ITMO UNIVERSITY

SoMin

Demystifying Psychographic Marketing

Multi-View Learning as a New Social Media User Profiling Standard

by Aleksandr Farseev

http://farseev.com

http://somin.ai

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

Multiple social networks describe user behavior from multiple views

Some facts about social networks…

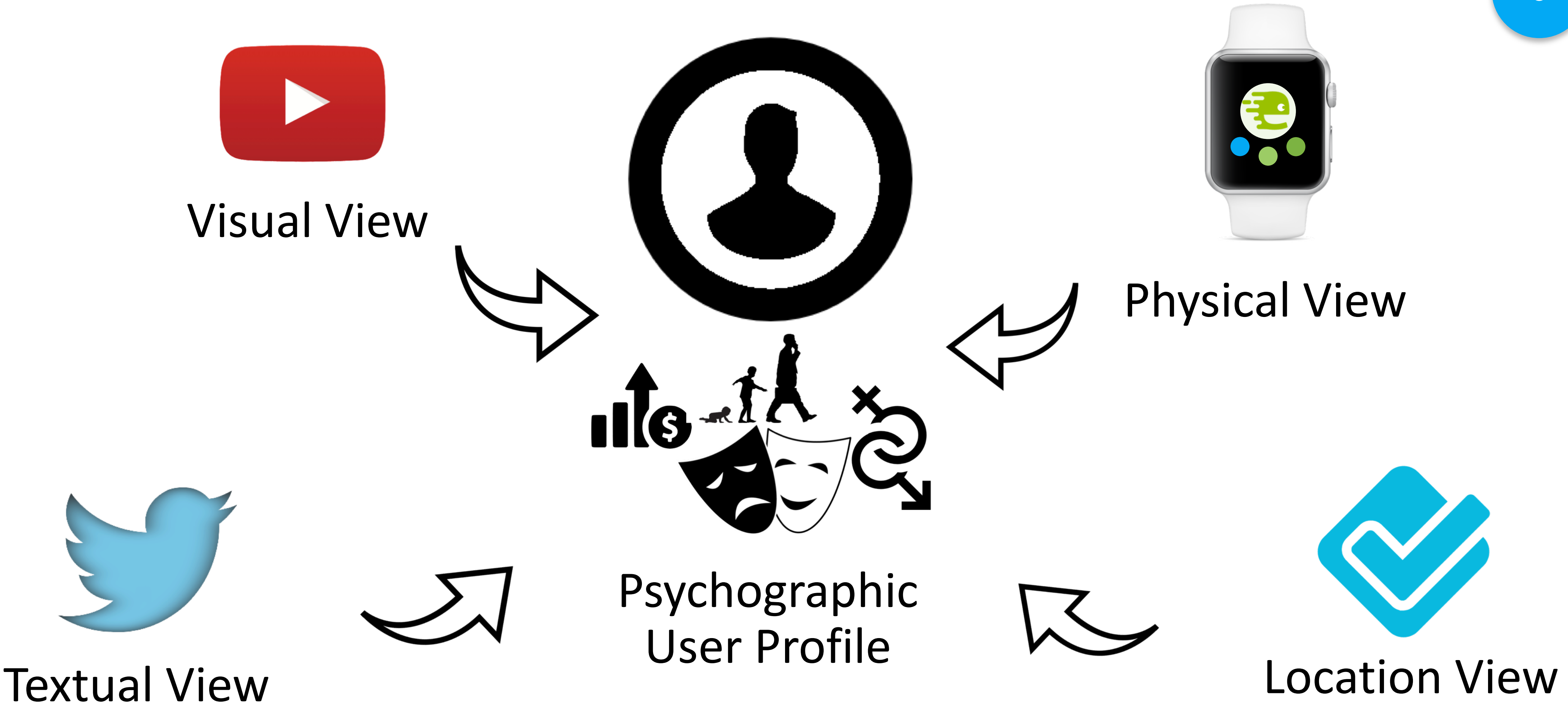
More than 50% of online-active adults use more than three social networks in their daily life*

*According Paw Research Internet Project's Social Media Update 2017 (www.pewinternet.org/fact-sheets/social-networking-fact-sheet/)

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

Different data modalities describe users from multiple views

Indeed, they are:



