Data-Driven Behavioral Analytics:
Observations, Representations and Models

Meng Jiang (UIUC)
Peng Cui (Tsinghua)
Jiawei Han (UIUC)

http://www.meng-jiang.com/tutorial-cikm16.html

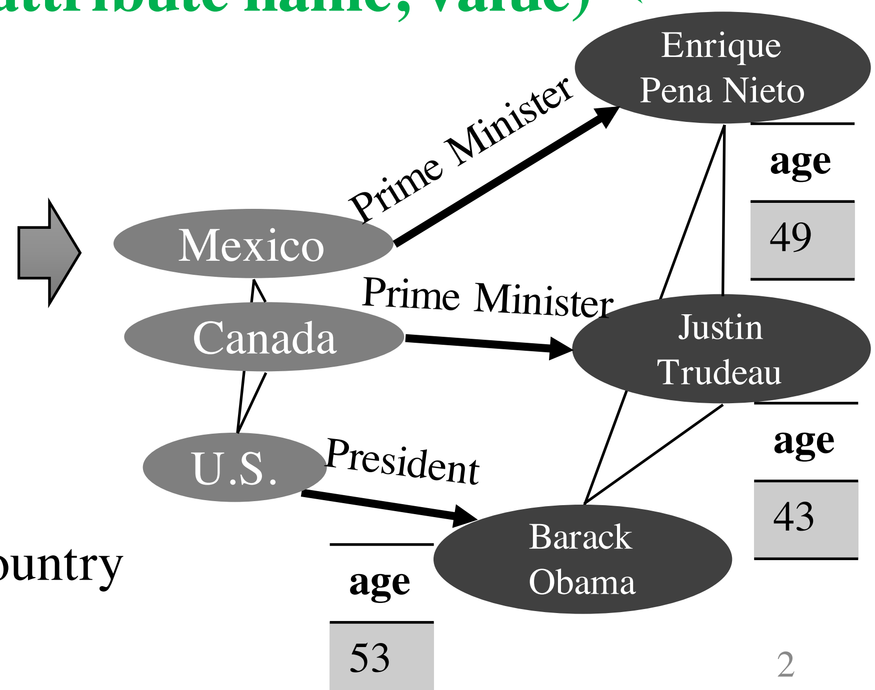
Tutorial in CIKM 2016, October 24, Indianapolis, IN

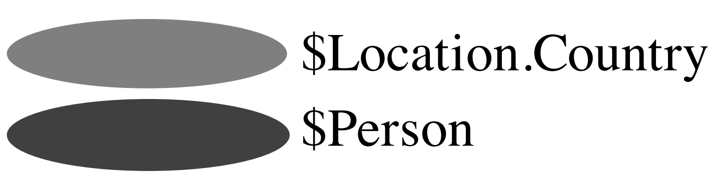
Construction of Heterogeneous Information Networks from Text

Philosophy: Not extensive "labeling" but exploring the power of massive text corpora!

- ❑ Mining phrases (the minimal semantic units)
- ❑ Entity recognition and typing
- ❑ **Attribute discovery (entity, attribute name, value)**

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…





Attribute Discovery

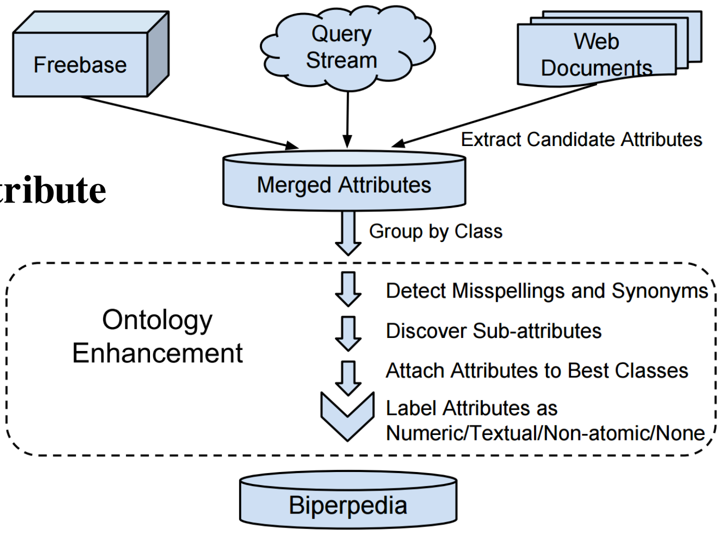
- Given text corpus (news, tweets, web documents)
  - U.S. President Barack Obama told reporters …
  - President Blaise Compaore of Burkina Faso said …
  - Canada 's Prime Minister Justin Trudeau and his wife Sophie arrived … Justin Trudeau, 43, ...
- Find
  - $COUNTRY: president, prime minister …
  - $PERSON: wife, age …
  - (U.S., president, Barack Obama)
  - (Burkina Faso, president, Blaise Compaore)
  - (Canada, prime minister, Justin Trudeau)
  - (Justin Trudeau, wife, Sophie), (Justin Trudeau, age, 43)

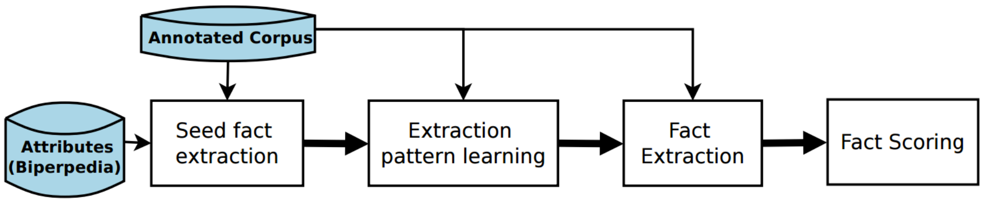
Google's Systems (Alon Halevy et al.)

- Biperpedia (VLDB'14): attribute name extraction
  - With query log ("canada prime minister"), replace entity mentions ("canada") with E and noun phrases ("prime minister") with A, and then find E-A patterns:
    - "E 's A", "E A", "A E", "A of the E", etc.
  - Take E-A patterns to web documents for attributes:
    - E: U.S., A: president ("E A")
    - E: Canada, A: prime minister ("E 's A")
  - Place the attributes on the hierarchy (Location.Country)
- ReNoun (EMNLP'14): slot filling
  - Pre-defined set of attributes, human annotations, learning

Google's Approaches on Attribute Extraction

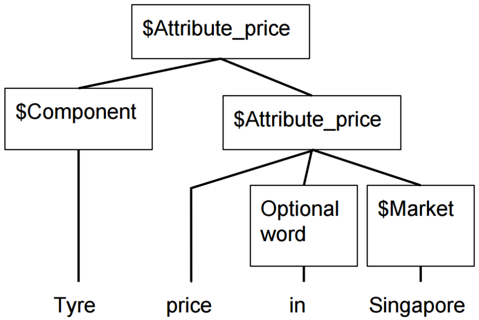
- Given Google's **query log**, web text and knowledge bases
  - "Obama wife name"… "Japan asian population", "Brazil female latino population", "Princeton economist"…
  - "Obama's wife, Michelle Obama, is a lawyer...", "Princeton economist Paul Krugman was awarded...”…
  - Obama: *$Person, $President*; Japan, Brazil: *$Location, $Country*; Princeton: *$Organization, $University*…
- Biperpedia (VLDB'14): **Attribute Name Extraction** from query log
  - $Person: wife name, daughter name
  - $Country: asian population, female latino population
  - $University: economist
- ReNoun (EMNLP'14): **Fact Extraction for Noun Phrase Attribute**
  - (Obama, wife, Michelle Obama)
  - (Princeton, economist, Paul Krugman)



