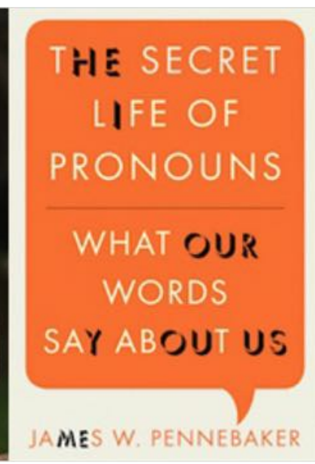
2018 Conference of
Online Social Behavior

Deep Learning and Computational Social Sciences

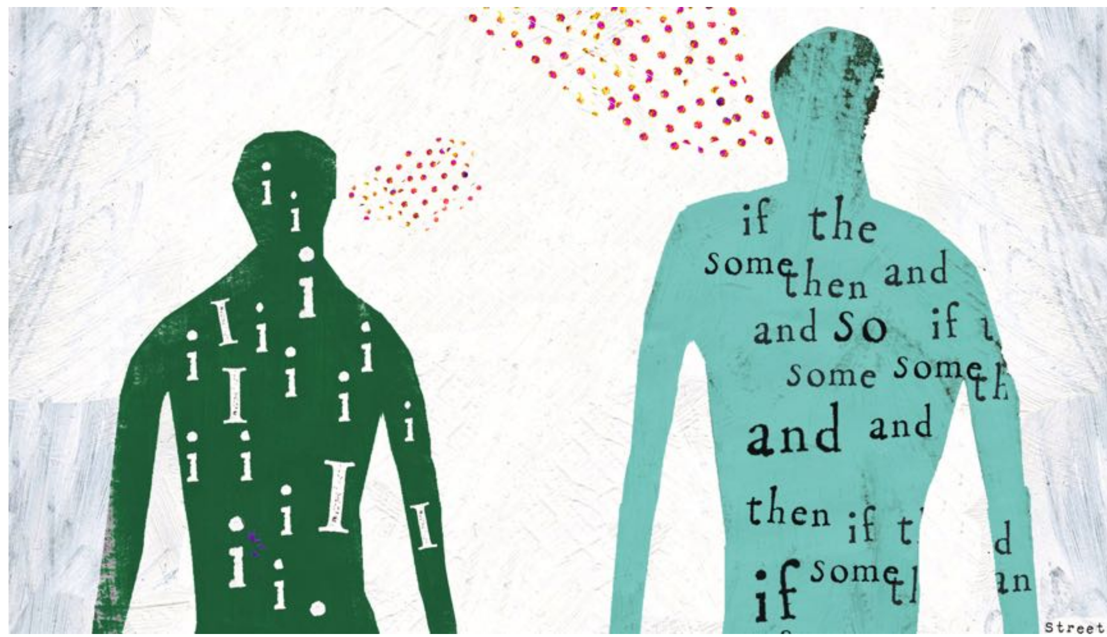
Tsinghua NLP Lab
Zhiyuan Liu

Language and Social Sciences

- Sociolinguistics and social psychology study human society by analyzing languages
- Example: Linguistic Inquiry and Word Count (LIWC)



James Pennebaker



Language and Social Sciences

- Linguistic Inquiry and Word Count (LIWC)

LIWC Results

Details of Writer: 40 year old Female
Date/Time: 6 January 2014, 1:02 am

LIWC results from input text

LIWC categories

LIWC results from personal text and formal writing for comparison

LIWC Dimension	*Your Data*	*Personal Texts*	*Formal Texts*
Self-references (I, me, my)	8.33	11.4	4.2
Social words	4.17	9.5	8.0
Positive emotions	2.08	2.7	2.6
Negative emotions	1.04	2.6	1.6
Overall cognitive words	3.12	7.8	5.4
Articles (a, an, the)	2.08	5.0	7.2
Big words (> 6 letters)	20.83	13.1	19.6

The text you submitted was 96 words in length.

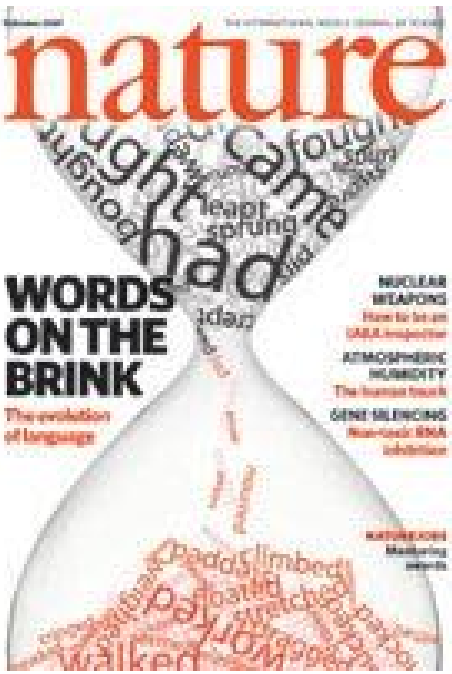
Your writing:

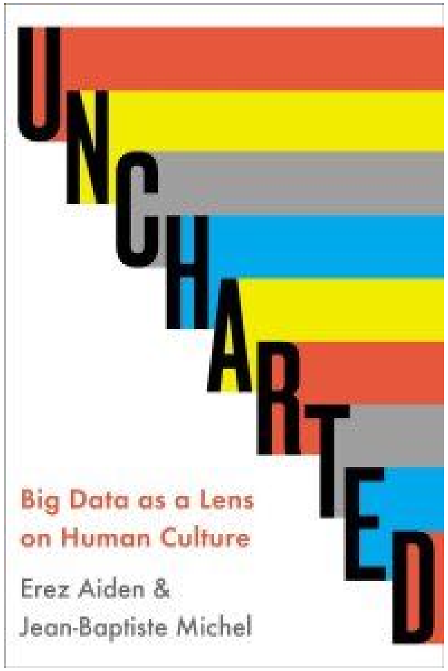
I'm newly diagnosed with type 2 diabetes. I also struggle with both calcium and uric acid kidney stones as well as the rare blood disorder LEIDEN FACTOR V. Is there anyone in this community who deals with Leiden as well as diabetes? If there is I would LOVE to be able to chat with you regarding diet and possible weight loss plans. I currently have no regular doctor and no insurance so my diabetes is uncontrolled at this time. I am working hard to educate myself AND make the necessary changes to improve my current health.

Input text: A post from a 40 year old female member in American Diabetes Association online community

Computational Social Sciences

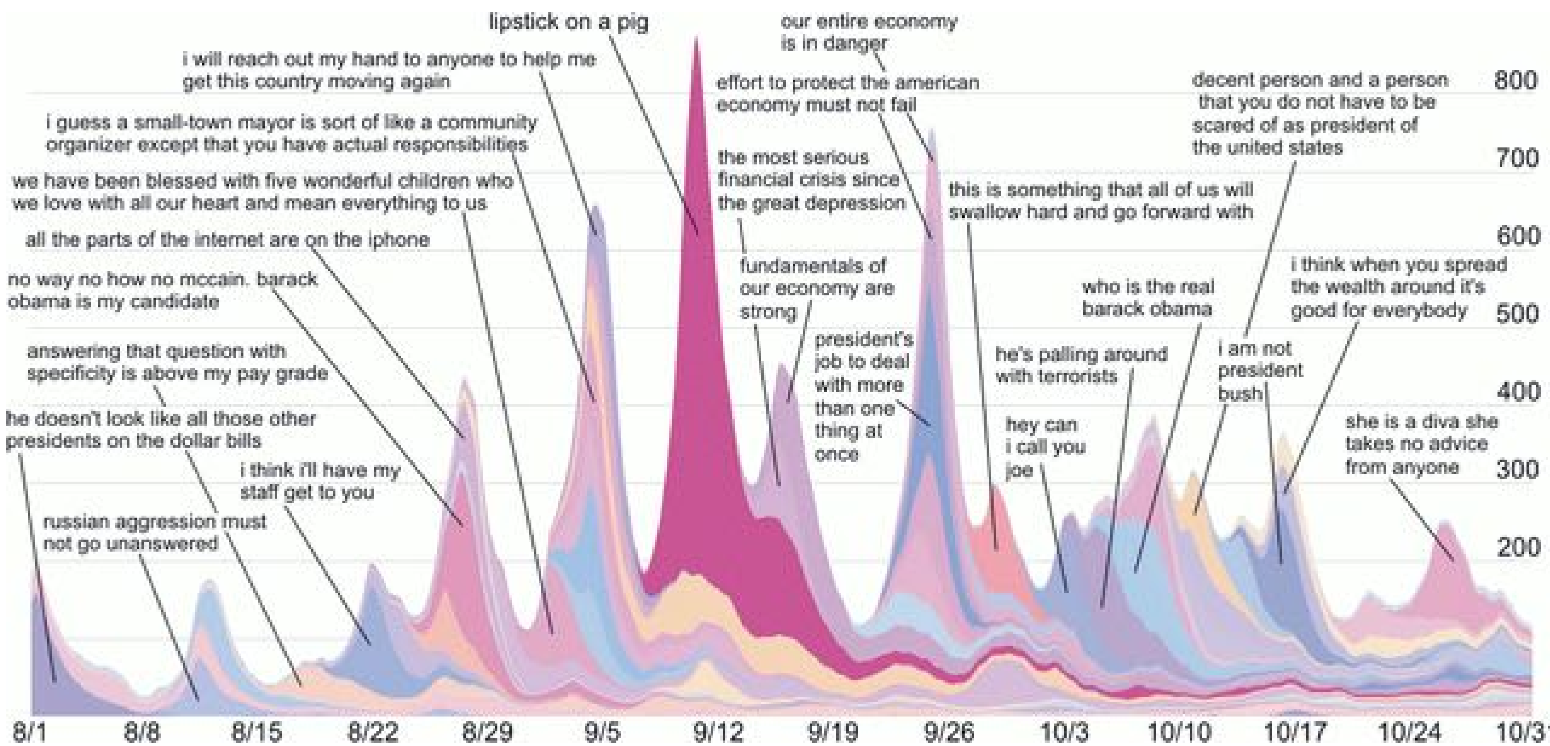
- Harvard Team collected 5 million Google Books (1800-2000) , and counted keyword frequencies to study human culture
- Culturomics: http://www.culturomics.org/
- Google Book N-grams: https://books.google.com/ngrams





