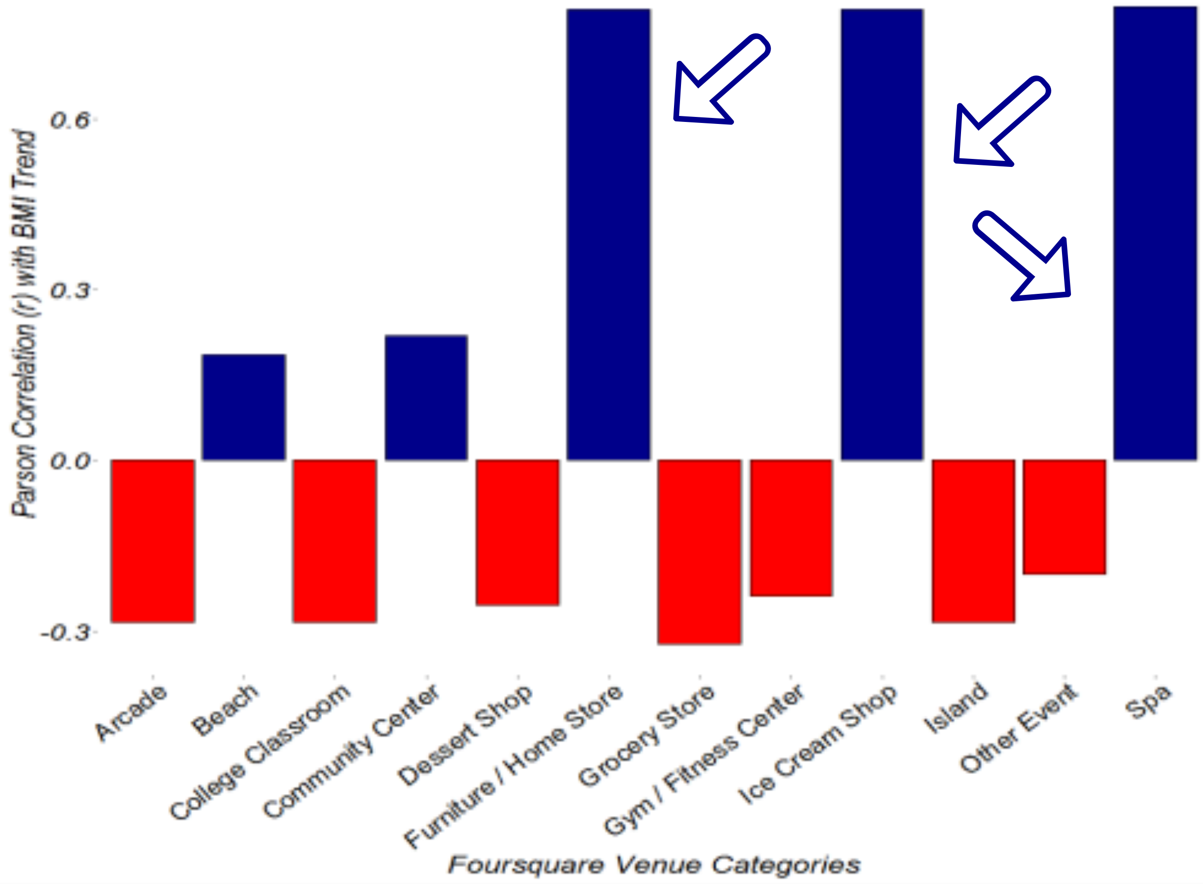
Examples of Wellness Correlations (4): Individual Profiling