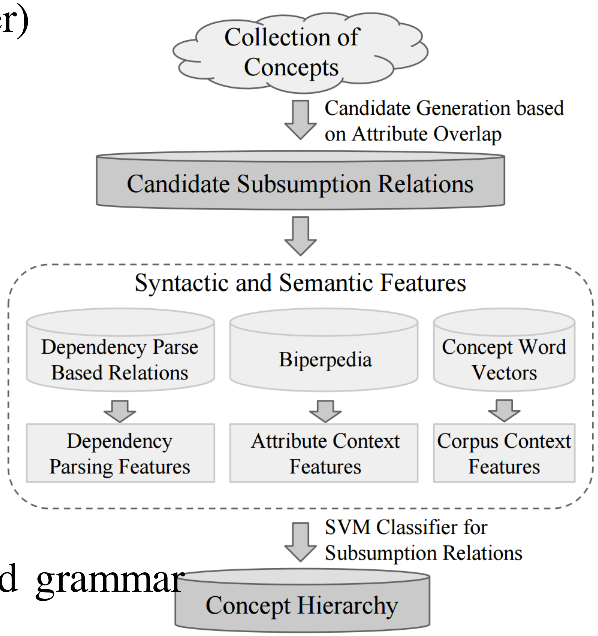


Google's Approaches on Attribute Extraction

- Latte (WebDB'15 Best Paper): **Concept (Type) Hierarchy Extraction** with attribute features
  - {country, address, zip code}: $University (sub) - $Location (super)
  - {online payment, non profit, tax return}: $University (sub) - $Organization (super)
  - {daughter name, wife name, age}: $President (sub) - $Person (super)




```
$Attribute_price ::= price
$Attribute_price ::= $Market $Attribute_price
$Attribute_price ::= $Attribute_price in $Market
$Attribute_price ::= $Component $Attribute_price
$Market          ::= Singapore | USA | Dubai | London | . . .
$Component       ::= battery | bumper | tyre | door | . . .
```

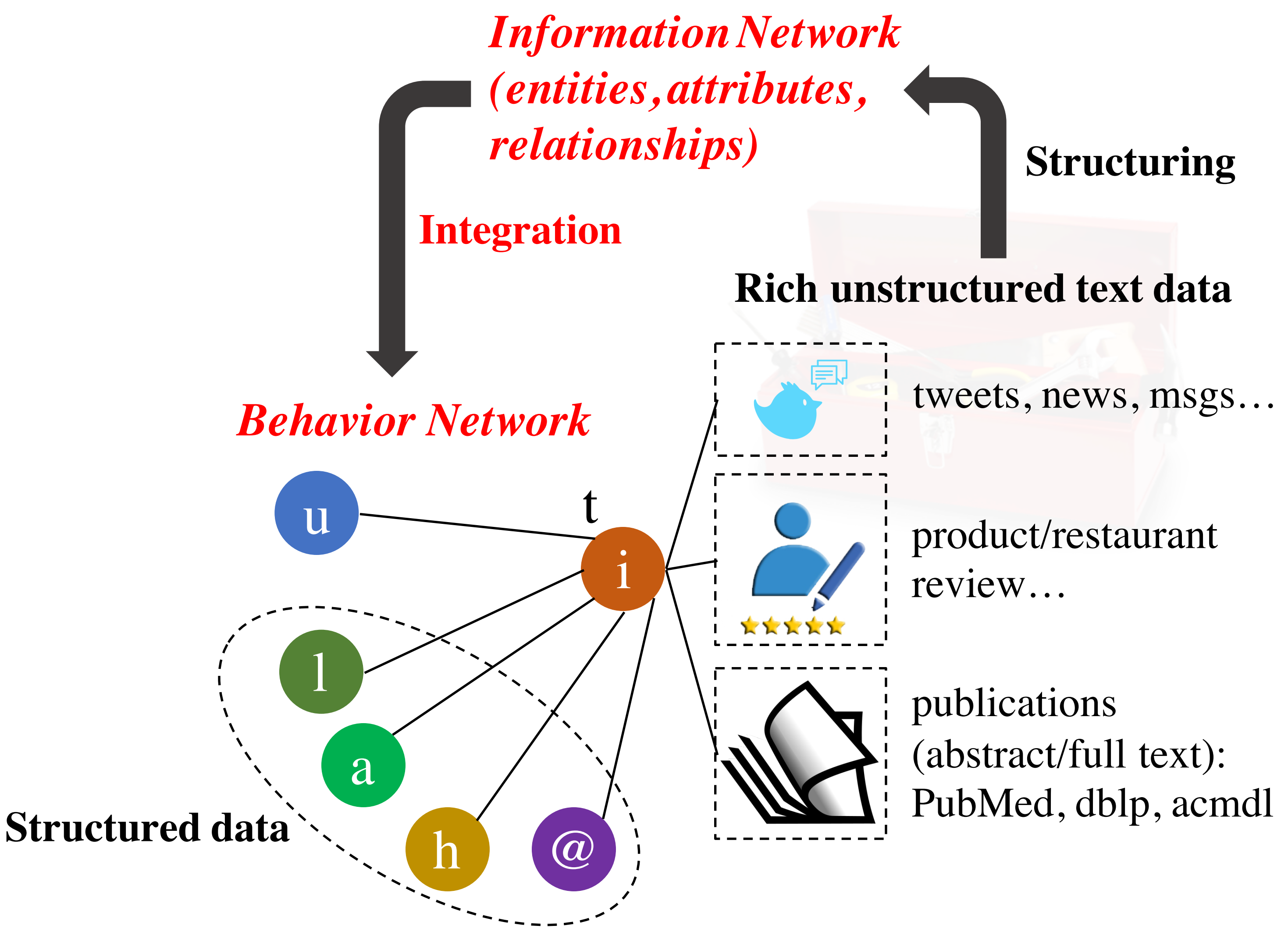


- ARI (WWW'16): **Attribute Name Structure Extraction** with rule-based grammar
  - Long-tail distribution of attribute names
  - $Person: $FamilyMember (name) - daughter, wife, mother, daughter name, wife name
  - $Country: ($Gender) ($Ethnicity) population - asian population, female latino population

Limitations

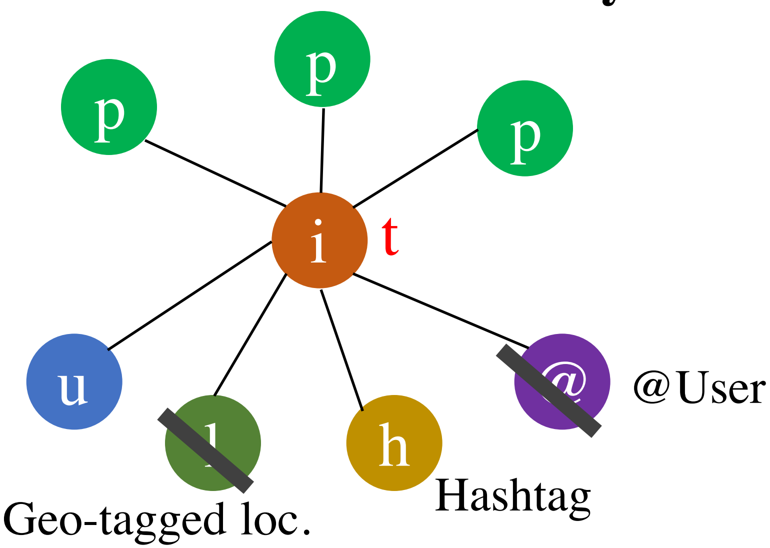
- Problem setting
  - Simultaneously extract attribute names and values?
  - Values for an open set of attribute names?
- Data sources
  - Query streams are unavailable.
  - Annotations are expensive: only for general-domain
  - Massive text corpora are unlimited!
- Poor precision and recall using E-A patterns
  - "A, E" pattern: "… yesterday, Obama …"
  - Missing long structure: "President Obama 's government of U.S." ("A1 E1 's A2 E2"???)