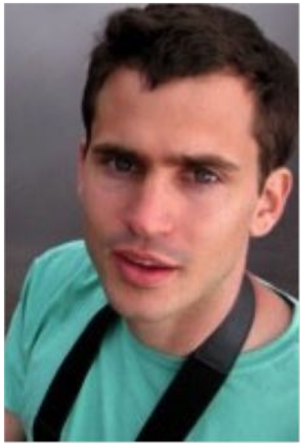
Keyword-based CSS

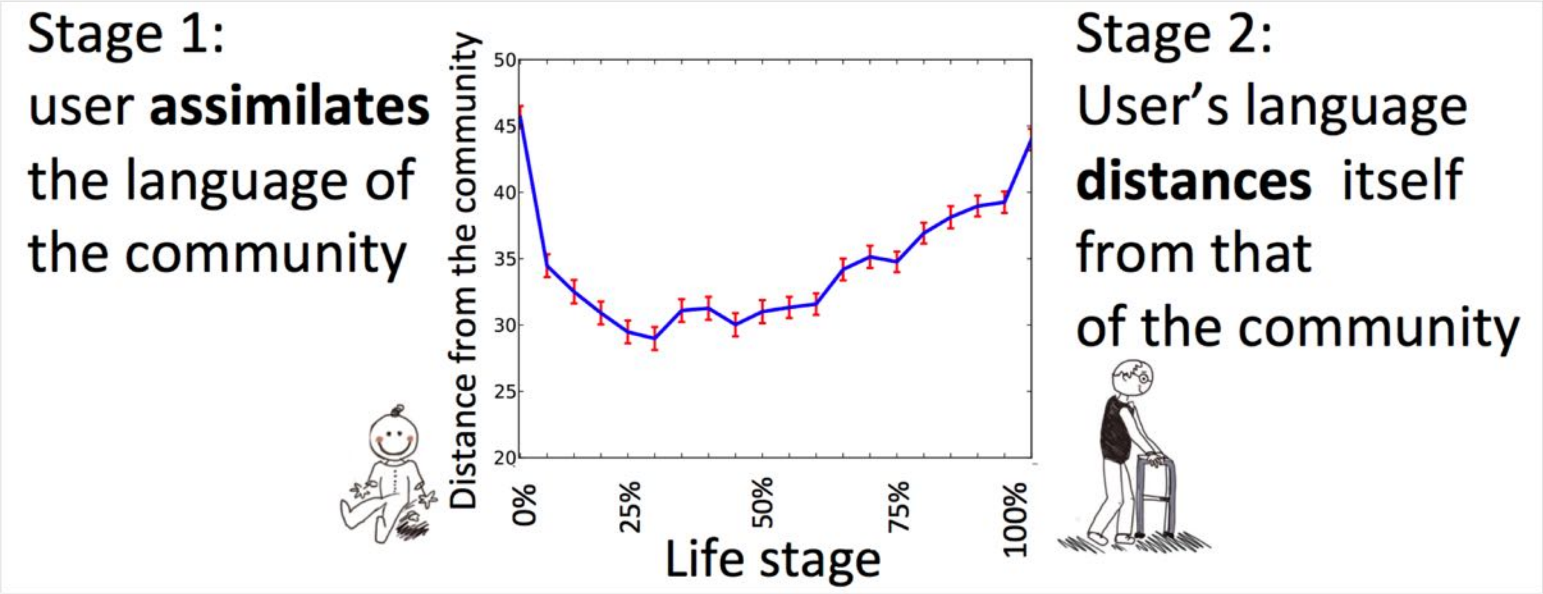
- Jure Leskovec at Stanford collected 90 million blogs and counted quotes as memes: http://www.memetracker.org/
- Example: “you can put lipstick on a pig” from Obama



Keyword-based CSS

- Cristian Danescu-Niculescu-Mizil at Cornell studied language style changes over times of online community users
- WWW 2013 Best Paper: No country for old members: User lifecycle and linguistic change in online communities





Keyword-based CSS

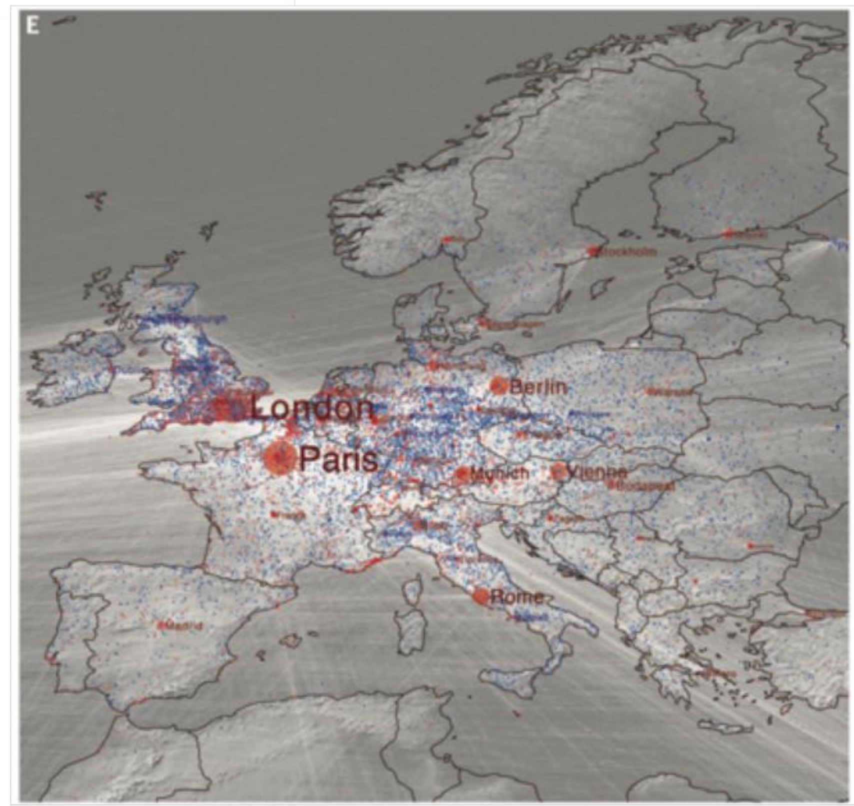
QUANTITATIVE SOCIAL SCIENCE

A network framework of cultural history

Maximilian Schich,[1,2,3*] Chaoming Song,[4] Yong-Yeol Ahn,[5] Alexander Mirsky,[2] Mauro Martino,[3] Albert-László Barabási,[3,6,7] Dirk Helbing[2]

Science 2014
Culture Center:
Birth Place→ Death Place

Winckelmann Corpus



Freebase

Keyword Extraction from Social Media

Keyword-based Occupation Prediction

- Use keywords in UGC as features for occupation prediction, with accuracy 83.8%
- User profiling in social computation

No.	Occupation	Precision	Recall	F
1	media	84.04%	90.60%	87.20%
2	government	94.03%	93.78%	93.90%
3	entertainment	84.78%	82.25%	83.49%
4	estate	88.22%	86.92%	87.57%
5	**finance**	68.86%	73.05%	70.90%
6	**IT**	72.93%	68.38%	70.58%
7	sports	94.05%	92.84%	93.44%
8	**education**	76.88%	73.80%	75.31%
9	fashion	84.84%	78.94%	81.78%
10	games	85.47%	84.19%	84.82%
11	literature	84.68%	75.99%	80.10%
12	**services**	65.32%	57.45%	61.13%
13	**art**	76.84%	69.92%	73.22%
14	healthcare	87.10%	87.50%	87.30%

No.	Occupation	Conj.	Interj.	M.P.
1	media	1.19%▽	0.22%△	2.16%△
2	government	1.29%	0.17%	1.70%
3	entertainment	1.08%▽	0.26%△	2.38%△
4	estate	1.26%	0.15%	1.72%
5	finance	1.39%△	0.15%▽	1.65%▽
6	IT	1.35%△	0.15%▽	1.66%
7	sports	1.04%▽	0.25%△	2.60%△
8	education	1.42%△	0.16%▽	1.55%▽
9	fashion	1.25%	0.22%	1.95%
10	games	1.34%	0.16%	1.26%▽
11	literature	1.31%	0.27%△	2.25%
12	services	1.29%	0.18%	1.94%
13	art	1.11%▽	0.22%△	2.06%△
14	healthcare	1.76%△	0.11%▽	1.15%▽

Event Detection in Social Media

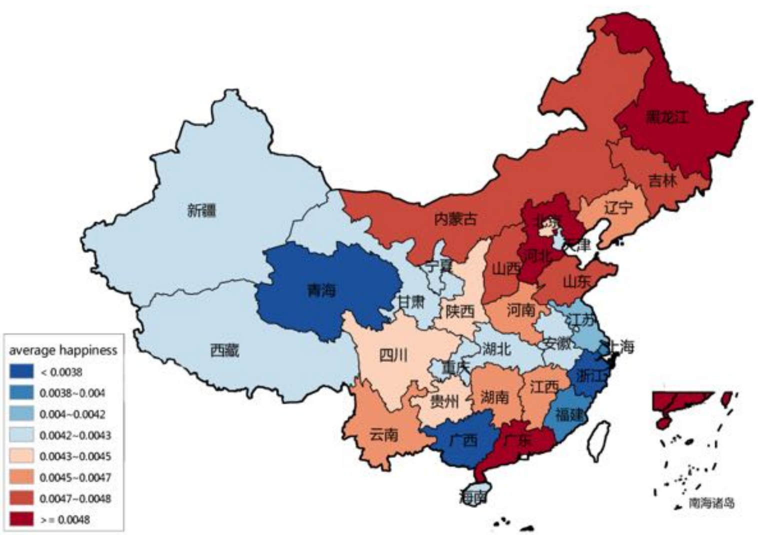
- Use keywords in UGC to detect big events of users with accuracy 75%
- Such as health (illness), marriage, life (buying house), and career

Class	Precision	Recall
Health	0.883	**0.782**
Love	**0.926**	0.543
Career	0.825	0.687
Life	0.807	0.758
Others	0.676	0.767

Happiness in Social Media

- Quantitively measure the happiness of Chinese based on the PERMA theory

Date	$h_{ave} \times 10^{-3}$	Remark	Date	$h_{ave} \times 10^{-3}$	Remark
11-24	6.849	Thanksgiving Day	07-25	0.989	7.23 highway accident
11-11	6.804	Single's Day	07-24	1.772	7.23 highway accident
05-08	6.687	Mother's Day	07-26	2.148	7.23 highway accident
01-01	6.552	New Year's Day	07-27	2.317	7.23 highway accident
09-12	6.513	Mid-autumn festival	03-11	2.504	Japan's 3.11 earthquake



Positive Factor	r
Commodity Retail Sales	0.773
Postal Packages	0.745
Total Retail Sales of Consumer Goods	0.727
Negative Factor	**r**
non-manufacturing PMI	-0.527
Railways Passenger-kilometers	-0.509
Inventory Index	-0.500

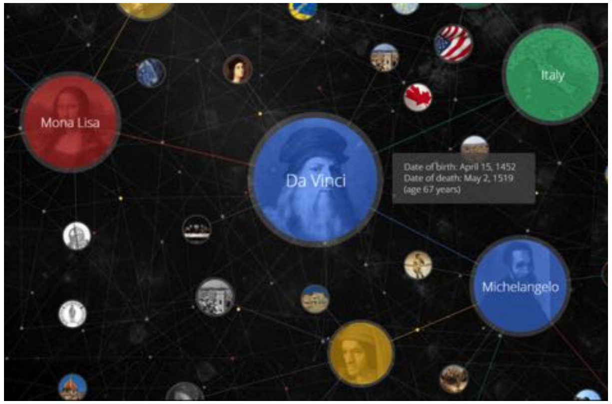
Symbol-based Representation

star [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...]

sun [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...]