Psychographic profiling in our works

Those attributes that we've inferred

360° User Profile

Group Profile

Individual Profile

User Communities

Wellness profile

Identity profile

Diabetes

Asthma

Obesity

Age

Gender

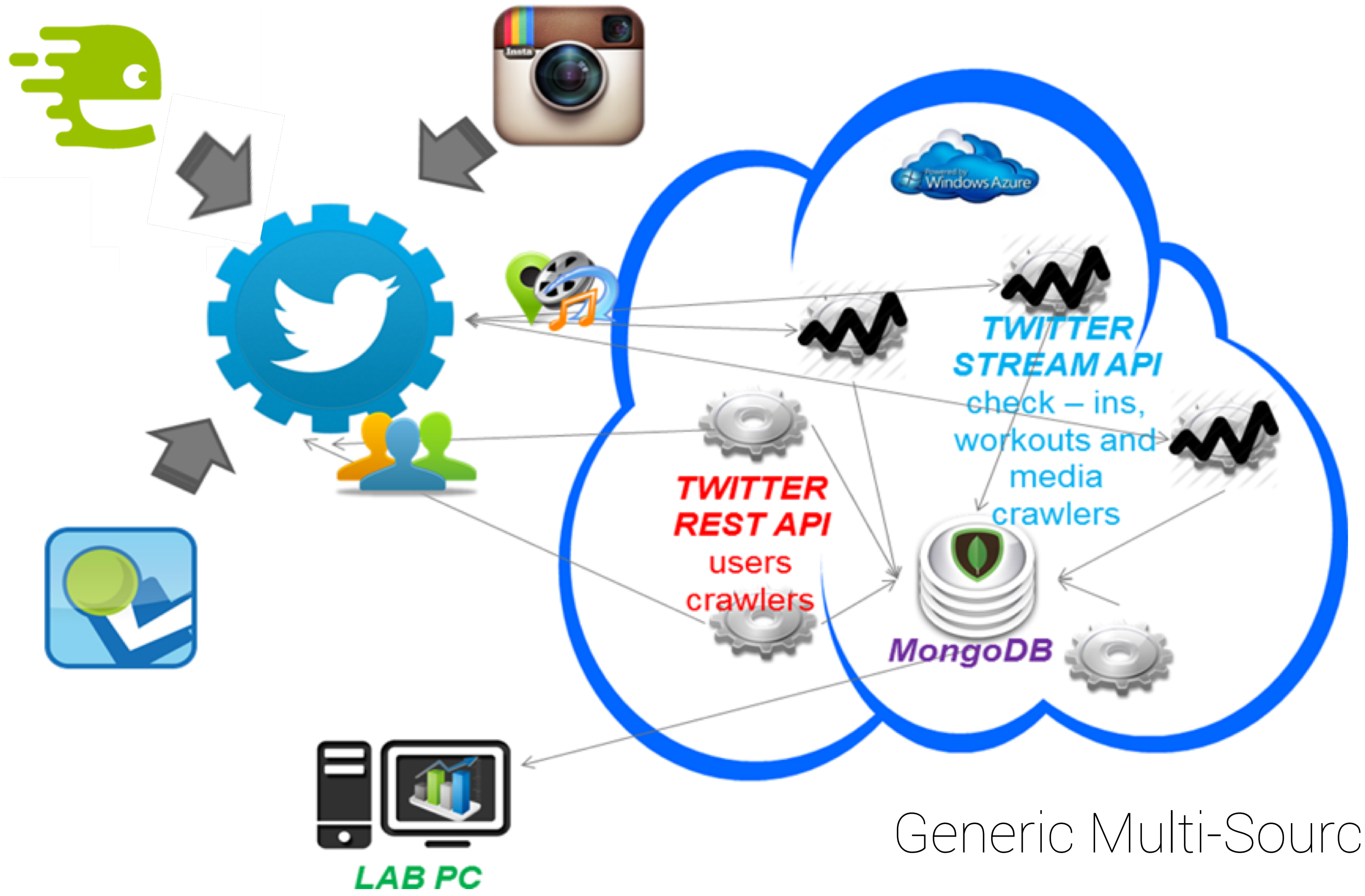
Personality

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.

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.

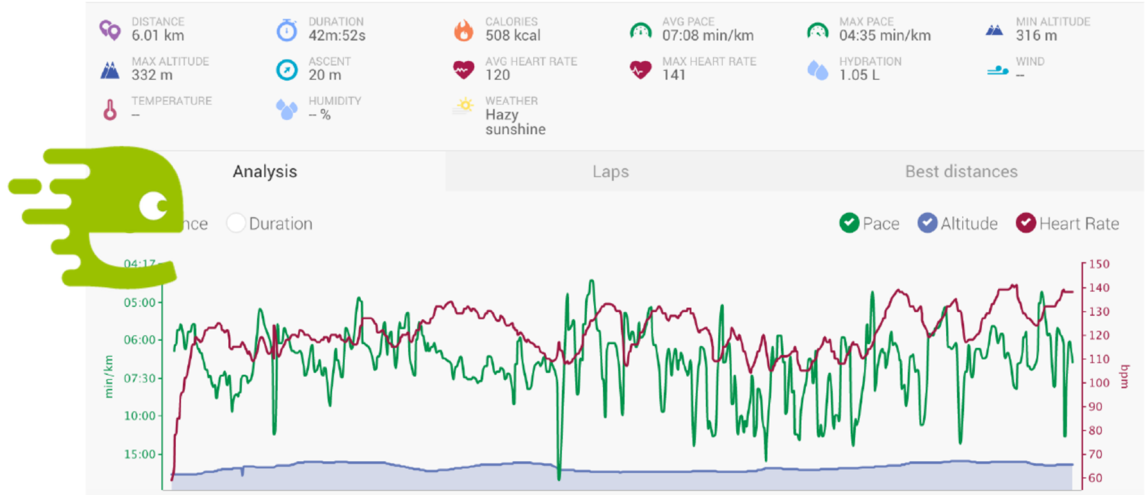
Cross-network ambiguity is resolved after collection of the first cross-network post.

Generic Multi-Source Data Gathering Approach

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

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

Data Representation: Summary

All data types together

TEXT Features:

Linguistic features: LIWC; Latent Topics

Heuristic features: Writing behavior

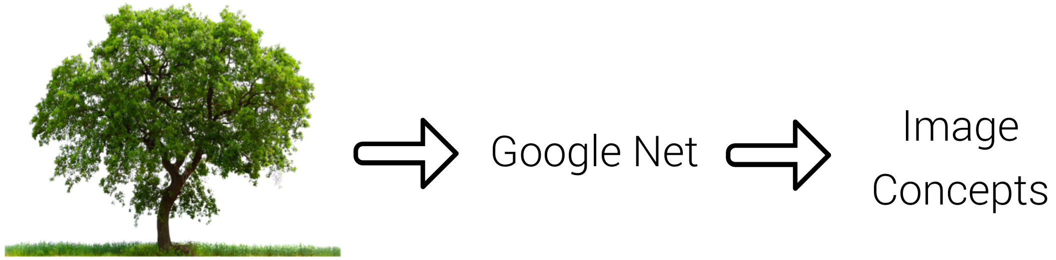
Location Features:

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

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

Our released large multi-source multi-modal datasets

NUS-SENSE
http://nussense.azurewebsites.net

Location	#users	#tweets	#check-ins	#images	#check-ins
Worldwide	5,375	16,763,310	19,743	48,137	140,926

NUS-MSS
http://nusmss.azurewebsites.net

Location	#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

Data was voluntarily publicly released by Twitter users and collected via official Twitter API
Datasets are released in a form of features thus user privacy is not affected.