Data to Network to Knowledge



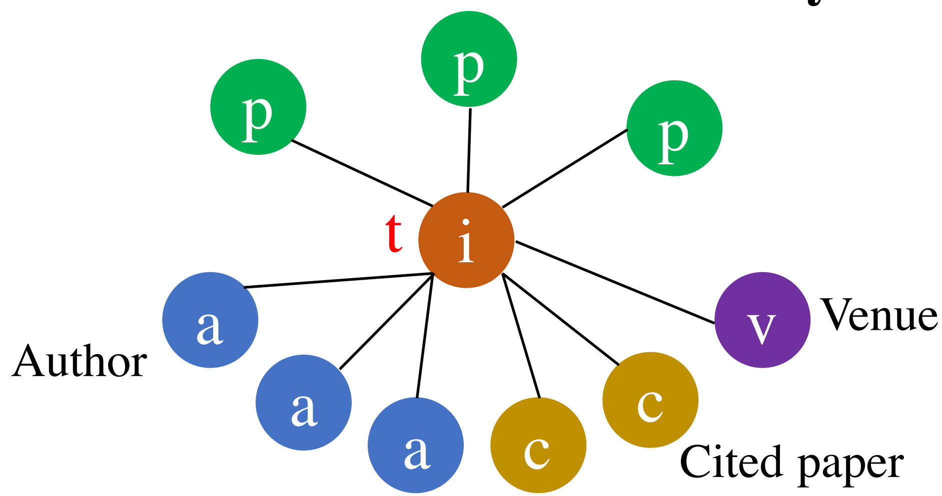
Bring Phrases to Behavior Modeling

- Tweeting behavior
  - Event **summary**



20:03:09 @ebekahwsm
this better be the best halftime show ever in the history of halftimes shows. ever. #SuperBowl

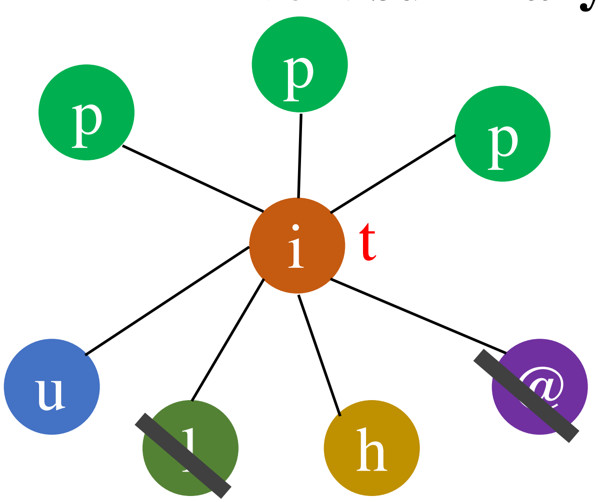
- Paper-publishing behavior
  - Research trend **summary**



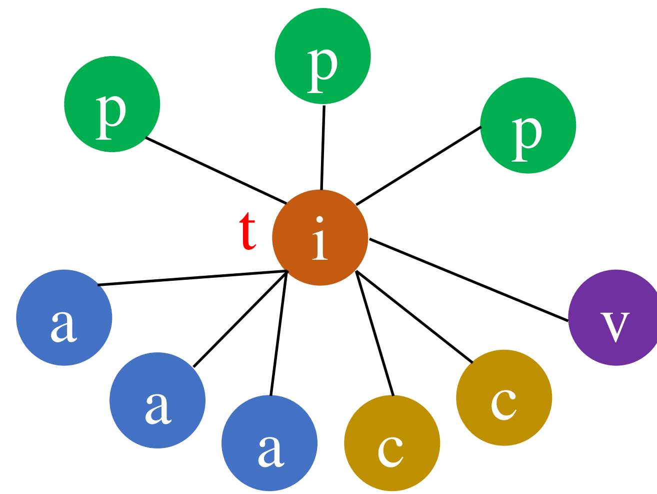
2009 P. Melville, W. Gryc, R. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", KDD'09. Refs: p81623, p84395…

Tensor Fails

- Tweeting behavior
  - Event **summary**



- Paper-publishing behavior
  - Research trend **summary**

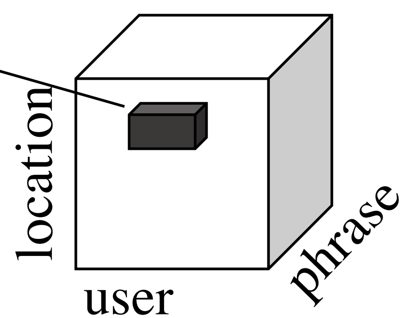


Q: How to represent and summarize **dynamic multi-contextual** behaviors?

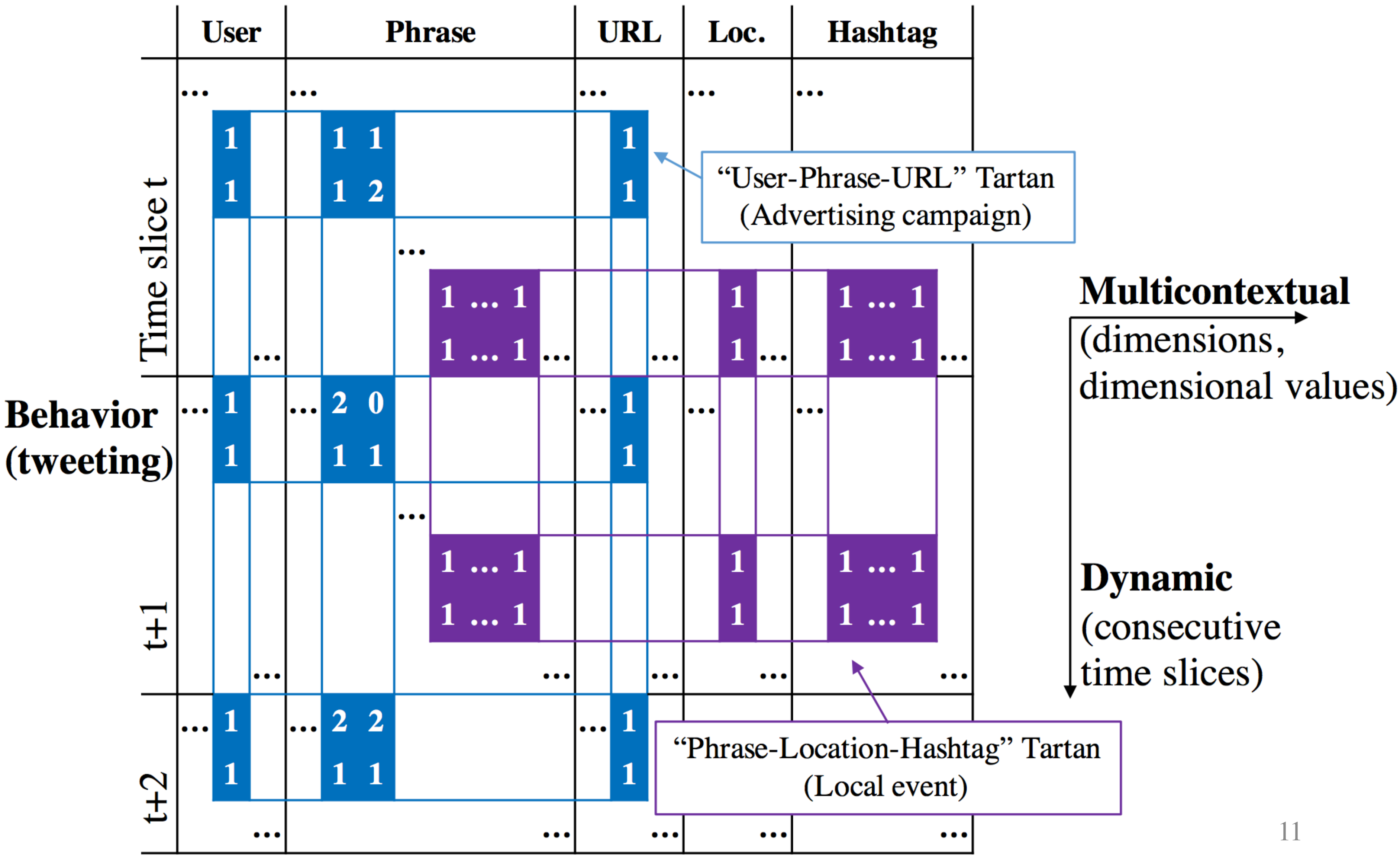
A set of values in dimensions (*one-guaranteed value*, *empty value*, *multi-values*)