sim(star, sun) = 0 ☹

Challenges in CSS



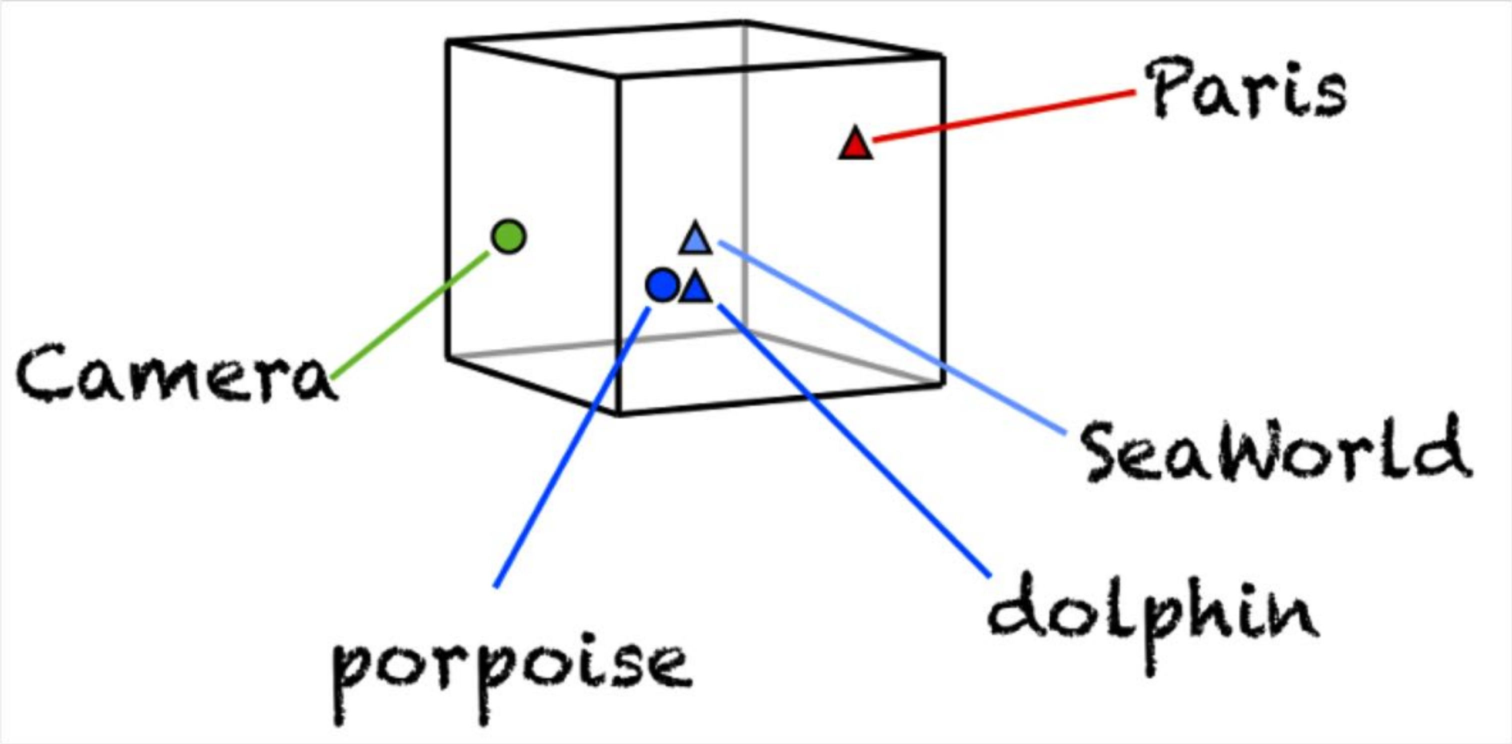
Social Networks UGC Knowledge

Key Challenge

How to compute semantic relations among heterogeneous information?

New Trend in Deep Learning

- Distributed Representation, i.e., embedding
- Each object is represented as a dense, real-valued and low-dimensional vector



Distributed Representation

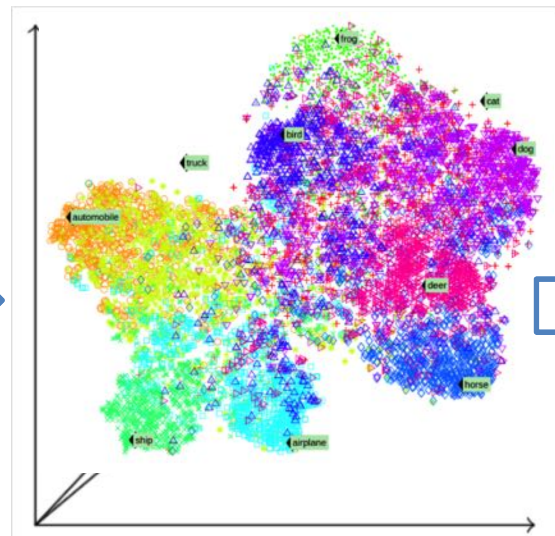
- Build a unified semantic space for heterogeneous information in social sciences

Knowledge

Items

Text

User



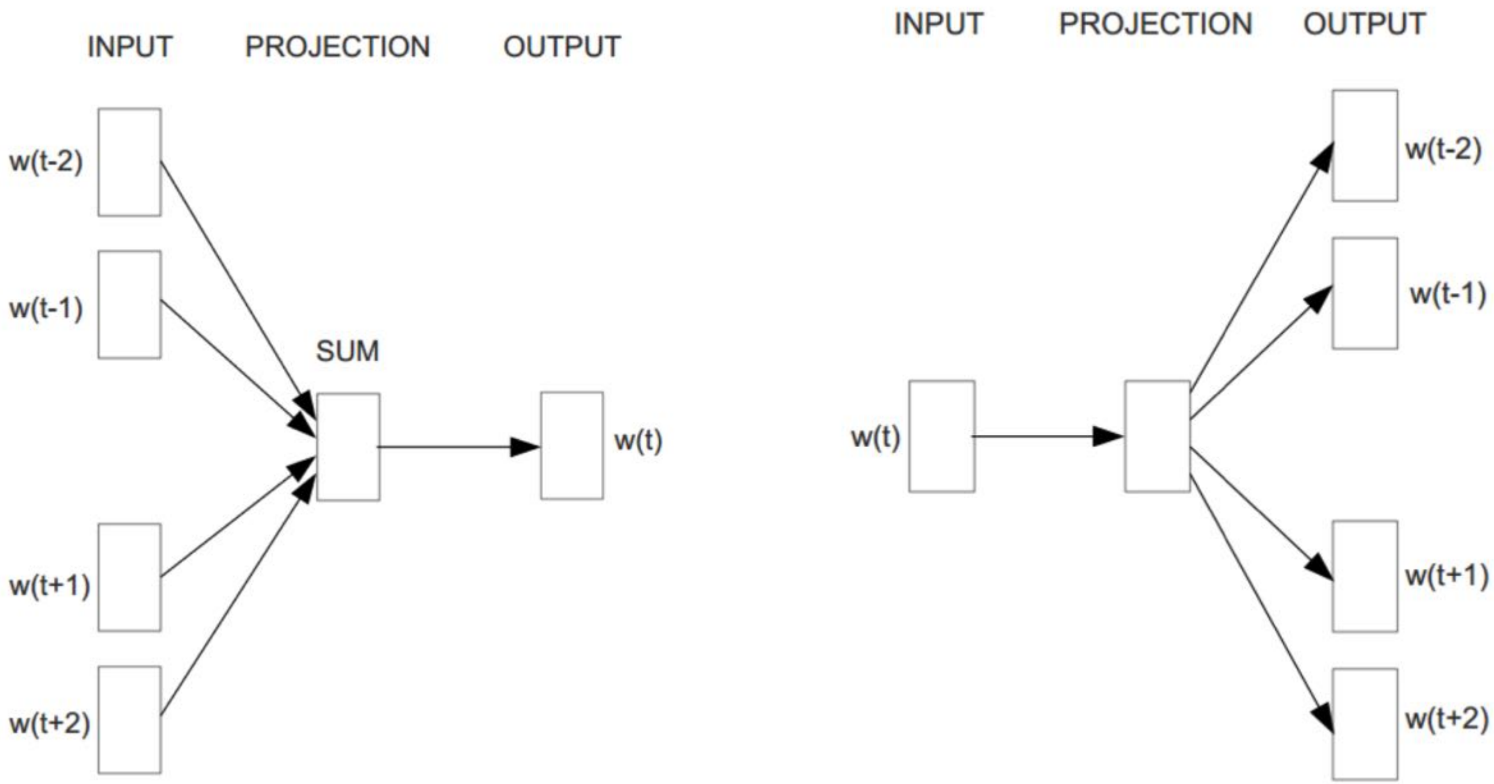
Unified Semantic Space

Recommendation

User Profiling

SNS

Word Embedding



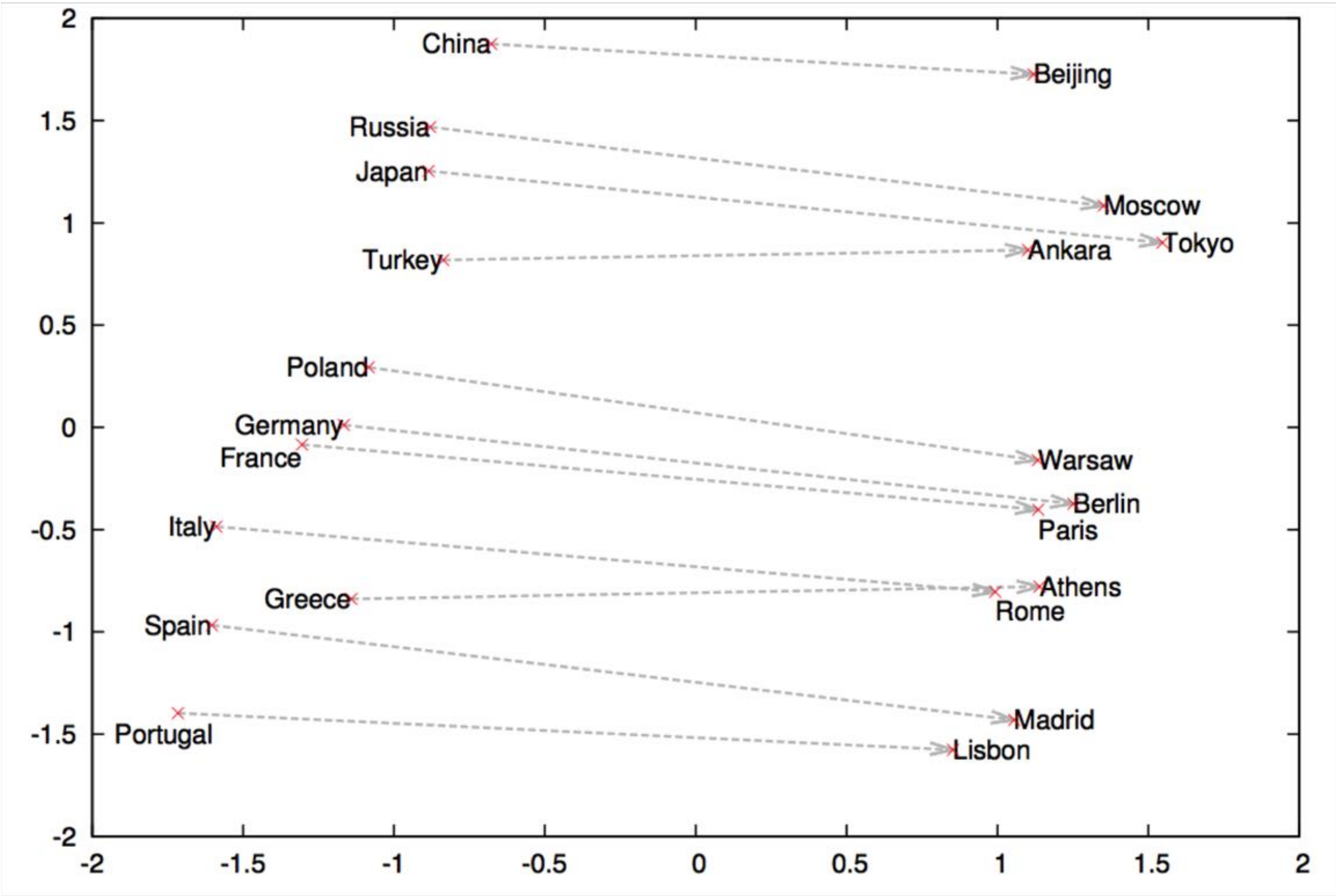
word2vec

Tomas Mikolov et al. Distributed representations of words and phrases and their compositionality. NIPS 2013.