Two Large Multi-Source Social Media & Sensor Datasets

Individual Multi-View Learning

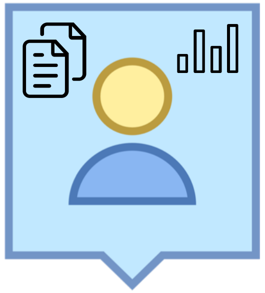
Part I: Demographic Profiling

*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.

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

On cross-domain 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

Research Questions

Question One

1

Is it possible to boost supervised machine learning for individual user profiling performance by incorporating multi-modal data from multiple social networks?

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

Contributions…

The First Work
On Multi-Source
Individual User Profiling via
Late Fusion

01

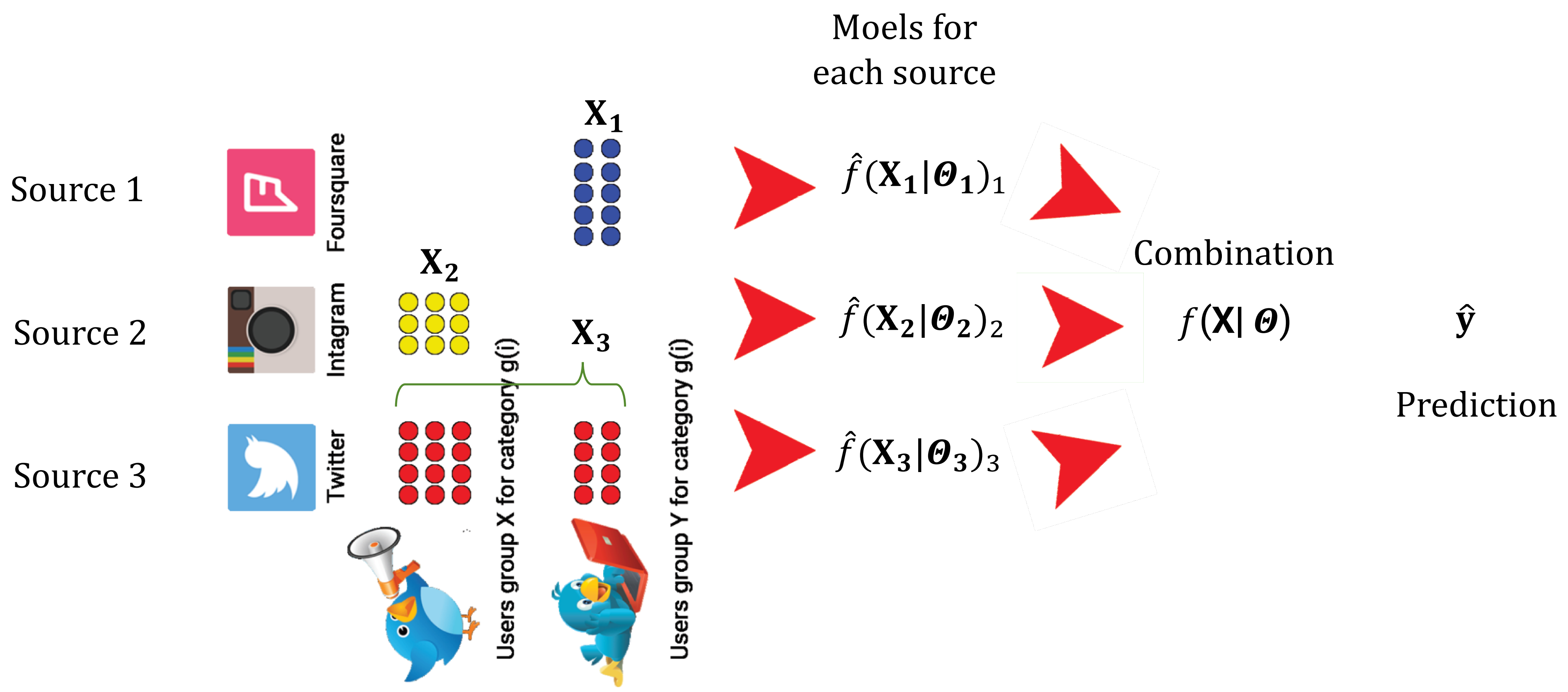
Methodology for Multi-Source Data Gathering
via Cross-posting for Arbitrary number
of Social Networks

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.