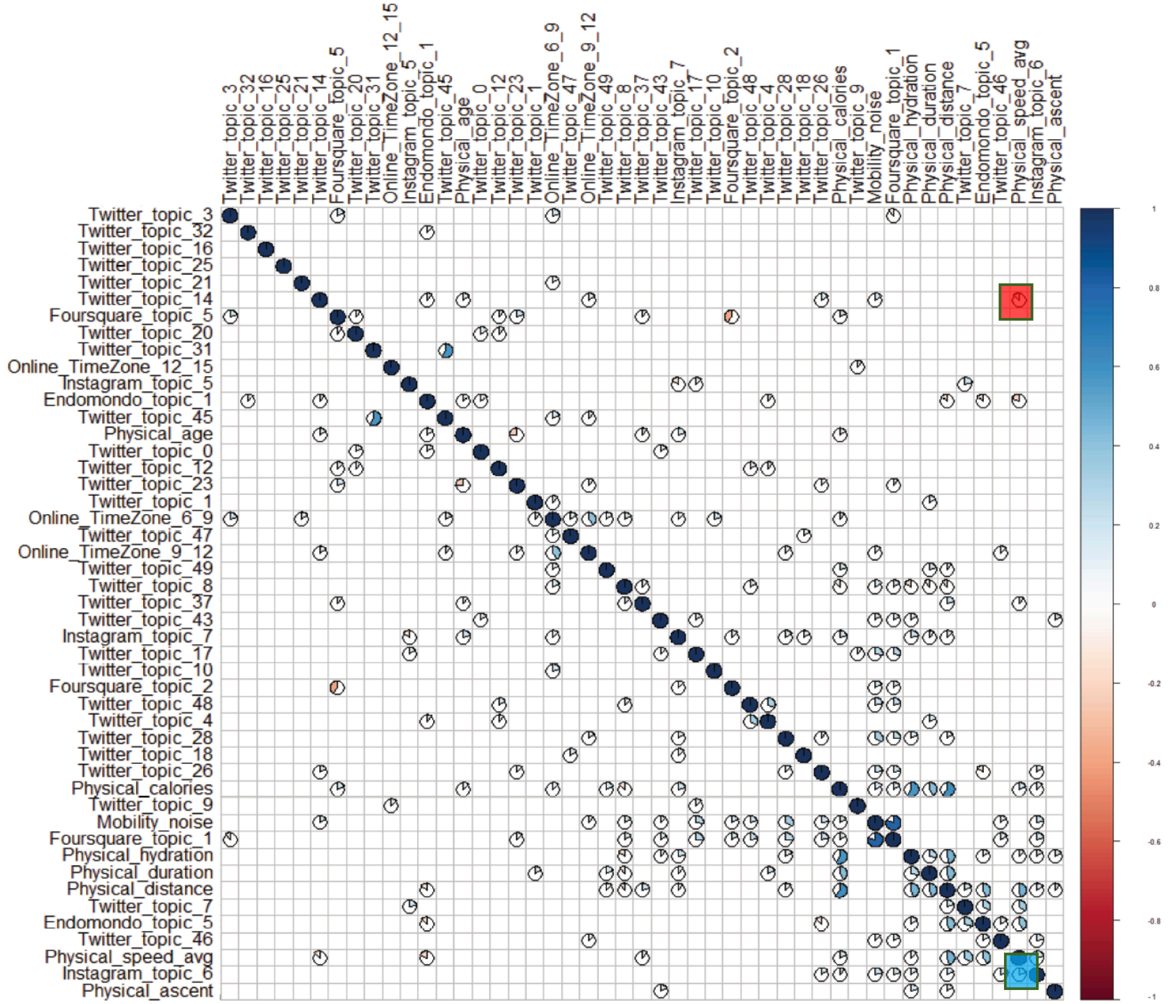
E-Mail: farseev@u.nus.edu | Website: http://farseev.com