Word Embedding for Computing Similarity

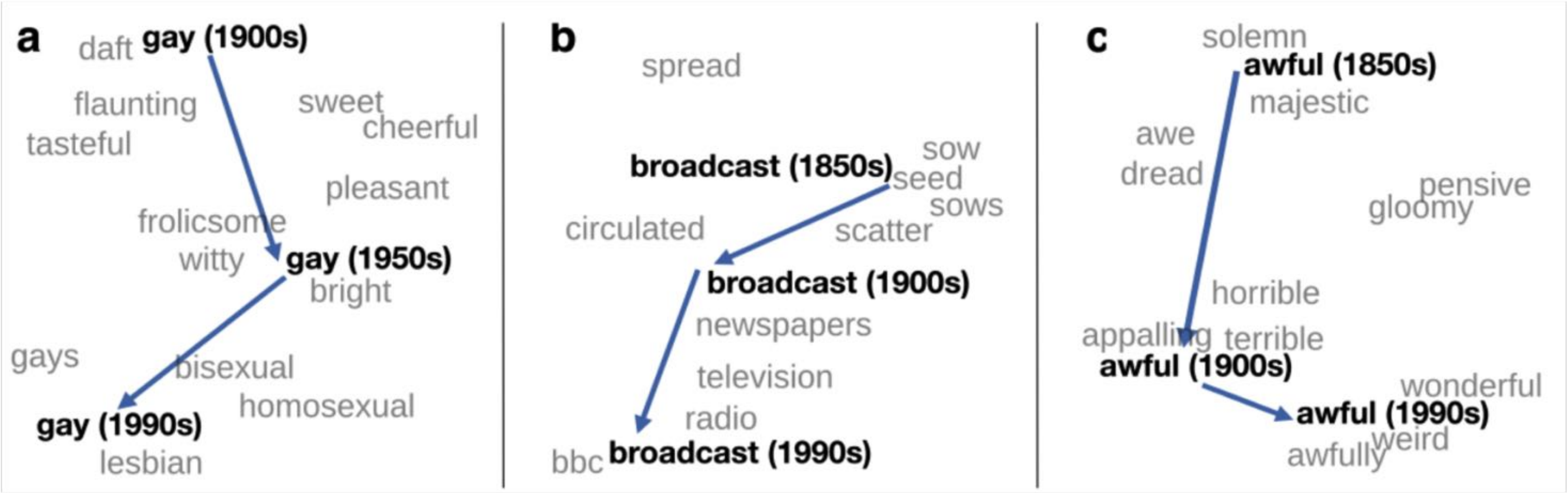
```
(EXIT to break): china

n vocabulary: 486

                              Word       Cosine distance
-----------------------------------------------------------
                            taiwan              0.768188
                             japan              0.652825
                             macau              0.614888
                             korea              0.614887
                               prc              0.613579
                           beijing              0.605946
                            taipei              0.592367
                          thailand              0.577905
                          cambodia              0.575681
                         singapore              0.569950
                          republic              0.567597
                          mongolia              0.554642
                           chinese              0.551576
```

Word Embedding for Implicit Relations



W(‘‘China“) − W(‘‘Beijing”) ≃ W(‘‘Japan“) − W(‘‘Tokyo")

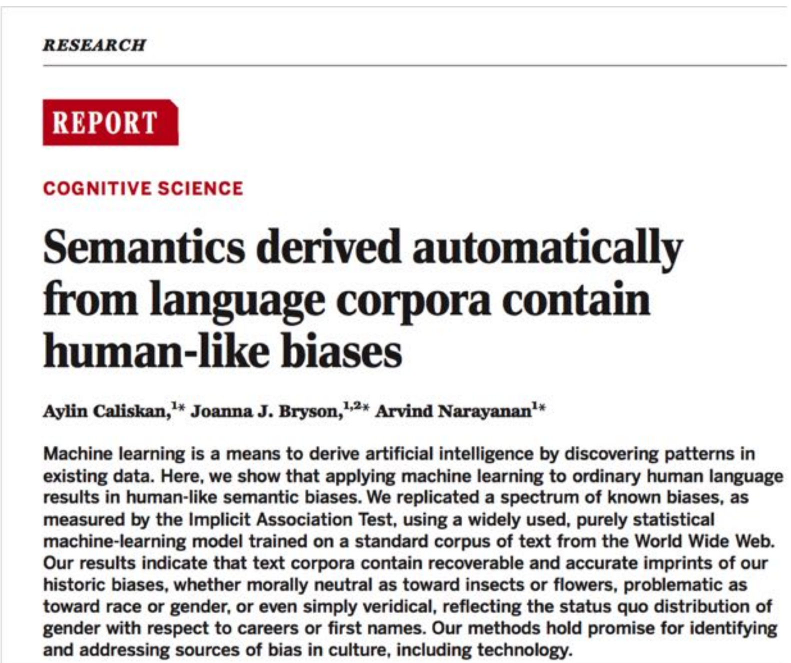
Word Embedding for Semantic Changes



William L. Hamilton, Jure Leskovec, and Dan Jurafsky. ACL 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.

Word Embedding for Political Biases

- Science Paper (2017) finds word embeddings learned from text corpora contain political biases



RESEARCH

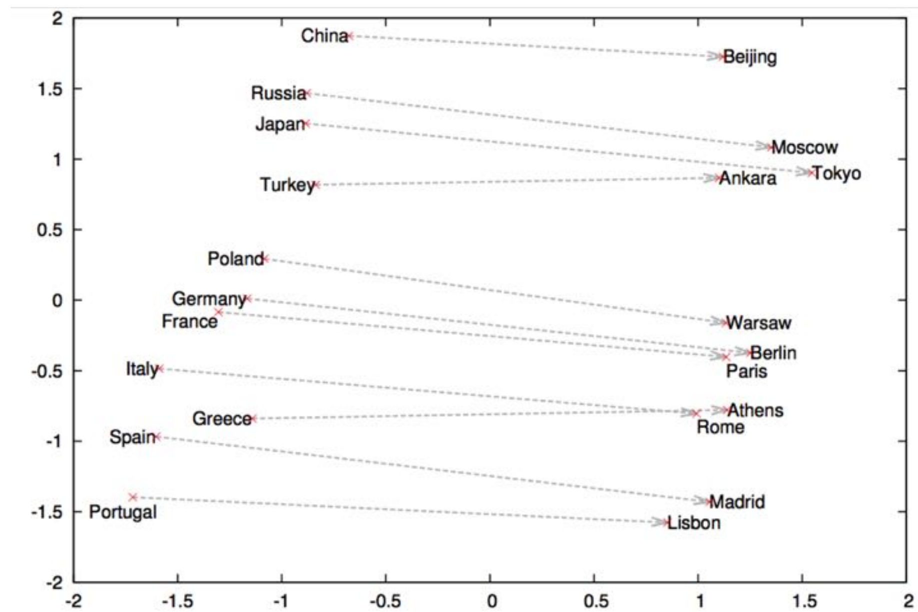
REPORT

COGNITIVE SCIENCE

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan,[1*] Joanna J. Bryson,[1,2*] Arvind Narayanan[1*]

Machine learning is a means to derive artificial intelligence by discovering patterns in existing data. Here, we show that applying machine learning to ordinary human language results in human-like semantic biases. We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology.



Word Embedding for Political Biases

- Science Paper (2017) finds word embeddings learned from text corpora contain political biases
- Consistent to the Implicit Association Test in Psychology

Target words	Attribute words	Original finding				Our finding			
		Ref.	N	d	P	N_T	N_A	d	P
Flowers vs. insects	Pleasant vs. unpleasant	(*5*)	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	(*5*)	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	(*5*)	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from (*5*)	(*7*)	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from (*9*)	(*7*)	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	(*9*)	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	(*9*)	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	(*10*)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	(*23*)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	(*9*)	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

Deep Learning for Depression Detection

- Apply neural network models to detect depressions based on UGC
- EMNLP 2017 Best Paper

Depression and Self-Harm Risk Assessment in Online Forums

Andrew Yates[†*] Arman Cohan[‡*] Nazli Goharian[‡]

[†]Max Planck Institute for Informatics,
Saarland Informatics Campus Saarbruecken, Germany

[‡]Information Retrieval Lab, Department of Computer Science,
Georgetown University, Washington DC, USA

ayates@mpi-inf.mpg.de

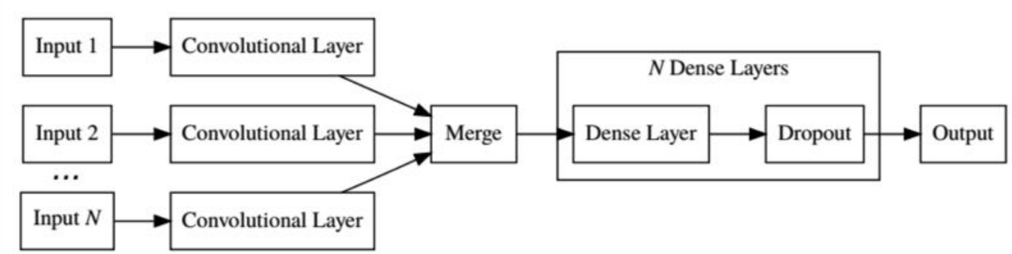
{arman,nazli}@ir.cs.georgetown.edu

Abstract

Users suffering from mental health conditions often turn to online resources for support, including specialized online support communities or general communities such as Twitter and Reddit. In this work, we present a framework for supporting and

well-being of families and on societies in general. Therefore identifying individuals at risk of self-harm and providing support to prevent it remains an important problem (Ferrari et al., 2014).

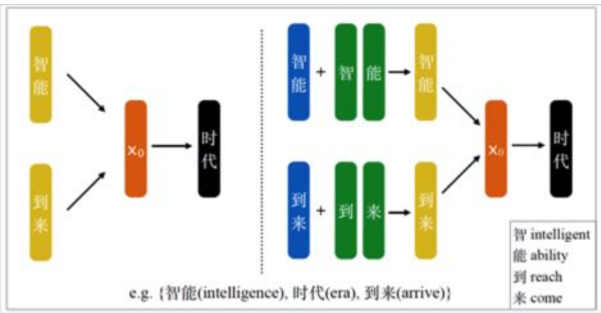
Social media is often used by people with mental health problems to express their mental issues and seek support. This makes social media a significant resource for studying language re-



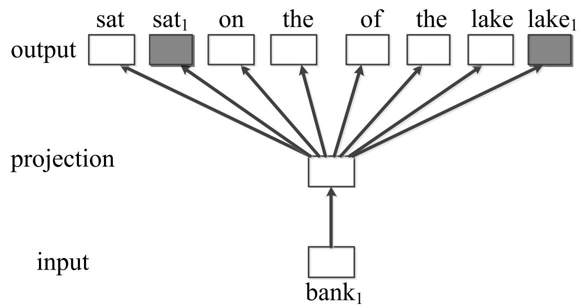
Method	Precision	Recall	F1
BoW - MNB	0.44	0.31	0.36
BoW - SVM	**0.72**	0.29	0.42
Feature-rich - MNB	0.69	0.32	0.44
Feature-rich - SVM	0.71	0.31	0.44
User model - CNN	0.59	**0.45**	**0.51**