✔	☐	☐

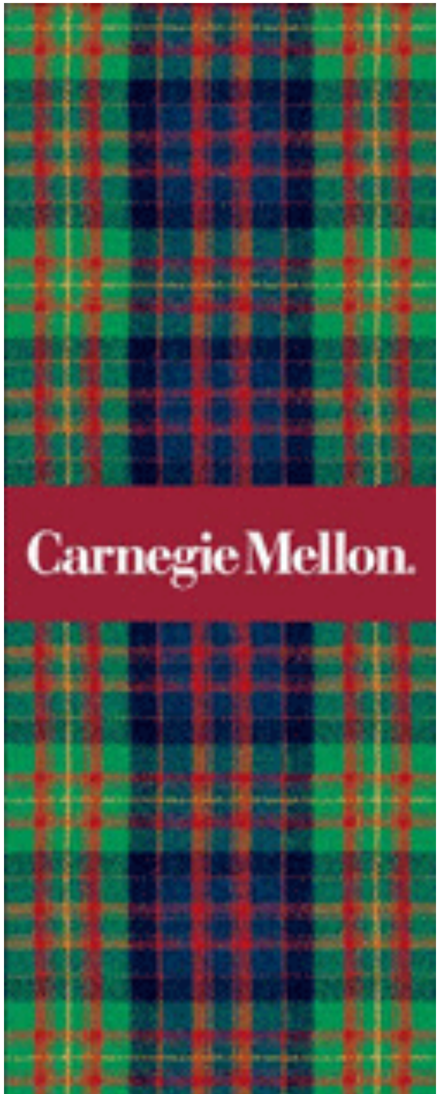
summary
(dense block)



Two-Level Matrix and “Tartan”



CMU Tartans

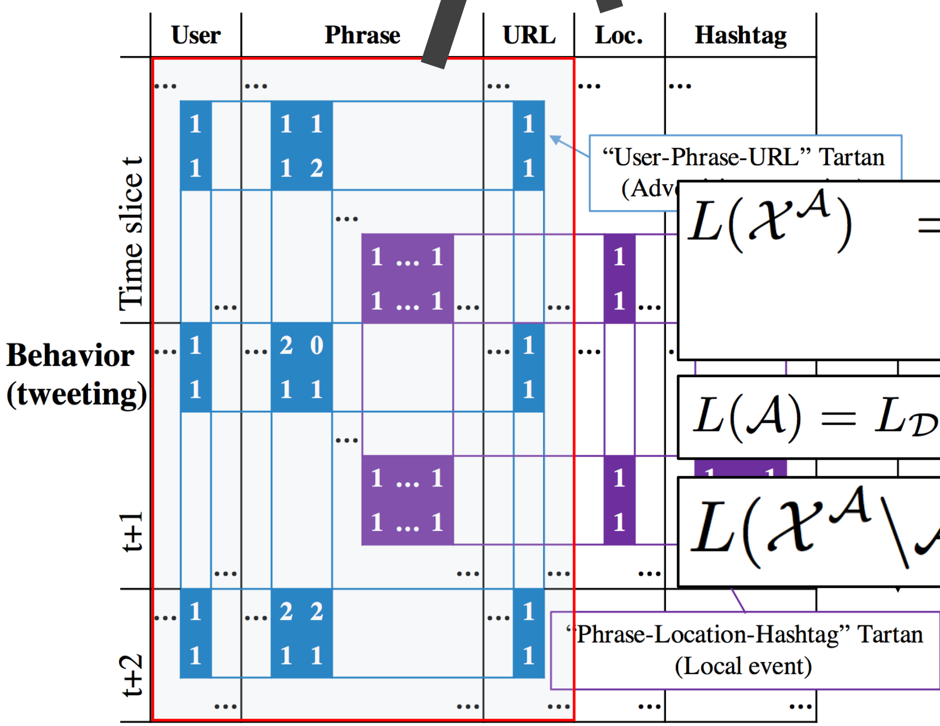


CARNEGIE MELLON UNIVERSITY
PITTSBURGH PENNSYLVANIA 1900

Optimize with MDL Principle

- *Maximize* the number of bits by encoding the Tartan

$$\boxed{f(\mathcal{A},\mathcal{X}) = L(\mathcal{X}^{\mathcal{A}}) - L(\mathcal{A}) - L(\mathcal{X}^{\mathcal{A}}\backslash\mathcal{A}).}$$

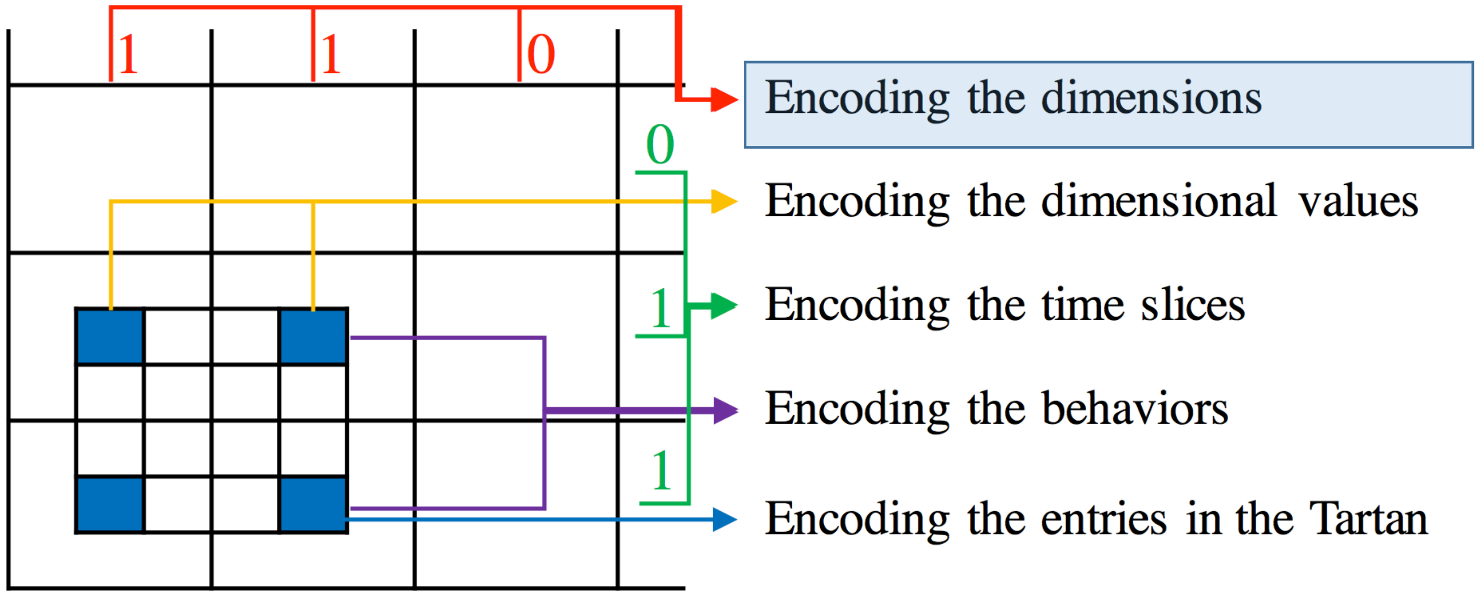


$$\begin{aligned} L(\mathcal{X}^{\mathcal{A}}) \quad &= \quad g(V+C, C) + L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) \\ &\quad + \textstyle\sum_{d\in\mathcal{D}} \log^* N_d + \sum_{t\in\mathcal{T}} \log^* E^{(t)}. \end{aligned}$$

$$L(\mathcal{A}) = L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{V}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) + L_{\mathcal{B}}(\mathcal{A}) + L_{A}(\mathcal{A}).$$