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

Intuition behind late-fused multi-source learning



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

Age and Gender Prediction

Running Random Forests With Random Restart

Sources 1..S

Random Forests for each S

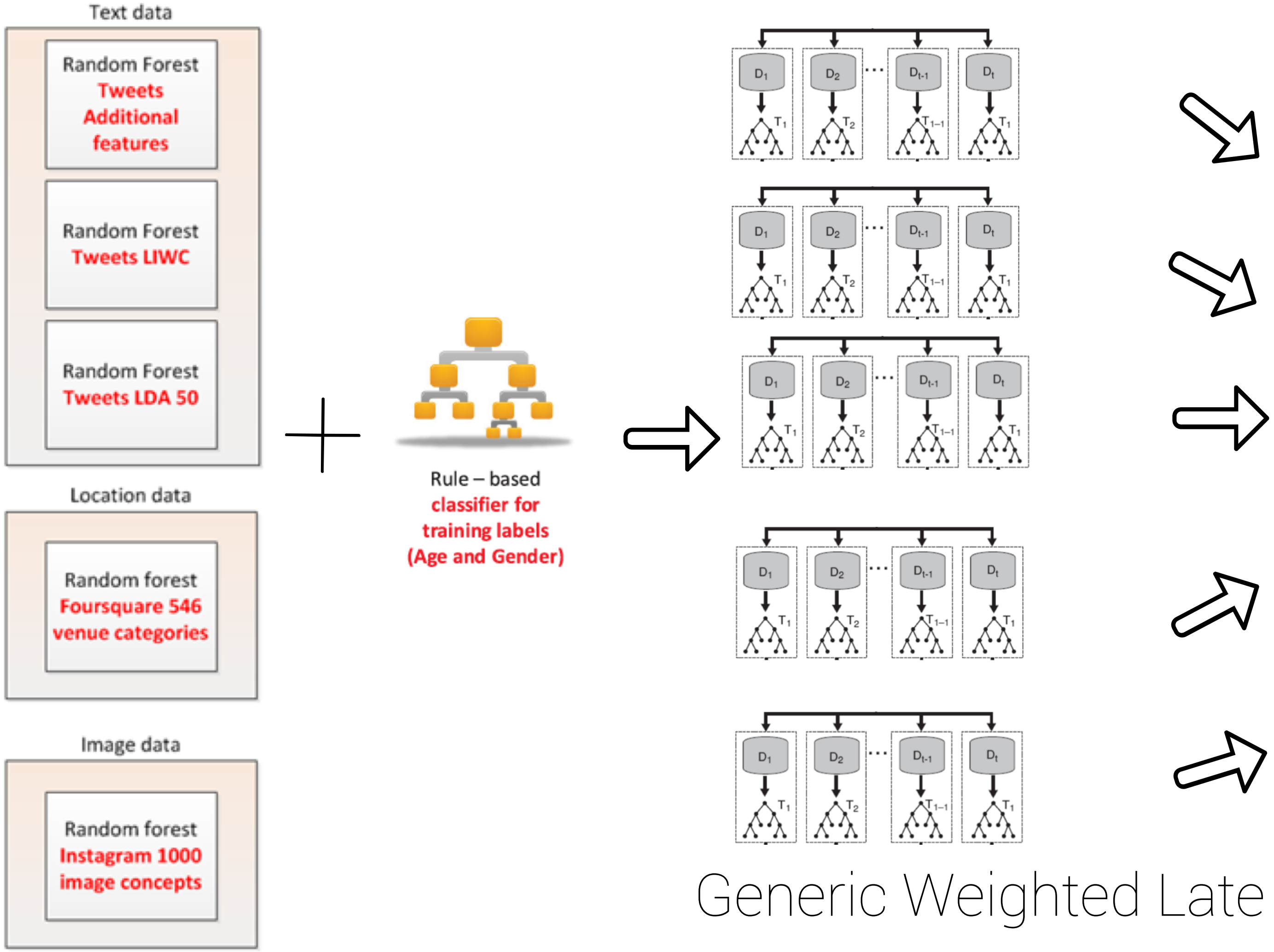
s_1

s_2

s_3

s_4

s_S



Weighted voting

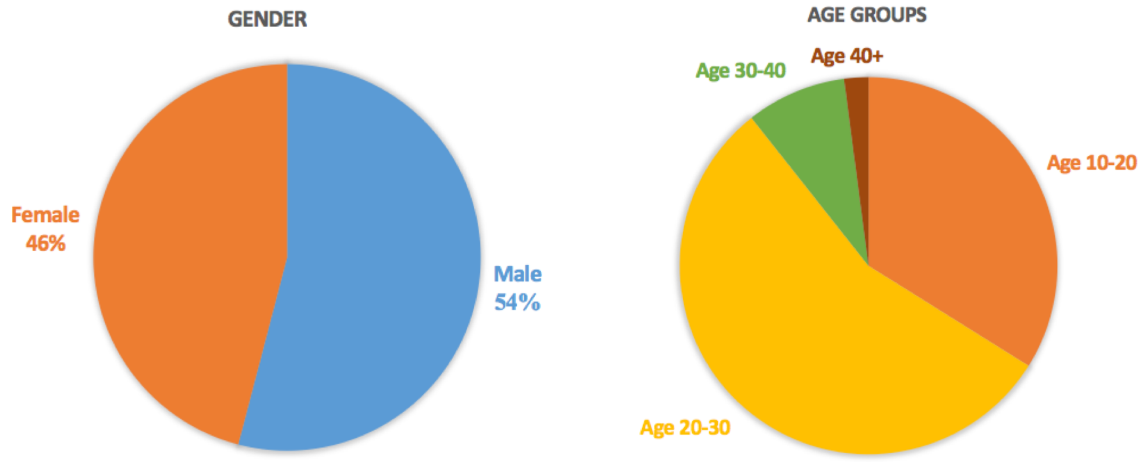
$$f_i(\mathbf{X}) = \frac{1}{S}\sum_{s=1}^{S} \hat{f}_i(\mathbf{X}_s)_s \times w_s$$

$\hat{f}_i(\mathbf{X}_s)_s$ - s-th model prediction
w_S - s-th view weight obtained by Stochastic Hill Climbing

Generic Weighted Late Fusion Approach

Age and Gender Ground Truth (NUS-MSS)

Attribute	Train (Age was Estimated from Education Path)	Test (Real Age Mentions)
Gender		
Male	2536	129
Female	2155	93
Age Groups		
10-20	360	181
20-30	589	28
30-40	91	8
40+	22	5



Note: Age ground truth is small
Solution: estimated age ground truth from users' Education and Occupation history

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

Baselines

Method	Gender	Age
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
Statistical Analysis		
Weighted Cohen's Kappa	$\kappa_w = 0.745, p < 0.01$	$\kappa_w = 0.297, p < 0.01$

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.