Language Representation Learning

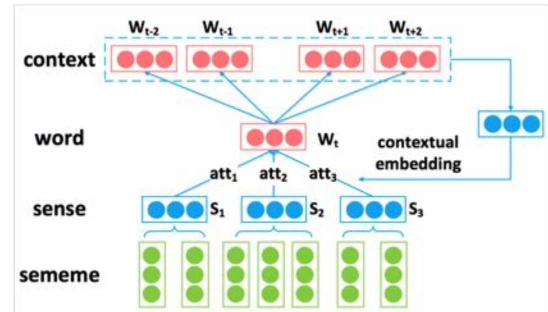
- Learn semantic representations of multi-grained language units



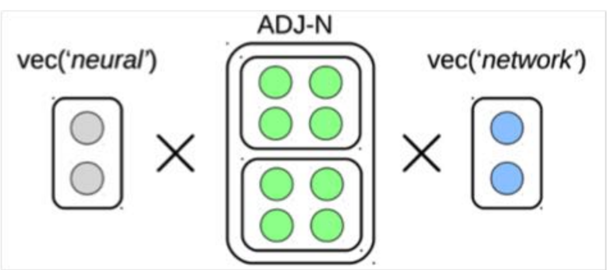
Character and Word Embedding (IJCAI 2015)



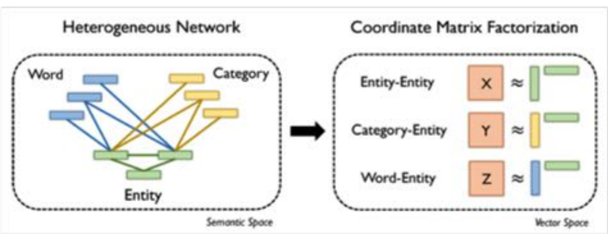
English Sense Embedding (EMNLP 2014)



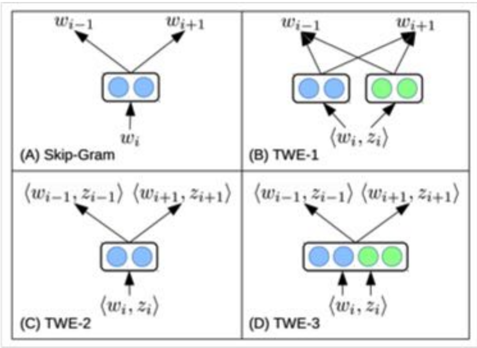
Chinese Sense Embedding (ACL 2017)



Phrase Embedding (AAAI 2015)



Entity Embedding (IJCAI 2015)



Document Embedding (IJCAI 2015)

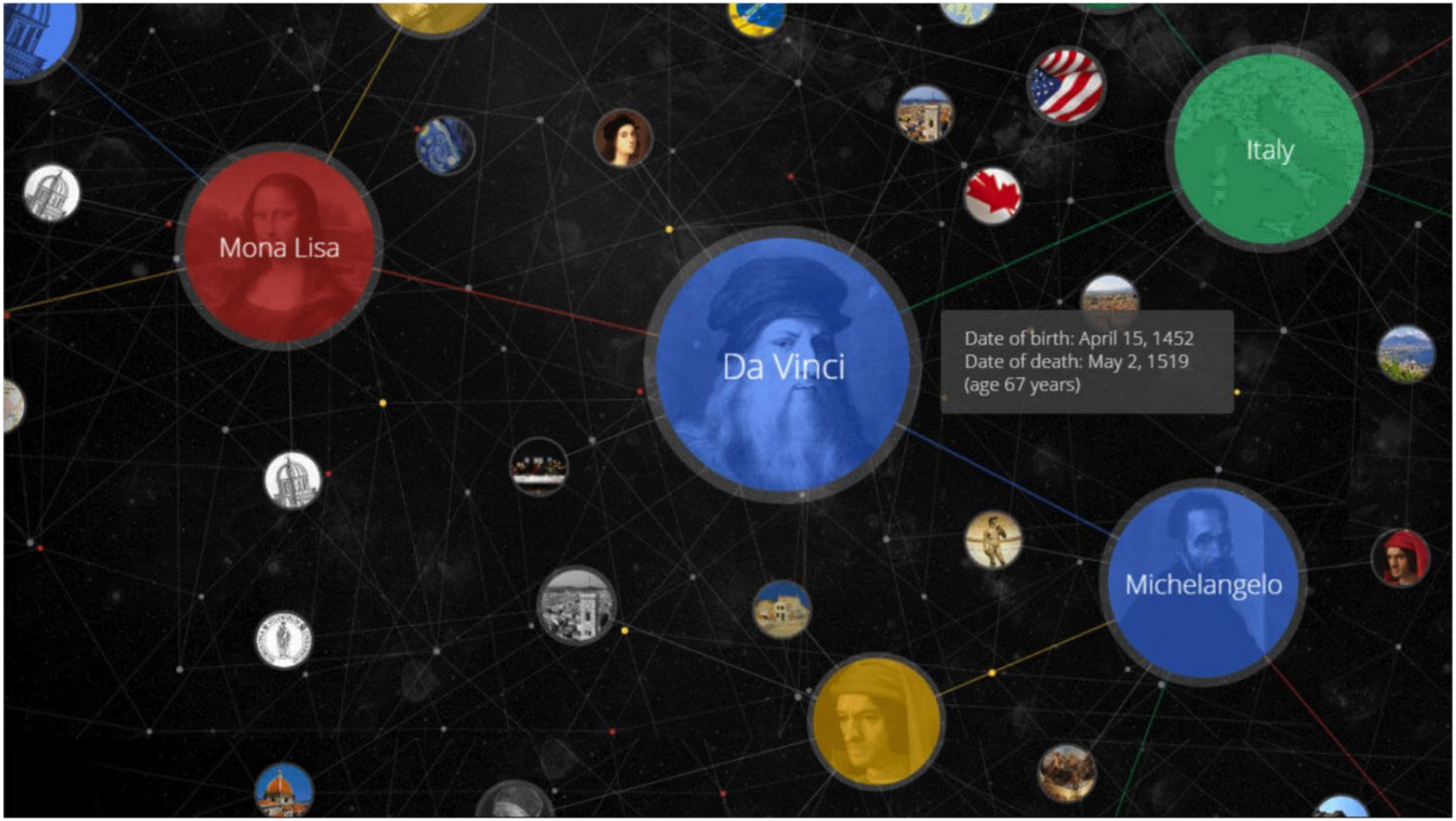
From Language to Knowledge

author

Shakespeare

Romeo and Juliet

Knowledge Graph

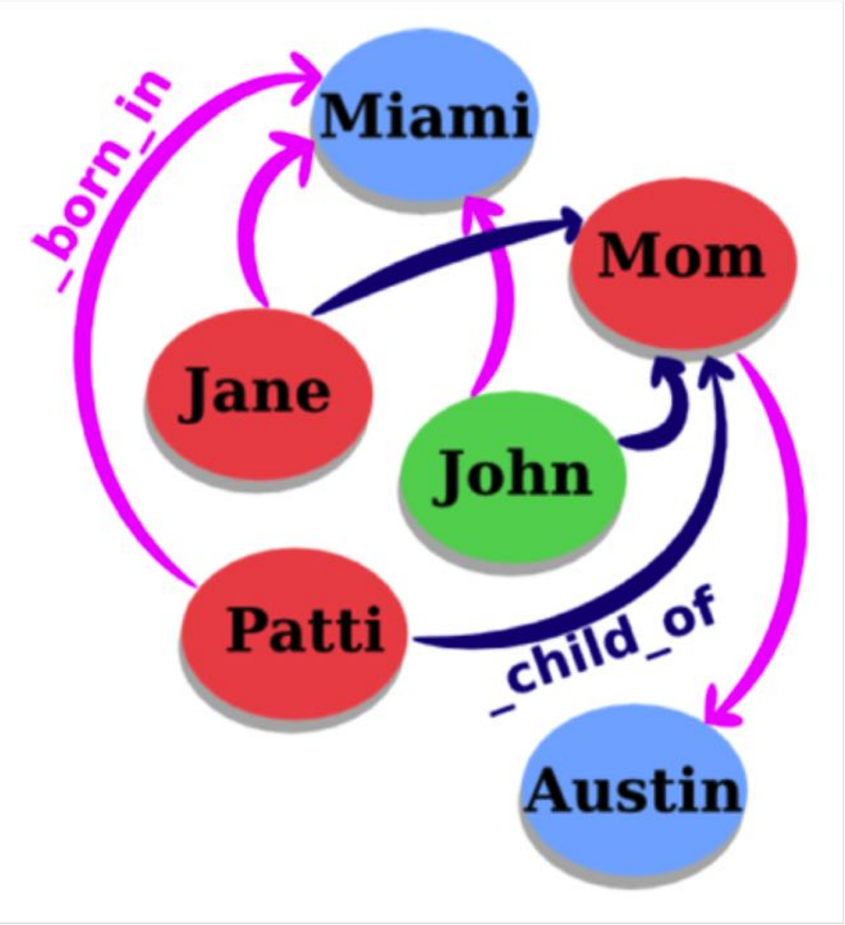


Knowledge Graph

- Entity as vertices and relations as edges

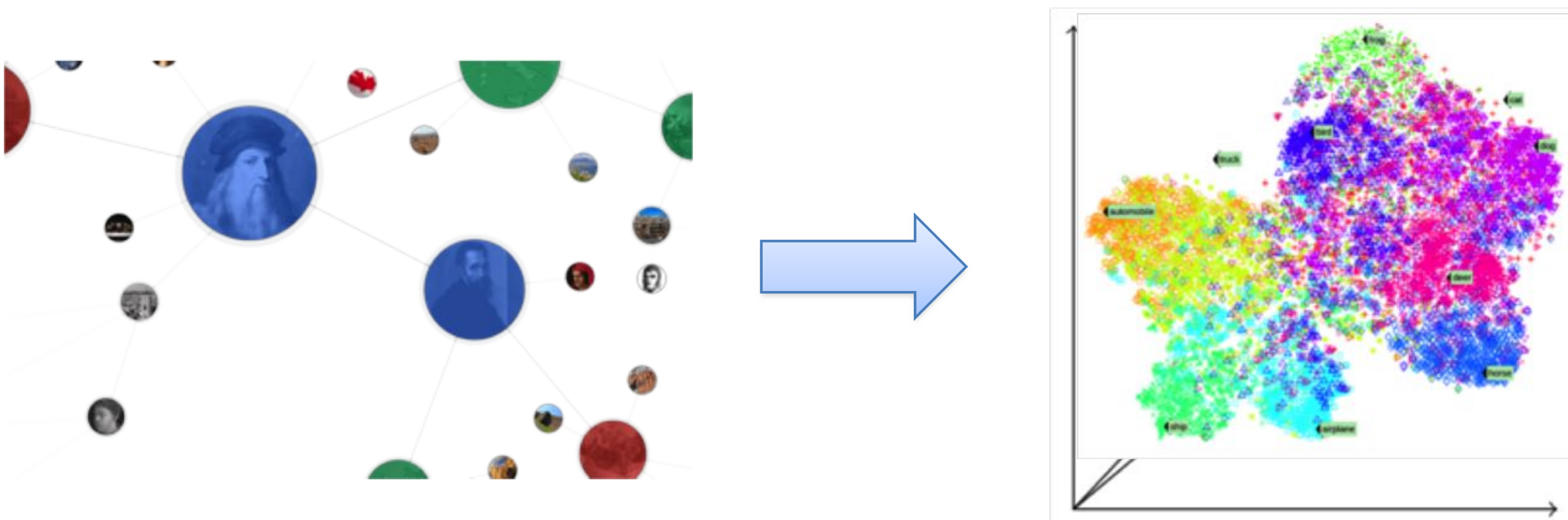
- Facts as triples
  - (head, *relation*, tail)