$$L(\mathcal{X}^{\mathcal{A}}\backslash\mathcal{A}) = g(V+C-v-c, C-c);$$

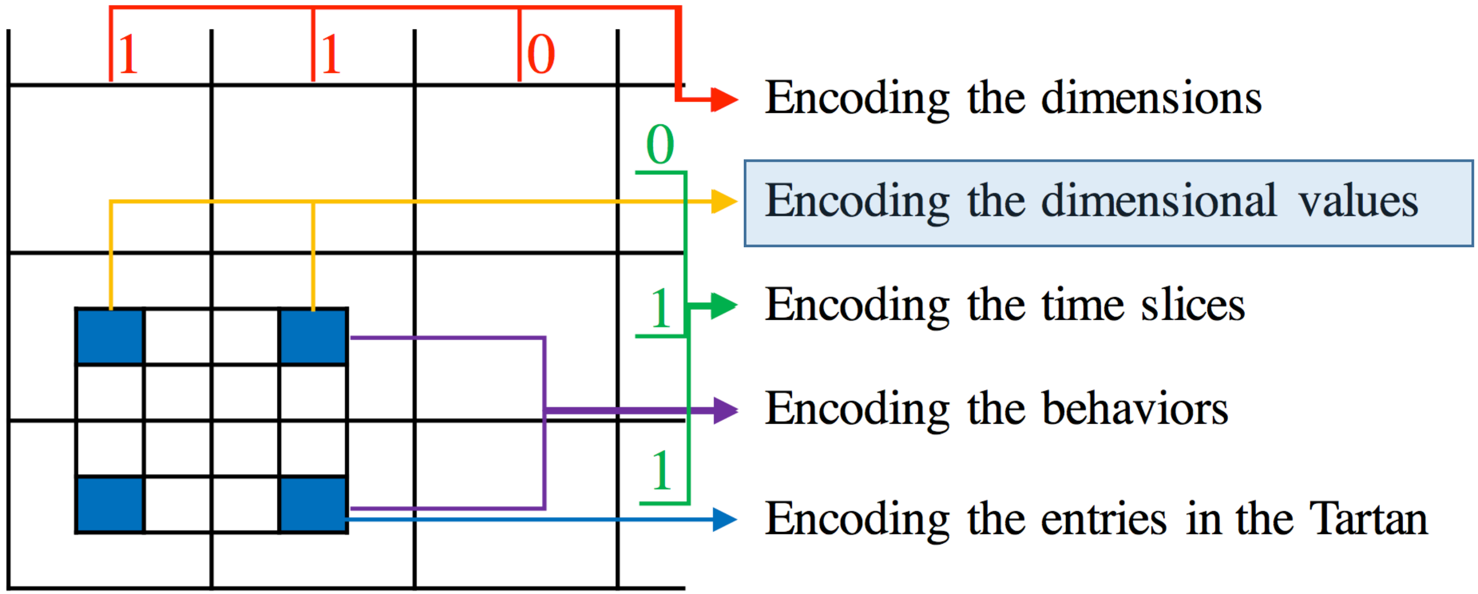
Encoding Tartan: Dimensions



$$
\begin{aligned}
H_{\mathcal{D}}(X) &= -\textstyle\sum_{x\in\{0,1\}} P(X=x)\log P(X=x) \\
&= -\left(\frac{D^{\mathcal{A}}}{D}\log\frac{D^{\mathcal{A}}}{D} + \frac{D-D^{\mathcal{A}}}{D}\log\frac{D-D^{\mathcal{A}}}{D}\right).
\end{aligned}
$$

$$
\begin{aligned}
L_{\mathcal{D}}(\mathcal{A}) &= \log^* D + \log^* D^{\mathcal{A}} + D\cdot H_{\mathcal{D}}(X) \\
&= \log^* D + \log^* D^{\mathcal{A}} + g(D, D^{\mathcal{A}}),
\end{aligned}
$$

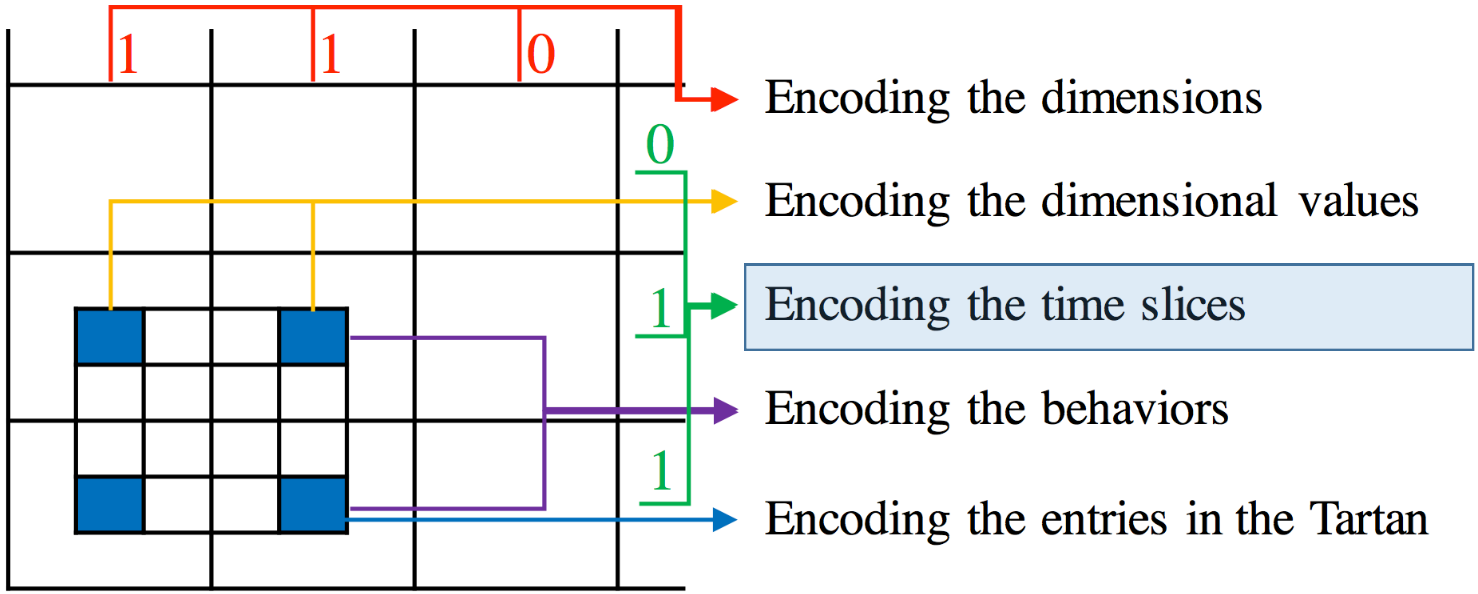
Encoding Tartan: Dimensional Values



$$H_{\mathcal{V}_d}(X) \quad = \quad -\left(\frac{n_d}{N_d}\log\frac{n_d}{N_d} + \frac{N_d - n_d}{N_d}\log\frac{N_d - n_d}{N_d}\right).$$

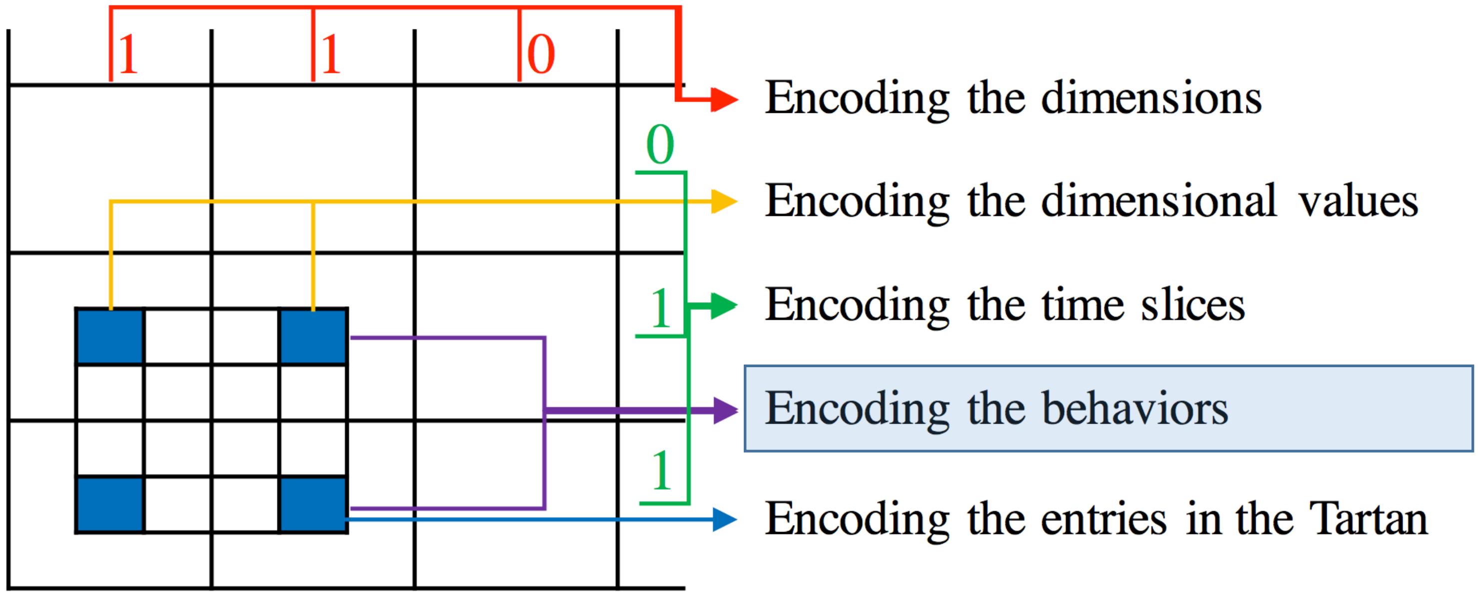
$$L_{\mathcal{V}}(\mathcal{A}) \quad = \quad \sum_{d\in\mathcal{D}}\left(\log^* N_d + \log^* n_d + g(N_d, n_d)\right).$$