Individual Multi-View User Profiling

Part II: Wellness Profiling

*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.

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 l Website: http://farseev.com

Research Questions

1

Question One

Is it possible to improve the performance of BMI category and "BMI Trend" inference by fusing multiple social media and sensor data?

2

Question Two

What is the contribution of sensor data towards BMI category and "BMI Trend" inference?

3

Question Three

Is it possible to improve the performance of BMI category and "BMI Trend" inference by incorporating inter-category relatedness into the learning process?

Contributions

Generic Model
For Supervised Joint Learning
From Multi-Source
Multi-Modal Incomplete Data

01

First Social-Sensor Dataset
NUS-SENSE

02

*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).

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

Unite Social Media And Wearable Sensors For Physical Attributes Inference

Just tweet to be fit….

FOURSQUARE

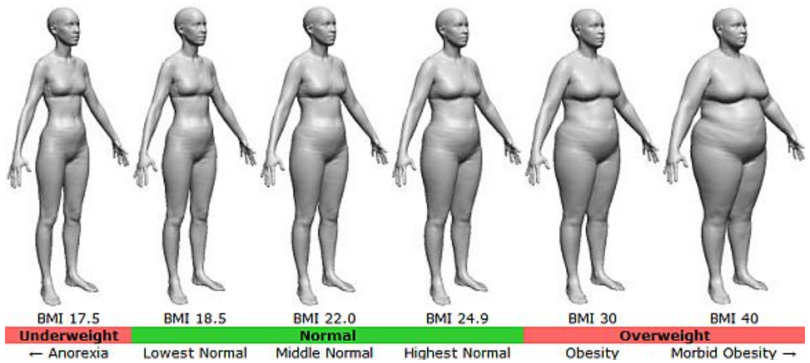
twitter

Instagram

\+

Weight Fluctuation Trend (BMI Trend)