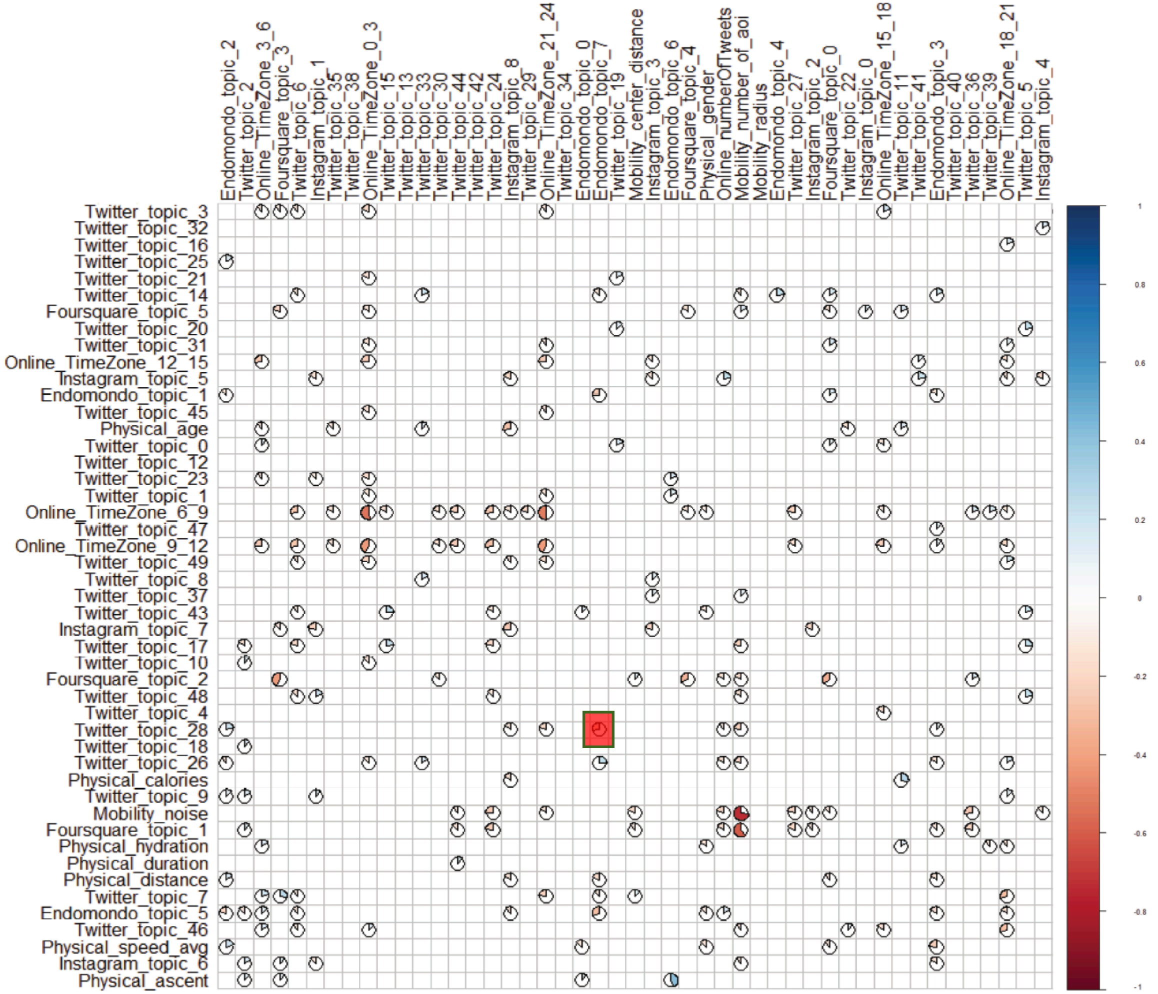
Examples of Wellness Correlations (5): Individual Profiling