Encoding Tartan: Time Slices



$$L_{\mathcal{T}}(\mathcal{A}) \quad = \quad \log^* T + \log^* T^{\mathcal{A}} + \log^* t_{start}$$

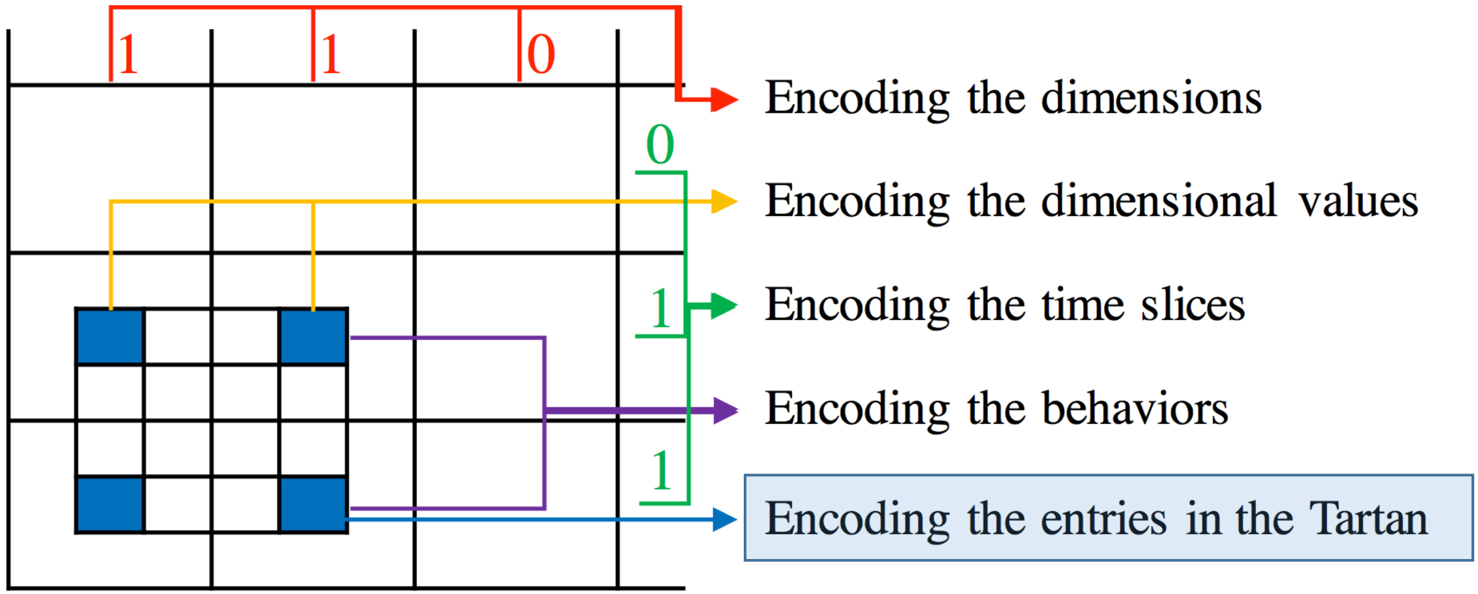
Encoding Tartan: Behaviors



$$H_{\mathcal{B}^{(t)}}(X) = -\left(\frac{e^{(t)}}{E^{(t)}} \log \frac{e^{(t)}}{E^{(t)}} + \frac{E^{(t)}-e^{(t)}}{E^{(t)}} \log \frac{E^{(t)}-e^{(t)}}{E^{(t)}}\right).$$

$$L_{\mathcal{B}}(\mathcal{A}) = \sum_{t\in\mathcal{T}} \left(\log^* E^{(t)} + \log^* e^{(t)} + g(E^{(t)}, e^{(t)})\right).$$

Encoding Tartan: Entries



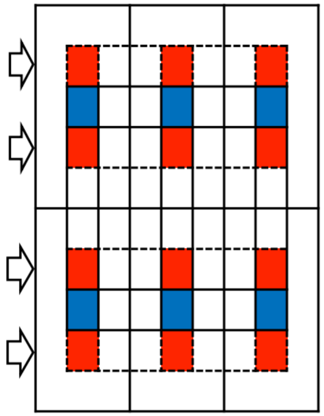
$$v = \left(\sum_{d \in \mathcal{D}} n_d\right)\left(\sum_{t \in \mathcal{T}} e^{(t)}\right).$$

$$c = \sum_{d \in \mathcal{D}, t \in \mathcal{T}} \sum_{b \in \mathcal{B}^{(t)}, i \in \mathcal{V}_d} \mathcal{X}_d^{(t)}(b, i).$$

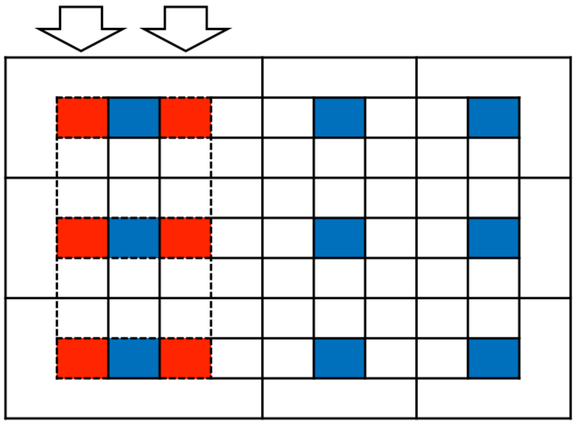
$$H_{\mathcal{A}}(X) = -\left(\frac{c}{v+c} \log \frac{c}{v+c} + \frac{v}{v+c} \log \frac{v}{v+c}\right).$$

$$L_{\mathcal{A}}(\mathcal{A}) = (v+c) H_{\mathcal{A}}(X) = g(v+c, c).$$

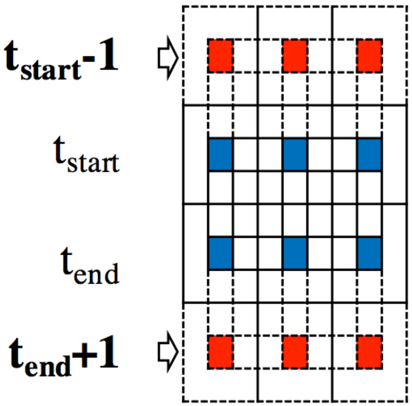
Greedy Search for the Local Optimum



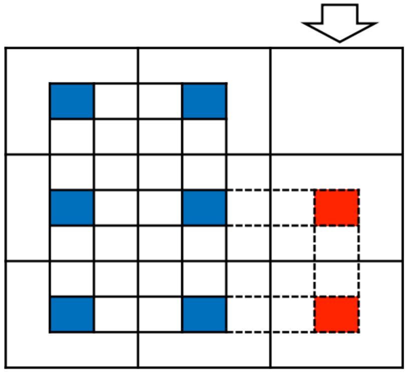
(a) Update the set of behaviors.