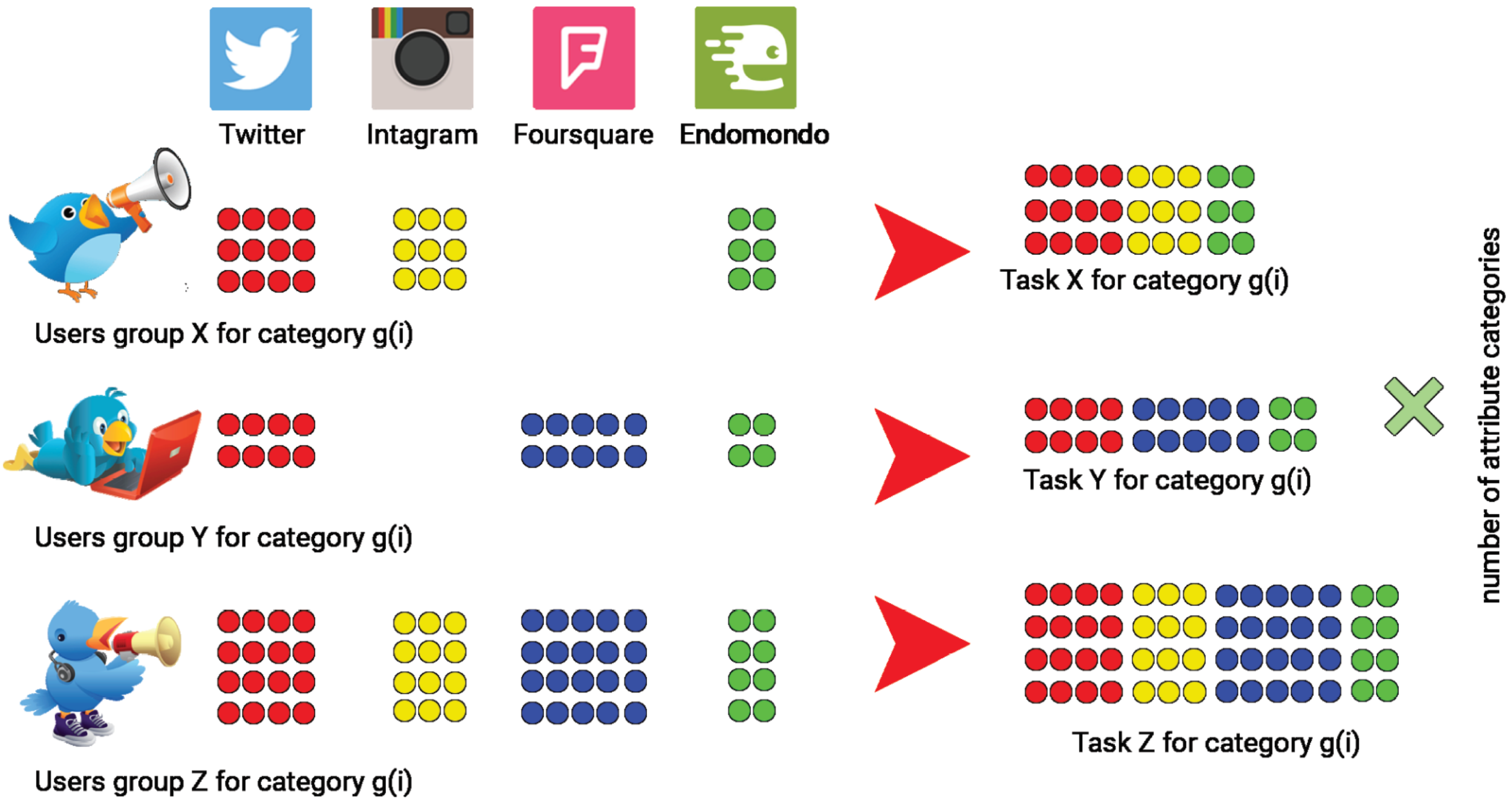
\+



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

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

Multi-Source Multi-Task Learning



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)

Generic Multi-View Hybrid Fusion Approach ☆

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

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

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

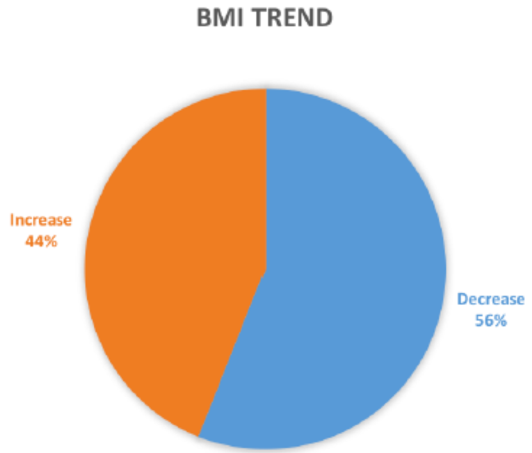
Inter-Category Smoothness Regularization ☆

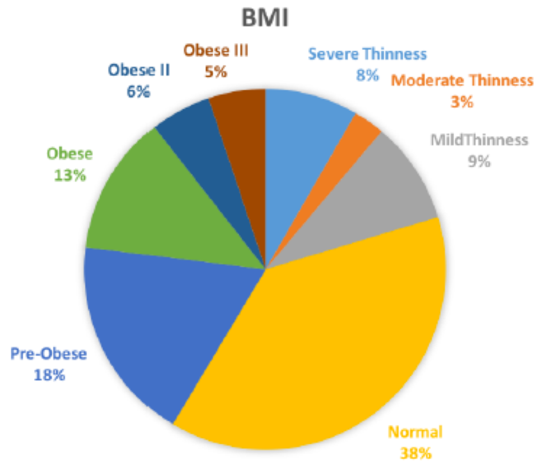
$$\Omega(\mathbf{W}) = \sum_{t=1}^{\hat{T}} \sum_{g \in C_{D_t}} \kappa_{g,\xi(t,g)} \left\| \mathbf{w}_g^t - \mathbf{w}_{\xi(t,g)}^t \right\|^2$$

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

BMI and BMI Trend Ground Truth (NUS-SENSE)

Attribute	Train	Test
BMI Trend		
Decrease	67	16
Increase	53	11
BMI		
Severe Thinness	71	16
Moderate Thinness	24	6
Mild Thinness	80	18
Normal	331	76
Pre Obese	157	36
Obese I	105	25
Obese II	47	11
Obese III	45	9





Note: data for some categories is not large.
Solution: applied SMOTE oversampling

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

BMI Category and BMI Trend Prediction: Results (1)

Data Source Combinations

Data Source Combination	BMI category prediction	
	R_{Mac}/P_{Mac}	$F_{1,Mac}$
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		"BMI Trend"	
	R_{Mac}/P_{Mac}	$F_{1,Mac}$	R_{Mac}/P_{Mac}	$F_{1,Mac}$
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

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

Source Importance Analysis: Feature level

Different feature types

Feature type	S. Th.	M. Th.	Md. Th.	Nrm.	P-Ob.	Ob. I	Ob. II	Ob. III
Latent topics	2	5	0	0	1	2	4	1
Lexicon	4	3	3	1	0	2	1	0
Writing style	2	1	1	1	0	2	1	3
Image con.	25	5	4	1	3	7	9	4
Venue sem.	19	5	1	2	11	5	11	2
Mob. & Tmp.	3	4	3	1	2	4	4	2
Work. sem.	1	0	1	0	1	2	2	2
Freq. domain	15	8	19	10	18	17	25	13
Work. stat.	1	2	2	1	1	2	2	3

1. **Text features are less useful as compared to others (consistent with cross-source experiment).**

2. **Image features are more helpful in distinguishing weight problems (abnormal BMI categories).**

3. **Venue categories (semantics) are more powerful for the whole BMI scale as compared to geographical mobility patterns.**

4. **Temporal workout features, are the most useful and absolutely necessary, while the type of exercise as well as exercise statistics play auxiliary roles.**

User Profiling Analytics

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

Research Questions

Question One

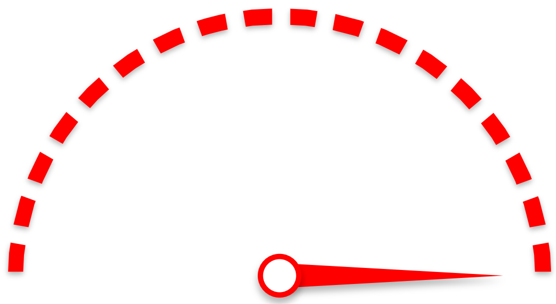
What is the relation between different data modalities, data sources, and individual user attributes?