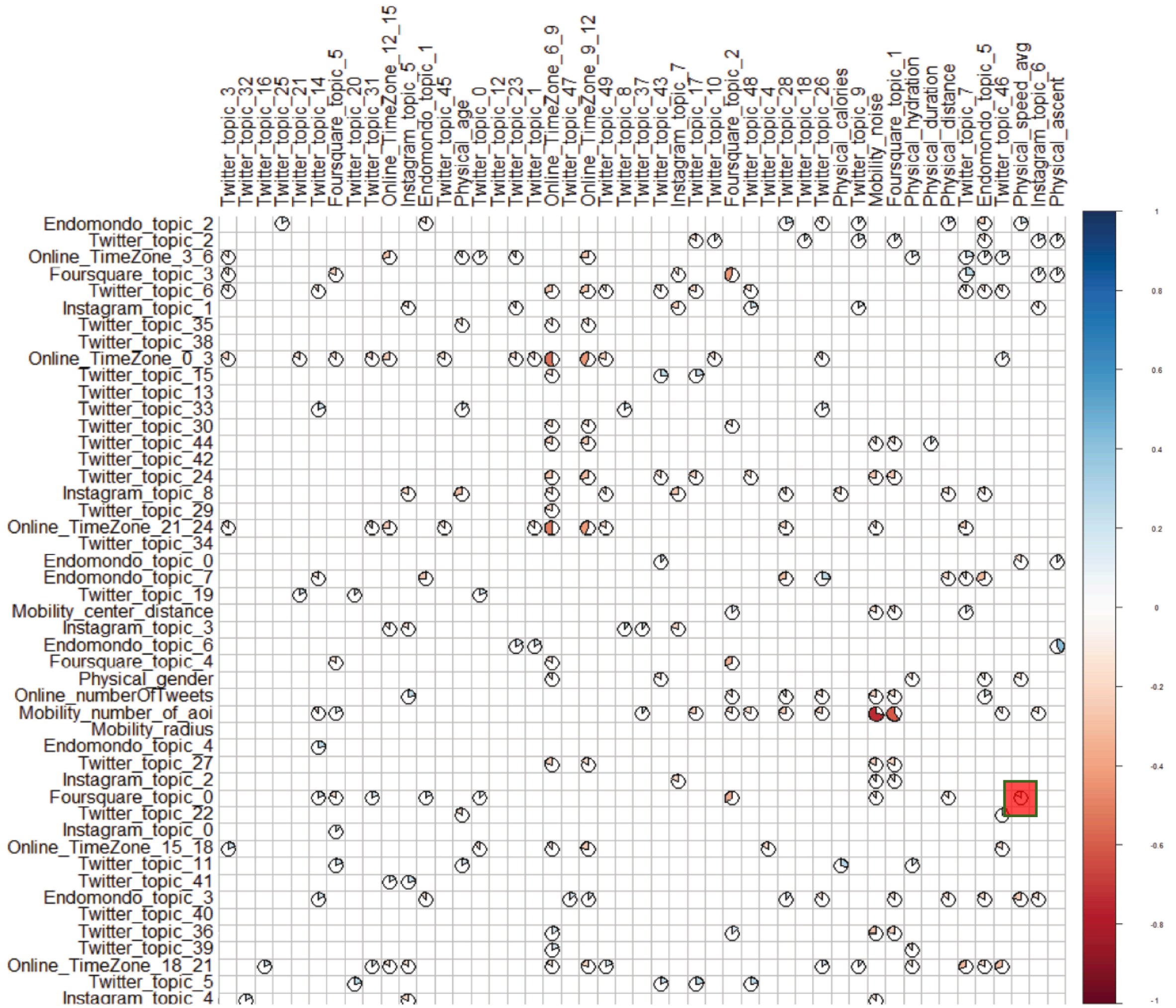


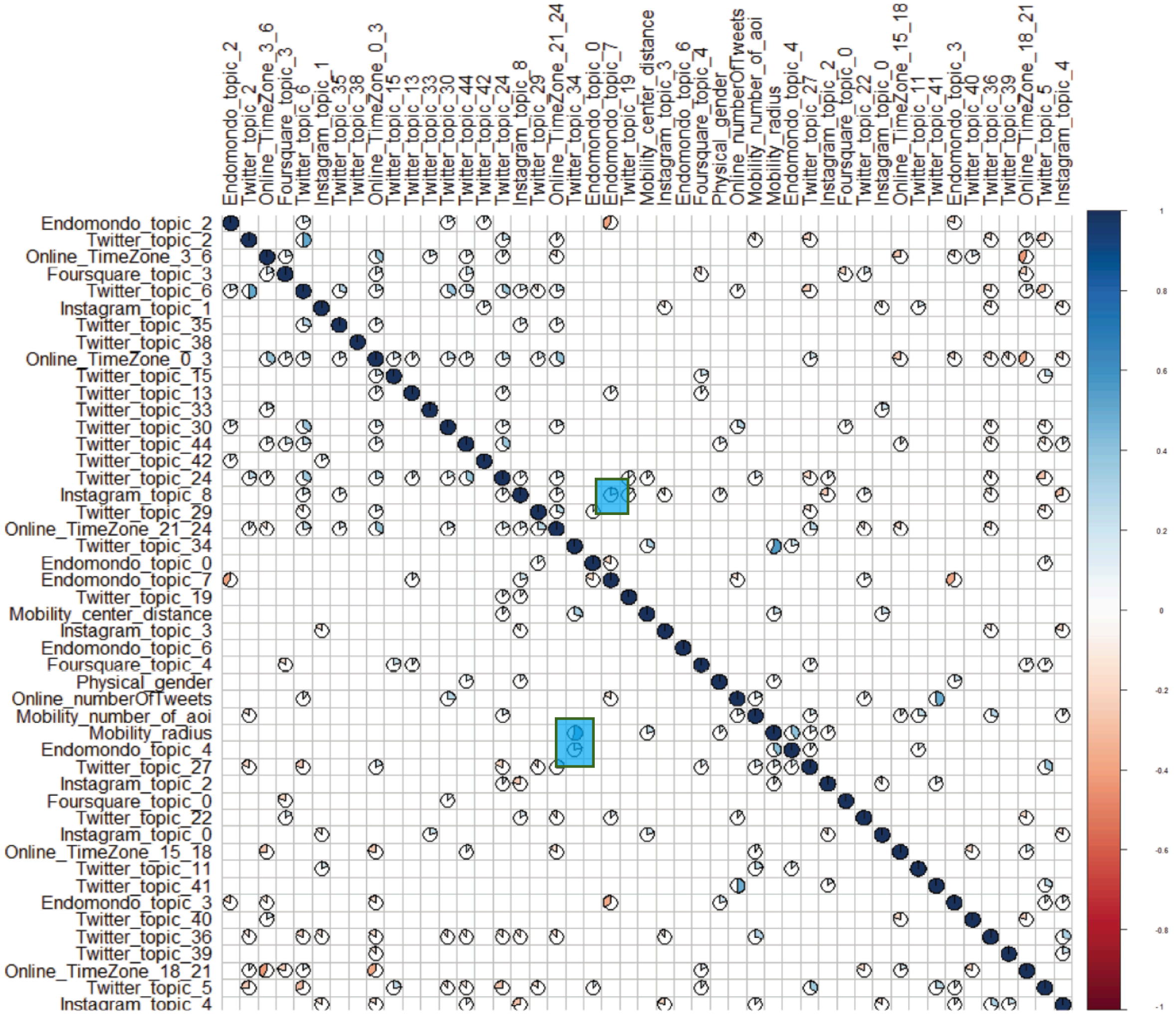
Examples of Inter-Source Correlations (I): Individual Profiling