(b) Update the set of values.



(c) Update the consecutive time slices.



(d) Update the set of dimensions.

Time complexity:

$$\mathcal{O}(\sum_d N_d \log N_d + \sum_t E^{(t)} \log E^{(t)})$$

Experimental Results

- DM/ML research trend summaries with DBLP data

1997 2000 2003 2006 2009 2012

Author	Venue	Keyword	Cited	#Paper
76 Cheng-xiang Zhai Hui Fang S. Kambhampati	**7** SIGIR VLDB TKDE	**7** "information retrieval" "data integration" "text classification"	**68** p56743[1] p62995 p76869	**32** 2003- 2007

[1] "A language modeling approach to information retrieval"

Venue	Keyword	#Paper
5 ICML NIPS …	**6** "reinforcement learning" "machine learning"	**40** 1997- 2002

Author	Venue	Cited	#Paper
6 Jiawei Han Xifeng Yan	**1** SIG- MOD	**1** p76095[2]	**22** 2004- 2010

[2] "Frequent subgraph discovery"

Venue	Keyword	#Paper
3 ICDM AAAI TKDE	**1** "anomaly detection"	**25** 2005- 2013

Author	Venue	Keyword	#Paper
27 C. Faloutsos J. Pei P. S. Yu X. Lin C. Aggarwal…	**6** KDD ICDM ICDE TKDE …	**12** "large graphs" "data streams" "evolving data" "evolving graphs" …	**70** 2006- 2013

Author	Venue	Keyword	Cited	#Paper
12 Ryen White Hang Li Tie-Yan Liu Zhaohui Zheng…	**5** SIGIR WWW WSDM CIKM…	**3** "web search" "click-through data" "sponsored search"	**12** p82630[3] p116290 p103899 p106191…	**32** 2006- 2013

[3] "Optimizing search engines using clickthrough data"

Author	Venue	Keyword	#Paper
8 Qiang Yang Dou Shen Sinno Pan…	**3** KDD PAKDD AAAI	**6** "transfer learning" "data mining" "localization models"	**17** 2007- 2010

Experimental Results

- Event summaries with Super Bowl 2013 tweets

16:30 | 17:00 | 17:30 | 18:00 | 18:30 | 19:00 | 19:30 | 20:00 | 20:30 | 21:00 | 21:30 | 22:00

16:30:31 My prediction Ravens 34 Niners 31 *16:30:57 Ready for the big game :D, my prediction 24-20 SF #SuperBowl* *16:31:14 My prediction for superbowl.. 48.. Jets over Bears 17-13 Mark Sanchez MVP* *16:32:24 I predict Baltimore Ravens will win 27 to 24 or 25 or 26. Basically it will be a close game.*	"my prediction"	user (3,325)	**phrase** 226	hashtag (0)	URL (0)	3,397 tweets	**Tartan #1: (1 dim)** 16:30-17:30
17:30:51 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist *17:31:01 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist* *17:31:16 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist* *17:31:19 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist*	"make your prediction"	user (196)	**phrase** 4	**RT @user** 1	**URL** 1	196 tweets	**Tartan #2: (3 dims)** 17:00-18:00
18:55:03 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47 *18:55:04 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47* *18:55:44 RT @Ravens: David Akers is good from 36 yards to make the score 7-3 Ravens. Nice job by the defense to tighten up in the red zone.*	"7-3", "1st Qtr"	user (213)	**phrase** 21	**RT @user** 3	URL (0)	215 tweets	**Tartan #3: (2 dims)** 18:30-19:30
20:20:01 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6 *20:20:02 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs* *20:20:04 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6* *20:20:05 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs*	halftime show"	user (617)	**phrase** 11	**RT @user** 4	**URL** 4	617 tweets	**Tartan #4: (3 dims)** 20:00-21:00
20:20:47 (Manhattan, NY)...and every one of those girls took #ballet #Beyonce #superbowl *20:22:01 (New York, NY) I have the biggest lady boner for Beyonce #BeyonceBowl #DestinyBowl #DestinysChild #SuperBowl* *20:24:32 (Manhattan, NY) No one can ever top that performance by Beyonce. EVER. #Beyonce #superbowl #halftimeshow*	"beyonce", #beyonce, #superbowl, #DestinysChild	**location** 2	**phrase** 55	**hashtag** 17	URL (0)	166 tweets	**Tartan #5: (3 dims)** 20:00-21:00
21:44:42 Ahora si pff #49ers 23-28 #Ravens *21:44:44 Baltimore #Ravens 28-23 San Francisco #49ers* *21:44:50 FG Akers #49ers 23-28 #Ravens 3Q 3:10 #SuperBolwXLVII #SuperBowl #NFL*	"28-23", #49ers, #Ravens	user (650)	**phrase** 69	**hashtag** 11	URL (0)	653 tweets	**Tartan #6: (2 dims)** 21:00-22:00
22:42:27 Congratulations Ravens!!!! *22:42:43 Congratulations Ray Lewis and the Ravens.* *22:42:43 Game over.! Ravens won ray got his retirement ring now all y'all boys and girls go to sleep .!* *22:42:52 "@LetThatBoyTweet: Game over. Ravens win the Super Bowl."*	"congratulations", "game over"	user (1942)	**phrase** 248	hashtag (0)	URL (0)	1,950 tweets	**Tartan #7: (1 dim)** 22:00-23:30

Summary

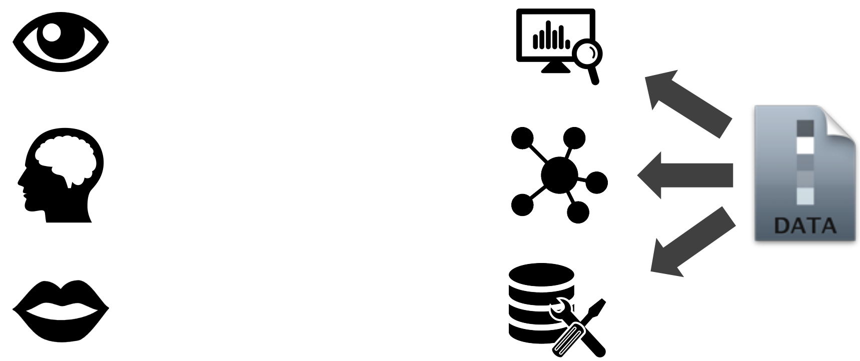
- Structuring text into heterogeneous information networks
- **Observations, Representations, Models**
  - **ToPMine/SegPhrase:** Quality phrase mining
  - **ClusType:** Entity recognition and typing
  - **MetaPAD:** Data-driven automatic attribute discovery for attributed network construction
    - Integrating text mining techniques
    - **Meta Pattern** Mining
- Integrating phrases into behavioral analysis
- **Observations, Representations, Models**
  - **CatchTartan:** Dynamic multicontextual. Tensor fails.