1

Contributions

First study on Cross-Modal Statistical Analysis of Users from Multiple Social Networks

01



*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).

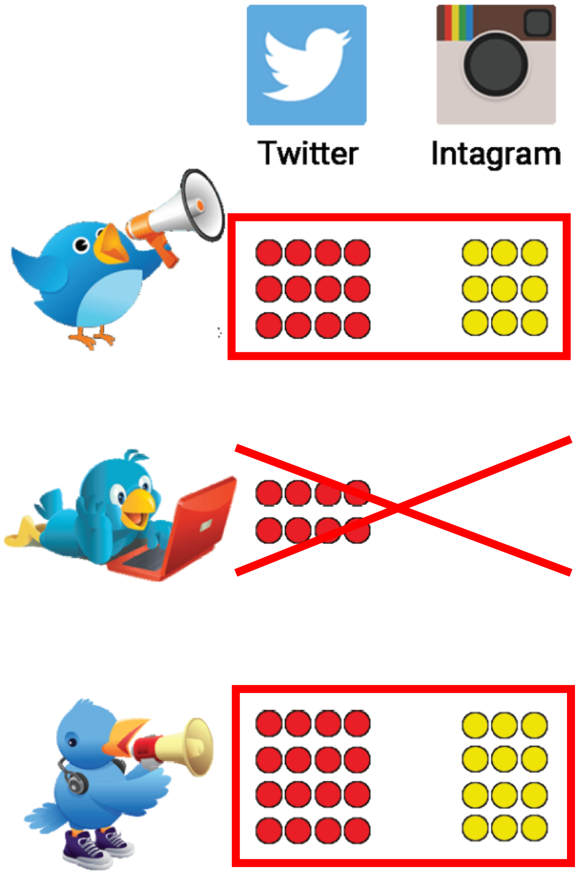
E-Mail: farseev@u.nus.edu I Website: http://farseev.com

Pearson Correlation to Visualize Significant Data Relationships

Sample correlation coefficient r – an estimate of the unknown correlation coefficient for a representative sample of size n:

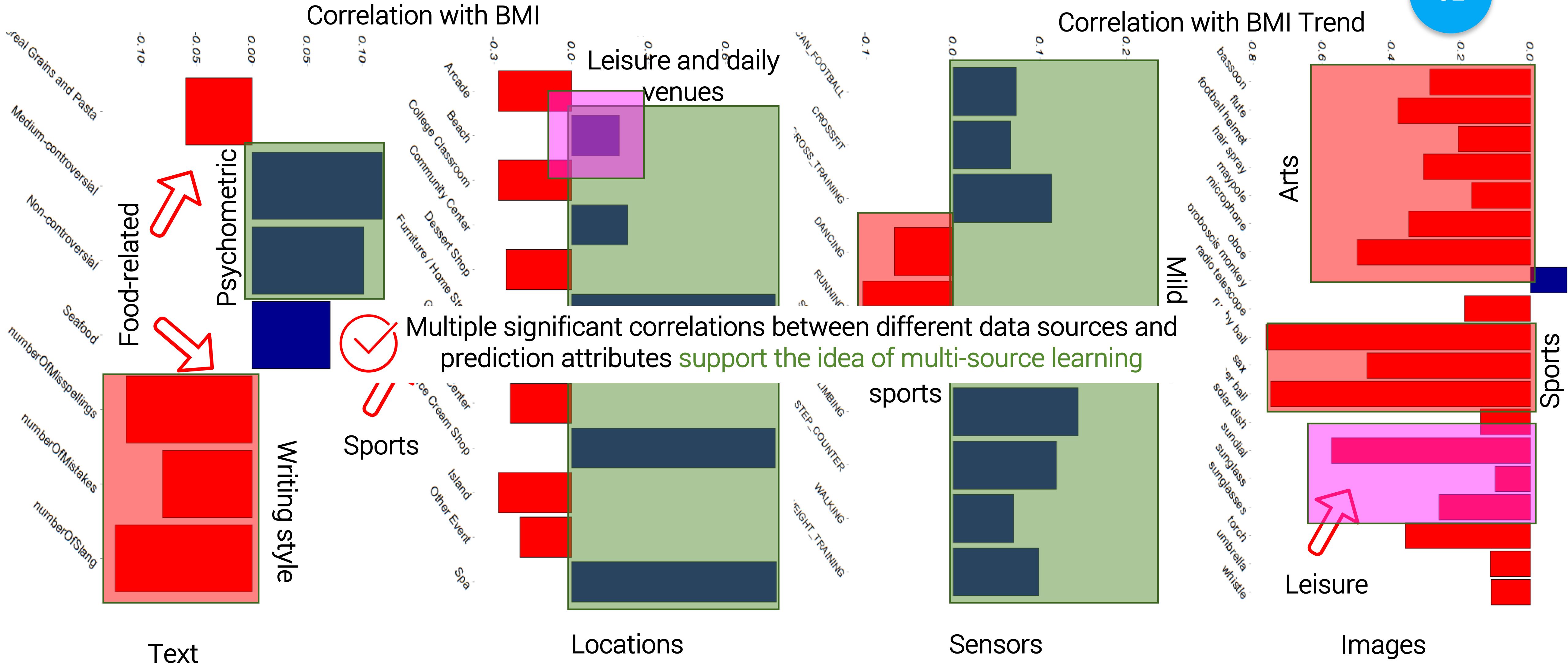
$$r = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

where x_i, y_i are i-th population samples and $\overline{x}$, $\overline{y}$ are the population means.