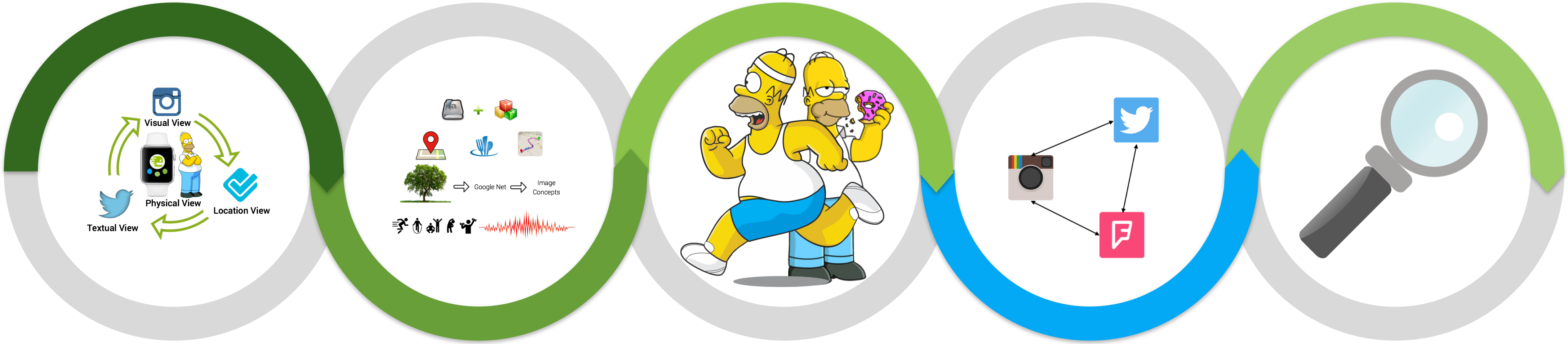
Examples of Inter-Source Correlations (II): Individual Profiling





E-Mail: farseev@u.nus.edu | Website: http://farseev.com

Main takeaways



Multi-Source Data Describes Users From Multiple Perspectives

Learning From Multi-Source Data Allows For Achieving Significant Individual User Profiling Performance Improvement

Join Learning From Social Media Data And Sensor Data Helps To Improve Wellness Profiling Performance

Considering Inter-Source Relationship is Useful For Group User Profiling

Qualitative Research On Multi-Source Data Relations is Promising

List of Involved publications

Buraya, K., Farseev, A., Filchenkov, A., & Chua, T. S. (2017 February). **Towards User Personality Profiling from Multiple Social Networks.** In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI.

Nie, L., Zhang, L., Wang, M., Hong, R., Farseev, A., & Chua, T. S. (2017). **Learning user attributes via mobile social multimedia analytics.** ACM Transactions on Intelligent Systems and Technology (TIST), 8(3), 36.