Observation

Representation

Models/Algorithms

Experience-Driven | **Data-Driven**

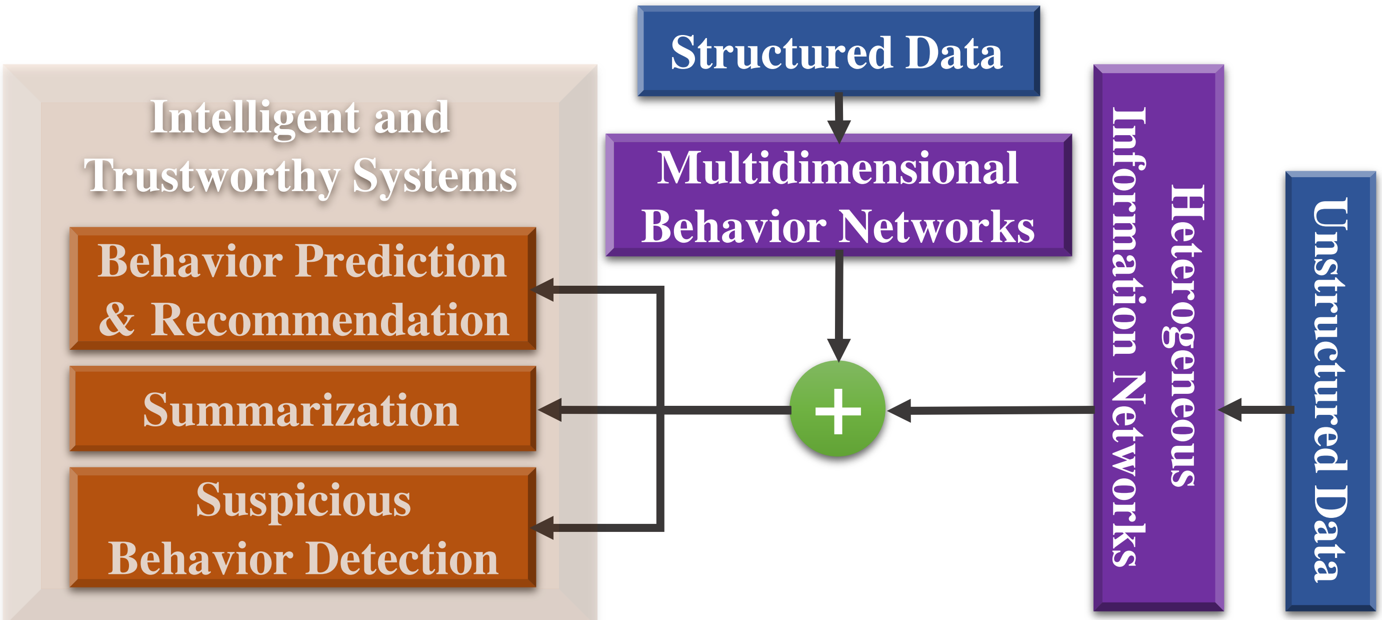


Conclusion

Data-Driven Behavioral Analytics

Data-Driven Behavioral Analytics

- Mining **behavior networks** with **social and spatiotemporal contexts** to support intelligent and trustworthy systems
  - Mining for **behavior prediction and recommendation**
  - Mining for **suspicious behavior detection**
- Structuring behavioral content and integrating behavioral analysis with **information networks**



Acknowledgement

NSF
NSFC
National Natural Science Foundation of China
Carnegie Mellon University
University of Maryland 1856
Microsoft Research
微软亚洲研究院
Tencent 腾讯

References

D. Blei, A. Ng, and M. Jordan. “Latent dirichlet allocation.” JMLR, 2003.

J. Herlocker, J. Konstan, L. Terveen, J. Riedl. “Evaluating collaborative filtering recommender systems.” ACM TOIS, 2004.

Y. Koren, R. Bell, C. Volinsky. “Matrix factorization techniques for recommender systems.” Computer, 2009.

Y. Koren. “Factorization meets the neighborhood: A multifaceted collaborative filtering model.” KDD, 2008.

Y. Koren. “Collaborative filtering with temporal dynamics.” CACM, 2010.

M. Balabanovic and Y. Shoham. “FAB: Content-based, collaborative recommendation.” CACM, 1997.

N. Liu and Q. Yang. “Eigenrank: A ranking-oriented approach to collaborative filtering.” SIGIR, 2008.

N. Liu, M. Zhao, and Q. Yang. “Probabilistic latent preference analysis for collaborative filtering.” CIKM, 2009.

References

H. Ma, H. Yang, M. Lyu, and I. King. "Sorec: Social recommendation using probabilistic matrix factorization." CIKM, 2008.

H. Ma, T. Zhou, M. Lyu, and I. King. "Improving recommender systems by incorporating social contextual information." ACM TOIS, 2011.

H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King. "Recommender systems with social regularization." WSDM, 2011.

J. Leskovec, A. Singh, and J. Kleinberg. "Patterns of influence in a recommendation network." PAKDD, 2006.

P. Massa and A. Paolo. "Trust-aware recommender systems." RecSys, 2007.

M. Jamali and E. Martin. "TrustWalker: A random walk model for combining trust-based and item-based recommendation." KDD, 2009.

H. Ma, I. King, and M. Lyu. "Learning to recommend with social trust ensemble." SIGIR, 2009.

H. Ma. I. King, and M. Lyu. "Learning to recommend with explicit and implicit social relations." ACM TIST, 2011.

References

M. Faloutsos, P. Faloutsos, and C. Faloutsos. "On power-law relationships of the internet topology." SIGCOMM, 1999.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Weiner. "Graph structure in the web." Computer Networks, 2000.

F. Chung and L. Lu. "The average distances in random graphs with given expected degrees." PNAS, 2002.

J. Kleinberg. "Authoritative sources in a hyperlinked environment." JACM, 1999.

H. Kwak, C. Lee, H. Park, and S. Moon. "What is Twitter, a social network or a news media?" WWW, 2010.

B. Hooi, H.A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. "Fraudar: Bounding graph fraud in the face of camouflage." KDD, 2016.

C. Aggarwal and J. Han. "Frequent pattern mining." Springer, 2014.

J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu. "FreeSpan: Frequent pattern-projected sequential pattern mining." KDD, 2000.

References

X. Yan and J. Han. "gspan: Graph-based substructure pattern mining." ICDM, 2003.

X. Yan and J. Han. "CloseGraph: Mining closed frequent graph patterns." KDD, 2003.

Y. Sun, J. Han, X. Yan, P.S. Yu, and T. Wu. "PathSim: Meta path-based top-k similarity search in heterogeneous information networks." VLDB, 2011.

Y. Sun, Y. Yu, and J. Han. "Ranking-based clustering of heterogeneous information networks with star network schema." KDD, 2009.

Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. "RankClus: Integrating clustering with ranking for heterogeneous information network analysis." EDBT, 2009.

Y. Sun, R. Barber, M. Gupta, C. Aggarwar, and J. Han. "Co-author relationship prediction in heterogeneous bibliographic networks." ASONAM, 2011.

A. El-Kishky, Y. Song, C. Wang, C.R. Voss, and J. Han. "Scalable topical phrase mining from text corpora." VLDB, 2014.

J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. "Mining quality phrases from massive text corpora." SIGMOD, 2015.

References

X. Ren, A. El-Kishky, C. Wang, F. Tao, C.R. Voss, and J. Han. "Effective entity recognition and typing by relation phrase-based clustering." KDD, 2015.

X. Ren, W. He, M. Qu, C.R. Voss, H. Ji, and J. Han. "Label noise reduction in entity typing by heterogeneous partial-label embedding." KDD, 2016.

C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han. "A phrase mining framework for recursive construction of a topical hierarchy." KDD, 2013.

E.E. Papalexakis, C. Faloutsos, N.D. Sidiropoulos. "ParCube: Sparse parallelizable tensor decompositions." PKDD, 2012.

D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos. "VOG: Summarizing and understanding large graphs." SDM, 2014.

R. Gupta, A. Halevy, X. Wang, S.E. Whang, and F. Wu. "Biperpedia: An ontology for search applications." VLDB, 2014.

M. Yahya, S. Whang, R. Gupta, and A. Halevy. "ReNoun: Fact extraction for nominal attributes." EMNLP, 2014.

A. Halevy, N. Noy, S. Sarawagi, S.E. Whang, and X. Yu. "Discovering structure in the universe of attribute names." WWW, 2016.

References

Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. “Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation.” SIGMOD, 2014.

Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. “A confidence-aware approach for truth discovery on long-tail data.” VLDB, 2014.

F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. “Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation.” KDD, 2015.

Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. “A survey on truth discovery.” KDD Explorations Newsletter, 2016.

S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. “Modeling truth existence in truth discovery.” KDD, 2015.

S. Kumar, R. West, and J. Leskovec. “Disinformation on the Web: Impact, characteristics, and detection of Wikipedia hoaxes.” WWW, 2016.

S. Kumar, F. Spezzano, and V.S. Subrahmanian. “Identifying malicious actors on social media.” ASONAM, 2016. (tutorial)

Thank you!

Data-Driven Behavioral Analytics:
Observations, Representations and Models