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

Individual Profile Analytics: Correlation with individual attributes



Multiple significant correlations between different data sources and prediction attributes support the idea of multi-source learning

Future Of User Profiling

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

Future of User Profiling

User profiling to be approached from Data and Modeling Perspectives

Future of User Profiling

Application Perspective		M. Learning Perspective		Data Perspective			
Domain-Specific Profiling	Content Actionability	Deep Learning	Temporal Learning	Video Streams	Mobility Data	Social Interactions	Asian Languages

What is Psychographics?

An example of demographic profiling

Customer Segment: Grand Parents

Age: 70+

Gender: Male

Marital Status: Married

Financial Status: Affluent

Expected Segment Needs:
Pension, healthcare, legacy

However, not all 70 Years old are the same!

Customer Segment: Grand Parents

Age: 71

Gender: Male

Marital Status: Married

Financial Status: Affluent

Segment Needs:
Staff Retention, legal representation, walls

"Think beyond demographic, connect through psychographic

– Loreal

Marketers need to immerse themselves and their teams in the various behaviors to find the right insight that could trigger an action

Interactive Magazine
8 June 2018

What is Psychographics?

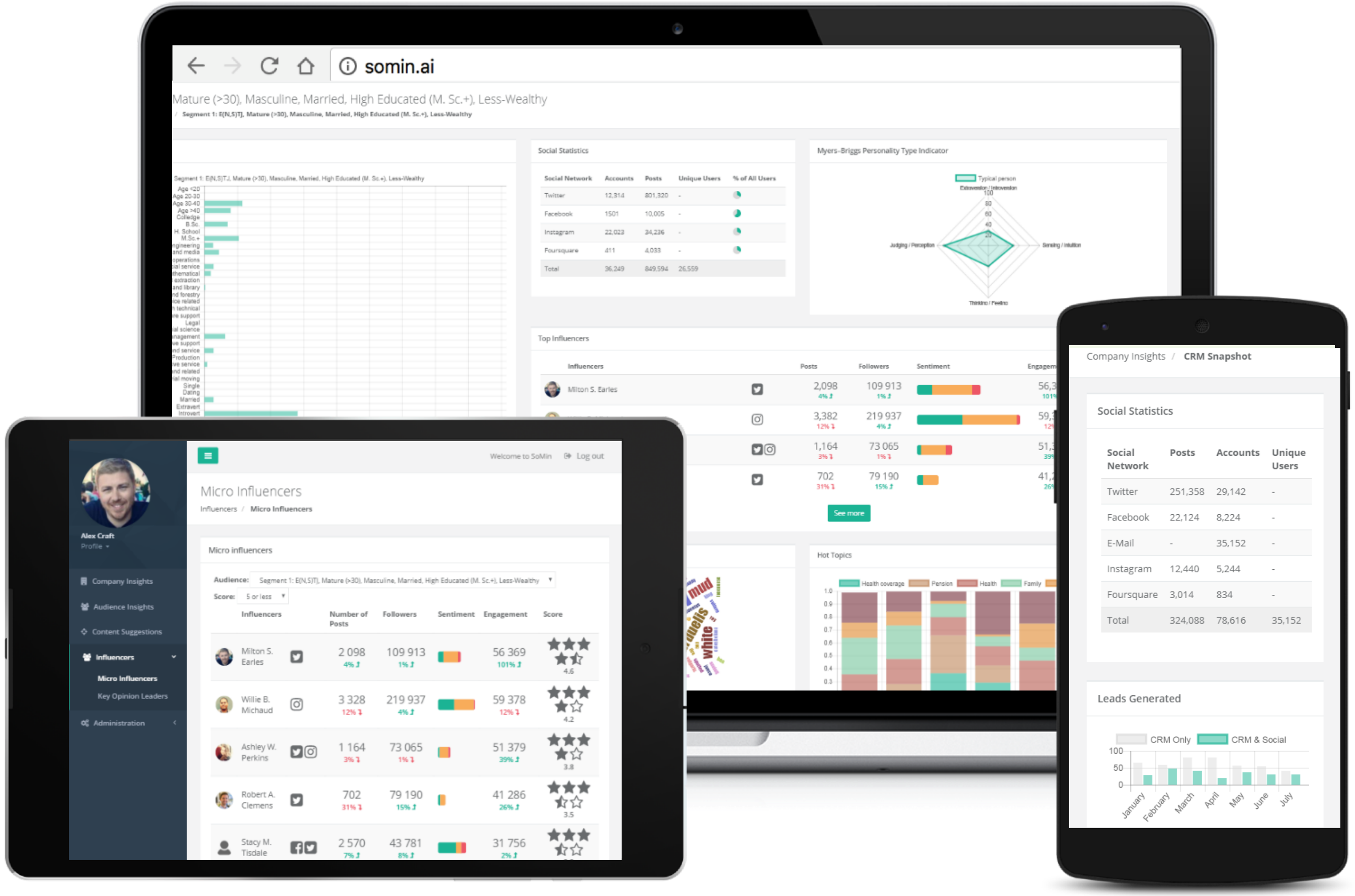
A consumer psychographic is a profile of a potential consumer based on interests, activities and personality. It is a snapshot into a consumer's lifestyle often used to quickly identify potential customers.

Companies then can use this information to create and implement highly targeted advertising and marketing campaigns

SoMin

the Secret Sauce :

Matching Customer Segments with Messages to inspire the action



Profiling API for Researchers: http://dev.somin.ai

Case Studies

Psycho-Emotional Trait Prediction for Career Planning Portal more...

Micro-Influencer Marketing Campaign for an International Restaurant Brand

more...

AI-Driven Social Media Campaigns for on of the Asia’s Leading Mega Gym

more...

Thank You

Questions?