Farseev, A., Nie, L., Akbari, M., & Chua, T. S. (2015, June). **Harvesting multiple sources for user profile learning: a big data study.** In Proceedings of the 5th ACM Conference on Multimedia Retrieval (pp. 235-242). ACM.

Farseev, A., Kotkov, D., Semenov, A., Veijalainen, J., & Chua, T. S. (2015, June). **Cross-social network collaborative recommendation.** In Proceedings of the 7th ACM Conference on Web Science (p. 38-39). ACM.

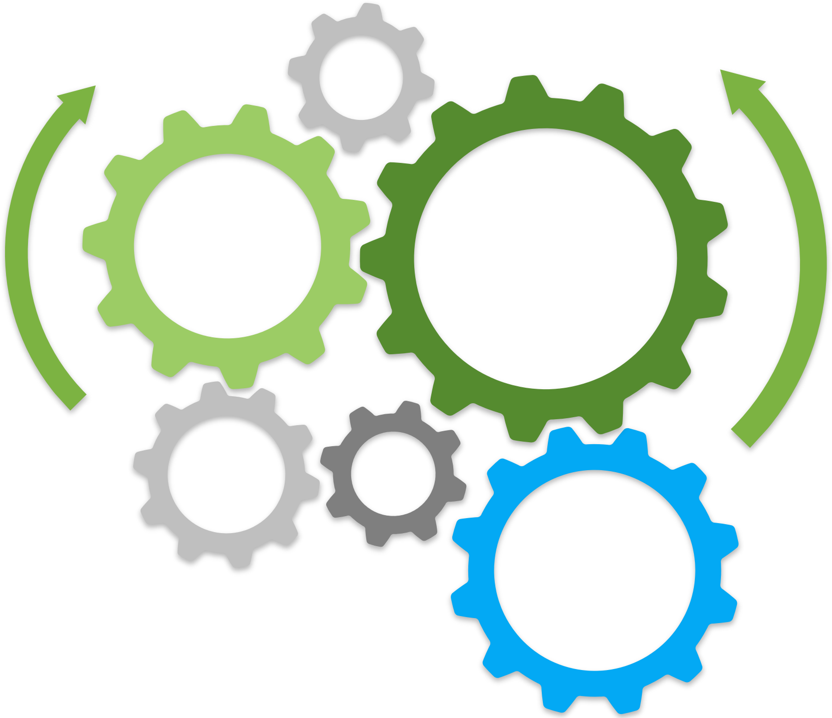
Farseev, A., Akbari, M., Samborskii, I., & Chua, T. S. (2016). 360° **user profiling: past, future, and applications.** by Aleksandr Farseev, Mohammad Akbari, Ivan Samborskii and Tat-Seng Chua with Martin Vesely as coordinator. ACM SIGWEB Newsletter, (Summer), 4.

Farseev, A., Samborskii, I., & Chua, T. S. (2016, October). **bBridge: A Big Data Platform for Social Multimedia Analytics.** In Proceedings of the 25th ACM Conference on Multimedia (pp. 759-761). ACM.

Farseev, A., & Chua, T. S. (2017 February). **TweetFit: Fusing Multiple Social Media and Sensor Data for Wellness Profile Learning.** In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (pp. 87-93). AAAI.

Farseev, A., & Chua, T. S. (2017). **Tweet can be Fit: Integrating Data from Wearable Sensors and Multiple Social Networks for Wellness Profile Learning.** ACM Transactions on Information Systems (TOIS).

Farseev, A., Samborskii, I., Filchenkov, A., & Chua, T. S. (2017 August). **Cross-Domain Recommendation via Clustering on Multi-Layer Graphs.** In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.



DATASETS

NUS-MSS

NUS-MSS – the largest available multi-source cross-region dataset.

Includes:

- **Foursquare** as a location data source;
- **Twitter** as a textual data source;
- **Instagram** as a visual data source;
- **Facebook** as a demographics-related ground truth source.

nusmultisource.azurewebsites.net

City	#users	#tweets	#check-ins	#images
Singapore	7,023	11,732,489	366,268	263,530
London	5,503	2,973,162	127,276	65,088
New-York	7,957	5,263,630	304,493	230,752

Ground Truth: Age, Education, Employment, Demography, Location, Relationship Status.