- Typical KG
  - Lexical KG: WordNet
  - World KG: Freebase



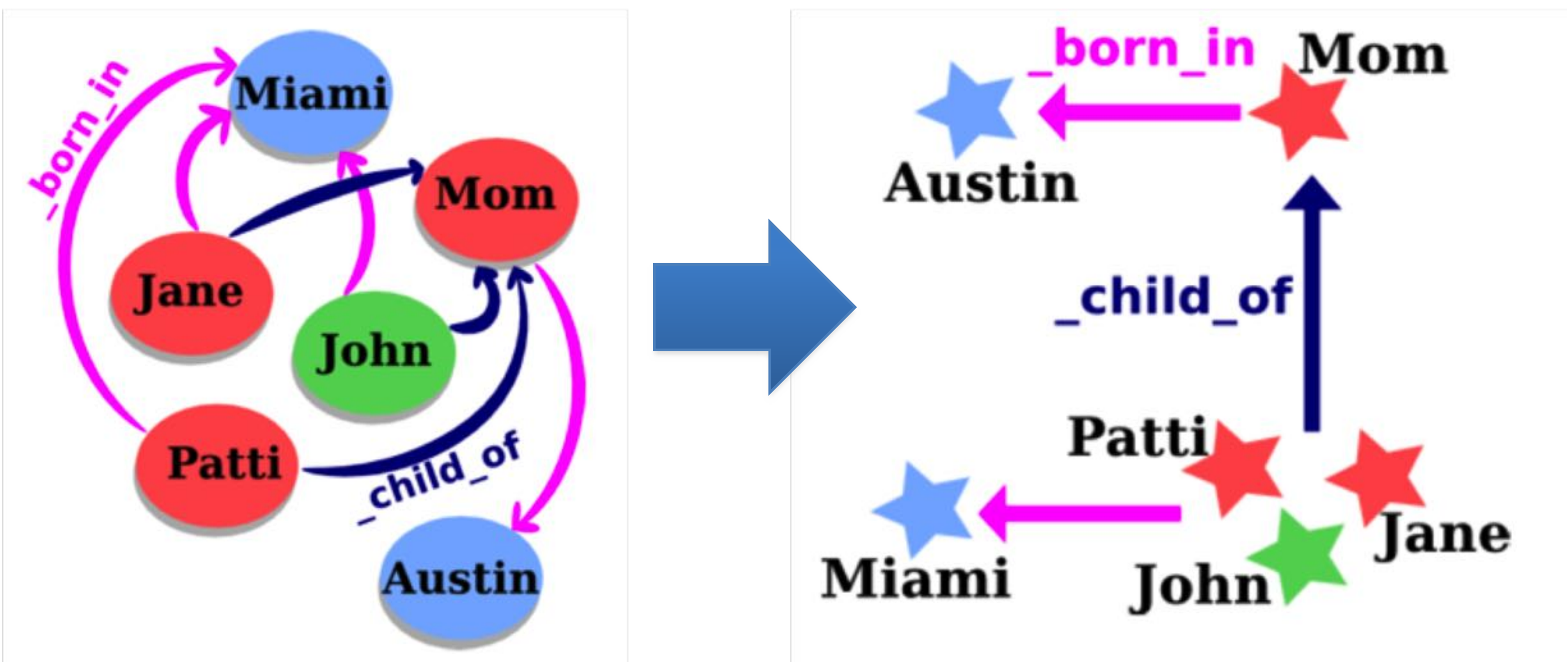
Knowledge Representation

- Symbol-based knowledge representation can not well compute semantic relations of entities
- Solution: project knowledge into low-dimensional space



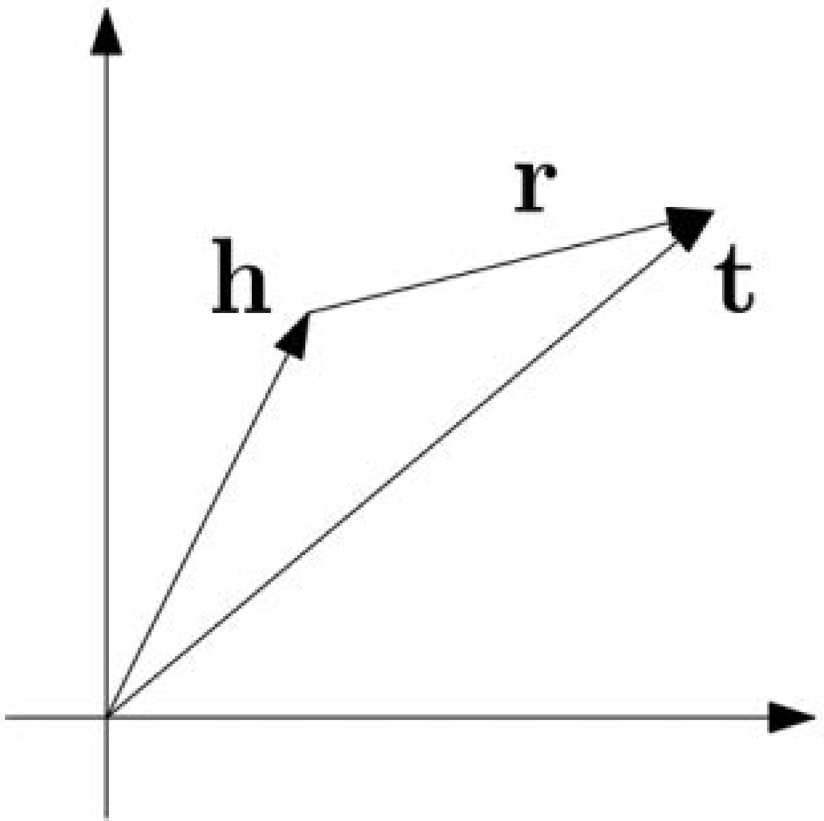
TransE

- For each fact (head, relation, tail), regard the relation as a translation from the head entity to the tail entity



TransE

- For each fact (head, relation, tail), regard the relation as a translation from the head entity to the tail entity



Learning objective

$$\mathbf{h} + \mathbf{r} = \mathbf{t}$$

Entity Prediction

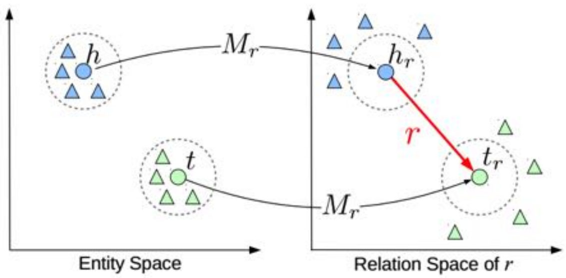
WALL-E _has_genre

- Animation
- Computer animation
- Comedy film
- Adventure film
- Science Fiction
- Fantasy
- Stop motion
- Satire
- Drama
- Connecting

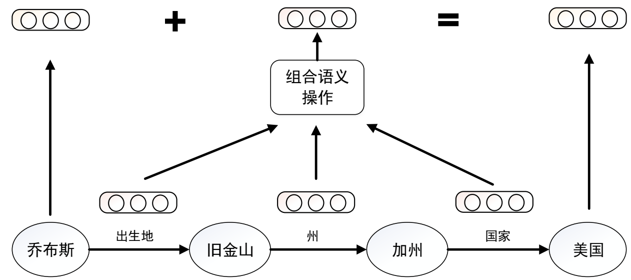
h + r = ?

Knowledge Representation Learning

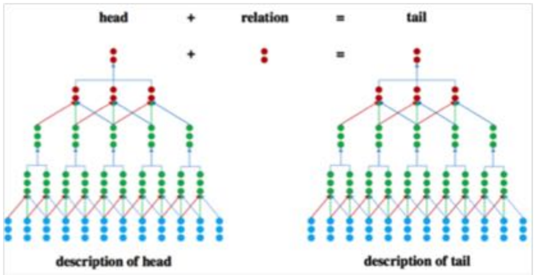
- Incorporate rich information in KG (such as description, class and images) for KRL



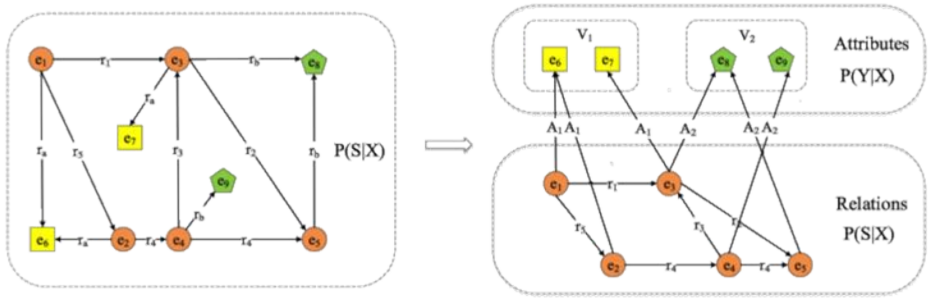
KRL with Complex Relations
TransR (AAAI 2015)



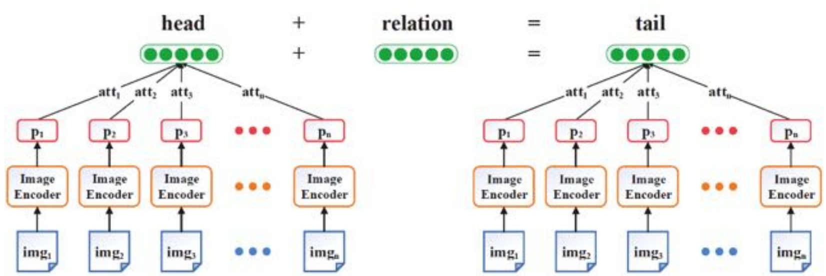
KRL with Relation Paths
PTransE (EMNLP 2015)



KRL with Entity Descriptions
DKRL (AAAI 2016)

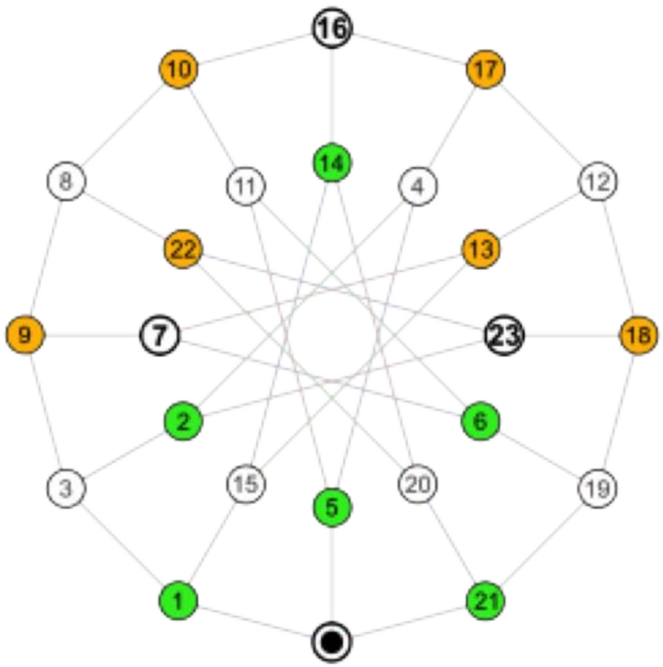


KRL with Entities, Relations and Attributes
KR-EAR (IJCAI 2016)