E-Mail: farseev@u.nus.edu l Website: http://farseev.com

NUS-SENSE

NUS-SENSE: The largest available multi-lingual dataset with personality ground-truth and Twitter IDs.

Includes:

- **Foursquare** as a location data source;
- **Twitter** as a textual data source;
- **Instagram** as a visual data source;
- **Endomondo** as a sensor data source;

nussense.azurewebsites.net

#users	#tweets	#check-ins	#images	#workouts
5,375	1,676,310	19,743	48,137	140,926

Ground Truth: BMI, BMI Trend, Exercise Type, Gender, Age

NUS-PERSONALITY

NUS-PERSONALITY:: The largest available multi-lingual dataset with personality ground-truth and Twitter IDs.

Includes:

- **Foursquare** as a location data source;
- **Twitter** as a textual data source;
- **Instagram** as a visual data source;

farseev@gmail.com

#users	#tweets	#check-ins	#images
18,164	22,255,445	420,603	4,465,565

Ground Truth: MBTI, Gender

NUS-GAMING

NUS-PERSONALITY:: The largest available multi-lingual dataset with personality ground-truth and Twitter IDs.

Includes:

- **Steam** users and their individual attributes

farseev@gmail.com

#users	#friends (avg.)	#favorite games
23,375	63	8

Ground Truth: Game Involvement, Achievements, friends, communication, game selection, etc.

Wellness Events in Microblogs

Wellness Events in Microblogs:: Personal Daily Events Detection

Includes:

- **Twitter** as a textual data source;

farseev@gmail.com

#users	#tweets
1, 987	13, 552

Ground Truth:

Diet (6), Exercise (5), Health (3) events, Diabetes

Thank You

Questions?

Backup I: Optimization of Multi-Source Multi-Task Objective

Smooth Reformulation of the objective:

$$f(\mathbf{W}) = \underset{\mathbf{W} \in \mathbf{Z}}{\arg\min}\ \Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \mu\Omega(\mathbf{W})$$

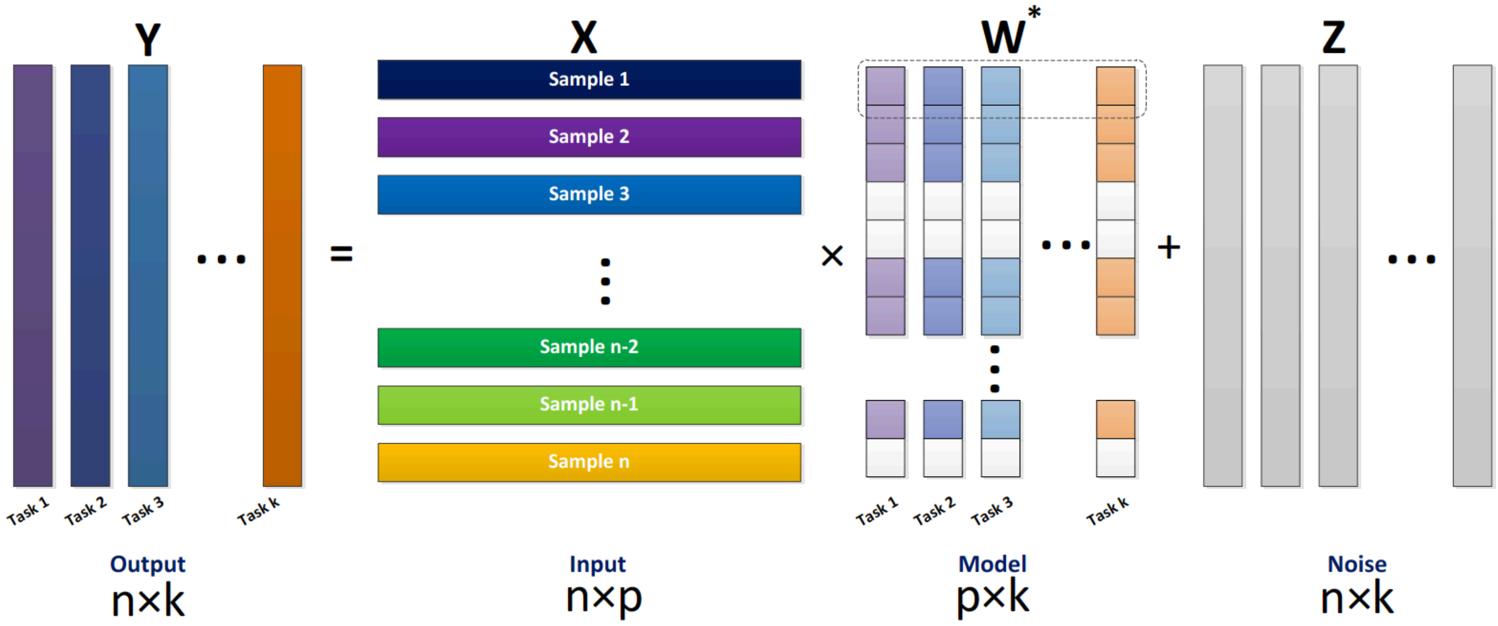
$$s.t.\ \mathbf{Z} = \left\{ \mathbf{W} \mid \|\mathbf{W}\|_{2,1} \leq z \right\},$$