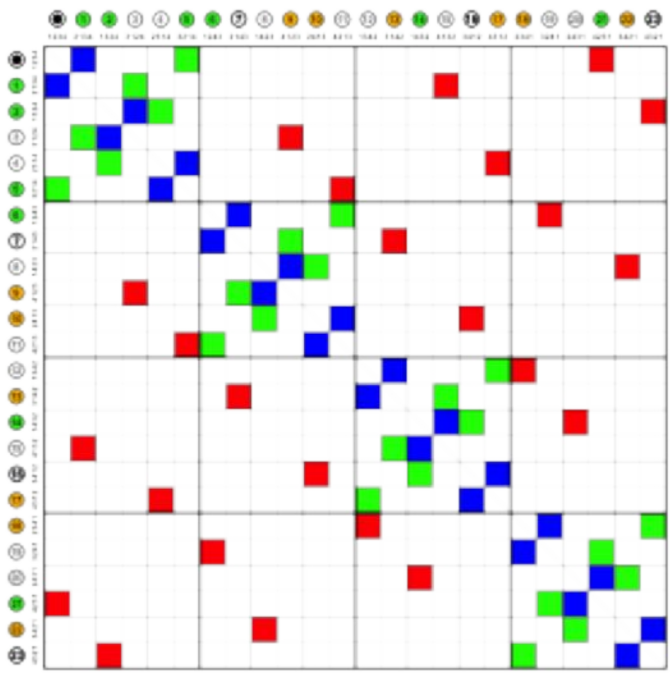


KRL with Entity Images
IKRL (IJCAI 2017)

Network Representation



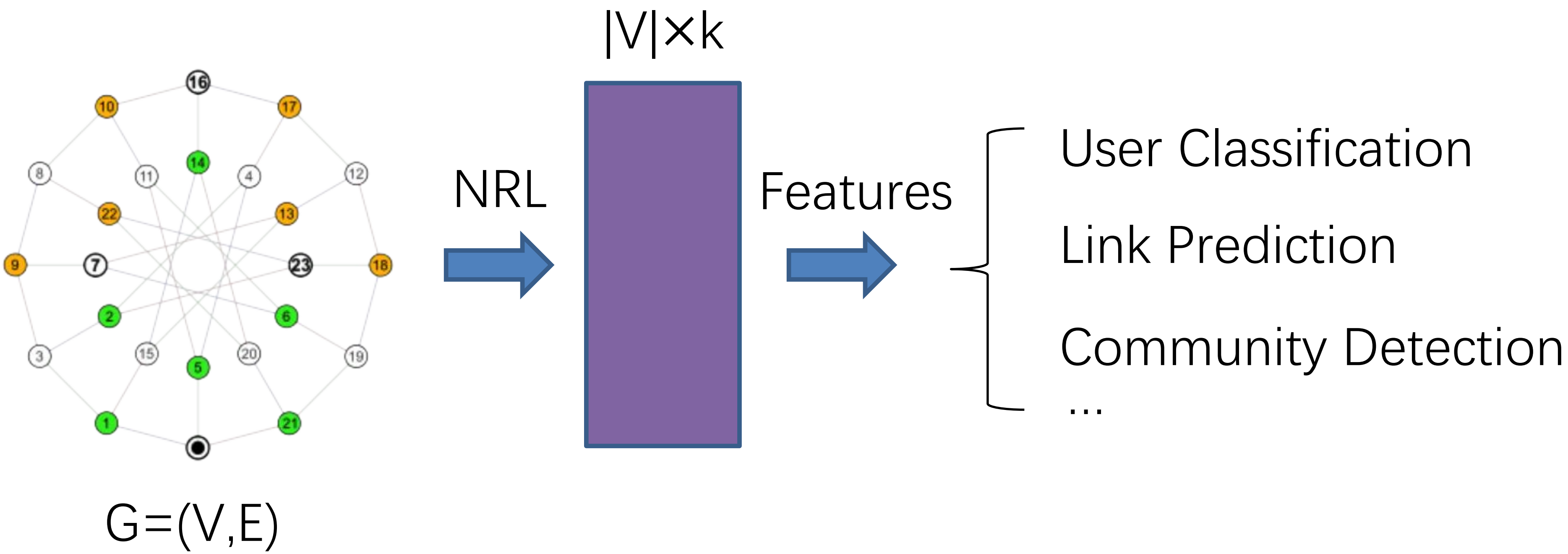
Social Networks



Adjacent Matrix

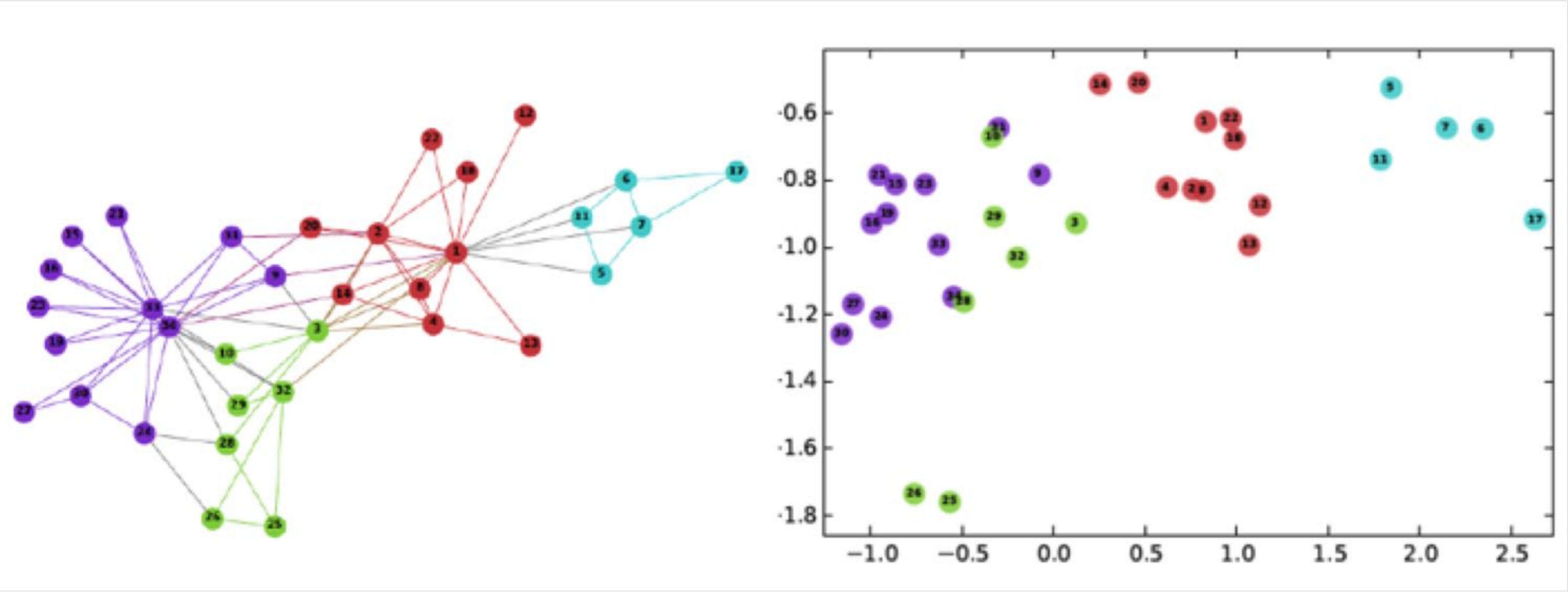
Network Representation Learning

- Project network vertices into low-dimensional space



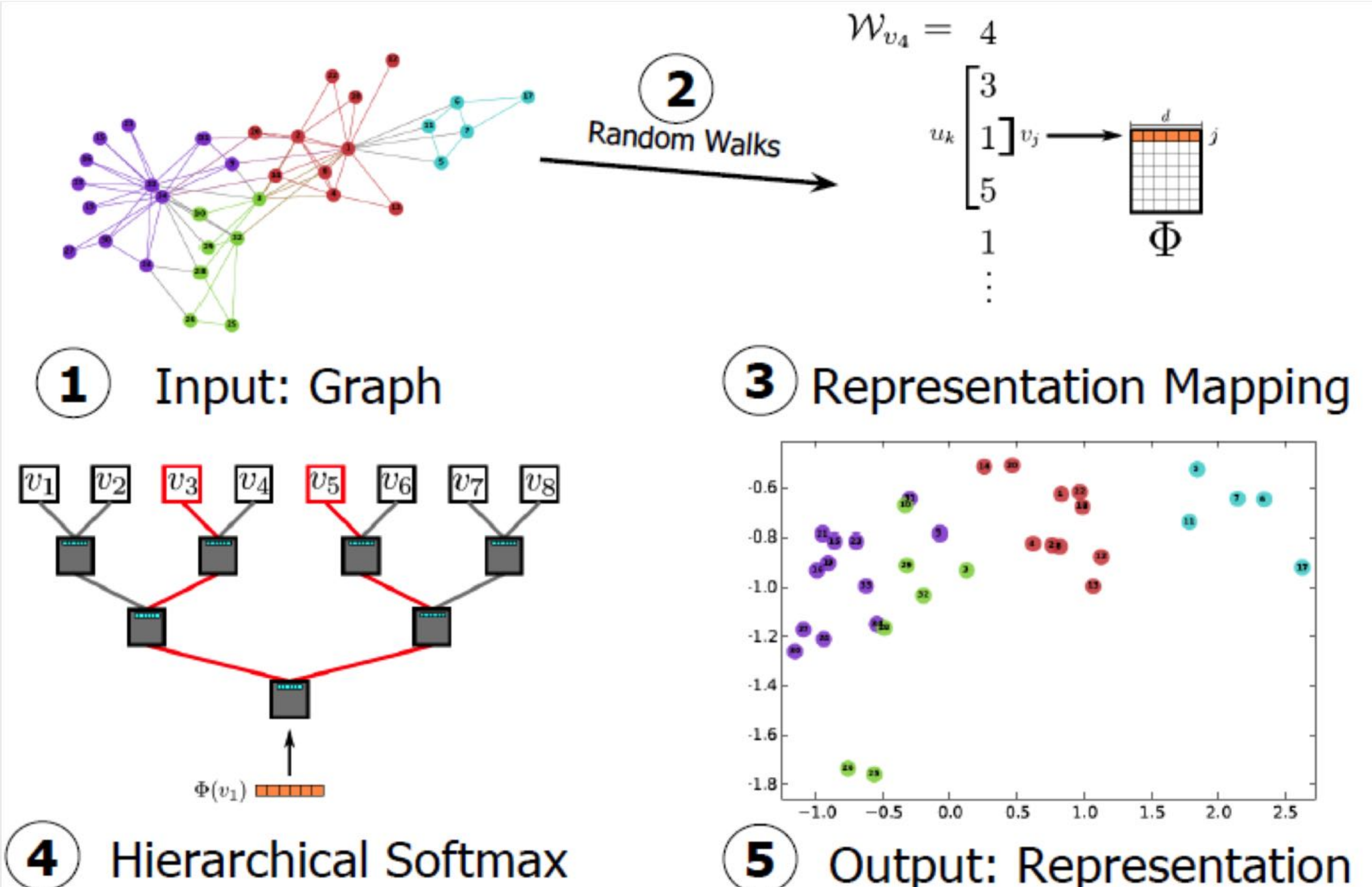
Network Representation Learning

- Karate Graph (k=2)



Perozzi et al. DeepWalk: Online Learning of Social Representations. KDD 2014

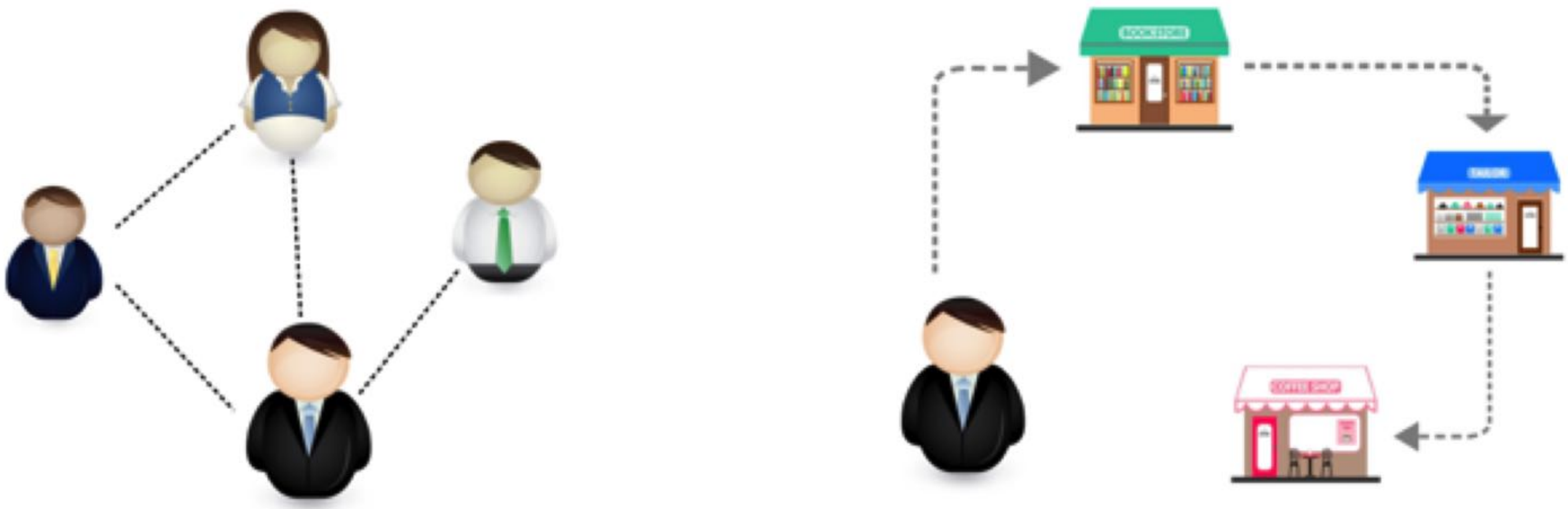
DeepWalk



1 Input: Graph
2 Random Walks
$\mathcal{W}_{v_4} =$ 4 3 1 5 1
u_k
v_j
d
j
Φ
3 Representation Mapping
v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8
$\Phi(v_1)$
4 Hierarchical Softmax
-0.6
-0.8
-1.0
-1.2
-1.4
-1.6
-1.8
−1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5
5 Output: Representation

Joint Model of Networks and Trajectories

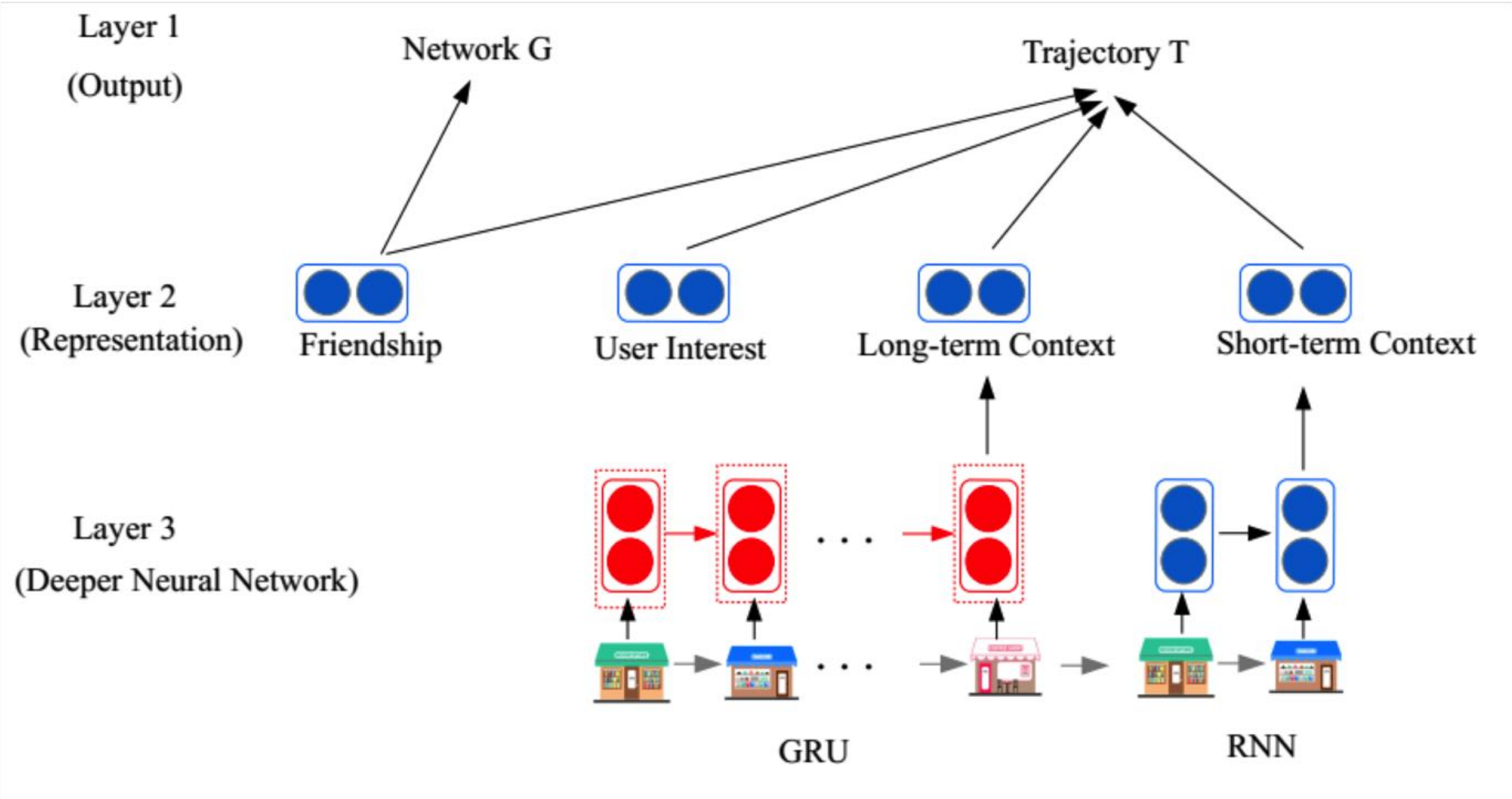
- Jointly modeling heterogeneous information of social networks and trajectories



(a) Friendship Network (b) User Trajectory

Joint Model of Networks and Trajectories

- The joint model can be achieved in the embedding space as multiple tasks



Experiment Results

- Next Position Prediction

Dataset	Brightkite			Gowalla		
Metric (%)	R@1	R@5	R@10	R@1	R@5	R@10
PV	18.5	44.3	53.2	9.9	27.8	36.3
FBC	16.7	44.1	54.2	13.3	34.4	42.3
FPMC	20.6	45.6	53.8	10.1	24.9	31.6
PRME	15.4	44.6	53.0	12.2	31.9	38.2
HRM	17.4	46.2	56.4	7.4	26.2	37.0
JNTM	**22.1**	**51.1**	**60.3**	**15.4**	**38.8**	**48.1**

- Friend Prediction

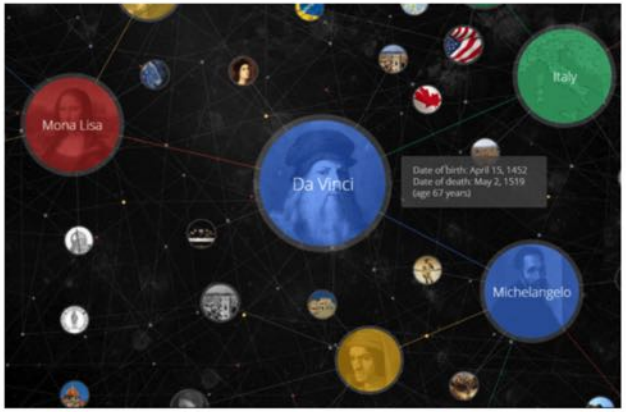
Training Ratio	20%		30%		40%		50%	
Metric (%)	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10
DeepWalk	2.6	3.9	5.1	8.1	**7.9**	**12.1**	**10.5**	**15.8**
PMF	1.7	2.4	1.8	2.5	1.9	2.7	1.9	3.1
PTE	1.1	1.8	2.3	3.6	3.6	5.6	4.9	7.6
TADW	2.1	3.1	2.6	3.9	3.2	4.7	3.6	5.4
JNTM	**3.8**	**5.5**	**5.9**	**8.9**	**7.9**	11.9	10.0	15.1

Summary

- Distributed representation is good at modeling semantic relations of heterogeneous information, with more insights about hidden semantics
- The key is how to apply it for innovative CSS

Social Networks

UGC



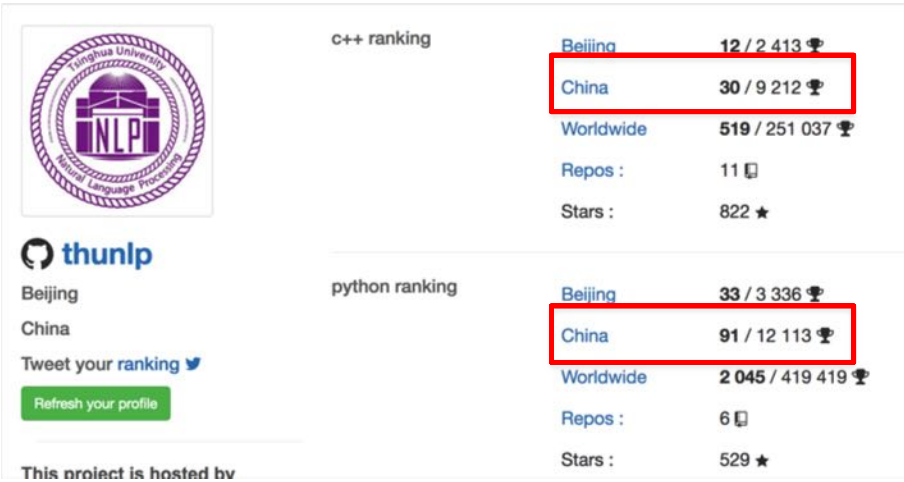
Knowledge

Open Source

- Packages for Chinese lexical analysis, keyword extraction, and representation learning

https://github.com/thunlp

THULAC： Chinese Lexical Analyzer
THUCTC： Chinese Text Classification
THUTAG： Keyword Extraction
OpenKE： Knowledge Representation Learning
OpenNE： Network Representation Learning
OpenNRE：Neural Relation Extraction
NSC： Neural Sentiment Analysis



OpenKE

http://openke.thunlp.org/

- Packages: Unified interface and implementation of the methods TransE, TransH, TransR, TransD, RESCAL, DistMult, HolE, ComplEx

- Embeddings: Learned knowledge embeddings for two widely-used large-scale KGs WikiData and Freebase

- Reading List: https://github.com/thunlp/KRLPapers

OpenNE

https://github.com/thunlp/OpenNE

- Packages: Unified interface and implementation of the methods DeepWalk, LINE, node2vec, GraRep, TADW and GCN

- Reading List: https://github.com/thunlp/nrlpapers

Thanks!

http://nlp.csai.tsinghua.edu.cn/~lzy/