Nesterov's Optimization Solution on Each Step, S_i computed from past solutions:

$$\mathbf{W}_{i+1} = \underset{\mathbf{W}}{\arg\min}\ M_{\gamma_i, S_i}(\mathbf{W}),$$

$$M_{\gamma_i, S_i}(\mathbf{W}) = f(\mathbf{S}_i) + \langle \nabla f(\mathbf{S}_i), \mathbf{W} - \mathbf{S}_i \rangle + \frac{\gamma_i}{2} ||\mathbf{W} - \mathbf{S}_i||^2, \quad \mathbf{S}_i = \mathbf{W}_i - \alpha_i(\mathbf{W}_i - \mathbf{W}_{i-1}).$$

Backup II: MTL with Joint Feature Learning -2



$$\min_{W} \frac{1}{2} \|XW - Y\|_F^2 + \lambda \sum_{i=1}^{p} \|\boldsymbol{w}_i\|_2$$

Using group sparsity: ℓ1/ℓ2-norm regularization

Backup III: Wellness Profiling Baselines

Wellness Profiling Baselines

- **aMTFL$_2$** [85] — the $\ell_{2,1}$ norm regularized multi-task learning with the least squares lost and $\alpha = 0.5$.
- **iMSF** [160] — the sparse $\ell_{2,1}$ norm regularized multi-source multi-task learning, with $\alpha = 0.4$;
- **MSESHC** — weighted ensemble proposed in [47]. The modality weights $\mathbf{s}$ were learned by Stochastic Hill Climbing (SHC): $\mathbf{s} : \{0.75, 0.2, 0.25, 0.45, 0.2, 0.3, 0.45, 0.2\}$; for venue categories, image concepts, behavioral text, LDA 50 text, sport categories, sensors freq. bins, workout statistics, and mobility features, respectively.
- **TweetFit** [44] — multi-source multi-task learning framework "TweetFit" (equivalent to **M^{2}WP** framework, trained without inter-category relatedness regularization (Equation 4.1)).

[44] A. Farseev and T.-S. Chua. Tweetfit: Fusing multiple social media and sensor data for wellness profile learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI, 2017.

[47] A. Farseev, L. Nie, M. Akbari, and T.-S. Chua. Harvesting multiple sources for user profile learning: a big data study. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. ACM, 2015.

[85] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009.

[160] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2012.

Backup IV: Detected User Communities

Name	Bag of Words for different modalities		
	Text	**Visual**	**Location**
Gadgets 832 users	device, launcher, android	mouse, digital clock, hard disc	electronics store, tech startup, technology building
Arts 538 users	painting, landscape, reflection	obelisk, paint-brush, pencil box	arts & crafts store, arts & entertainment, museum
Food 446 users	dining, cof-fee, cook-ing	pineapple, mi-crowave, fry-ing pan	italian restaurant, pizzeria, macanese restaurant

Backup V: Recommendation Evaluation Metrics

Average Precision

$$AP@p = \frac{1}{\sum_{i=1}^{p} r_i} \sum_{i=1}^{p} r_i \left(\frac{\sum_{j=1}^{i} r_j}{i} \right), r_i = \begin{cases} 1, & \text{item } i \text{ is relevant} \\ 0, & \text{otherwise.} \end{cases}$$

Normalized Discounted Cumulative Gain

$$NDCG@p = \frac{DCG@p}{IDCG@p}, DCG@p = \sum_{i=1}^{p} \frac{2^{rel_i}}{\log_2(i+1)}, rel_i = \frac{Cat_i}{N_{Cat}}$$