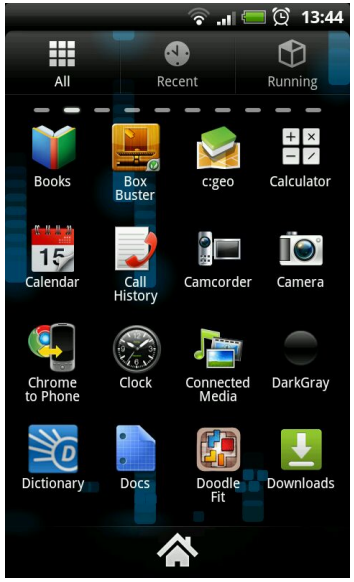
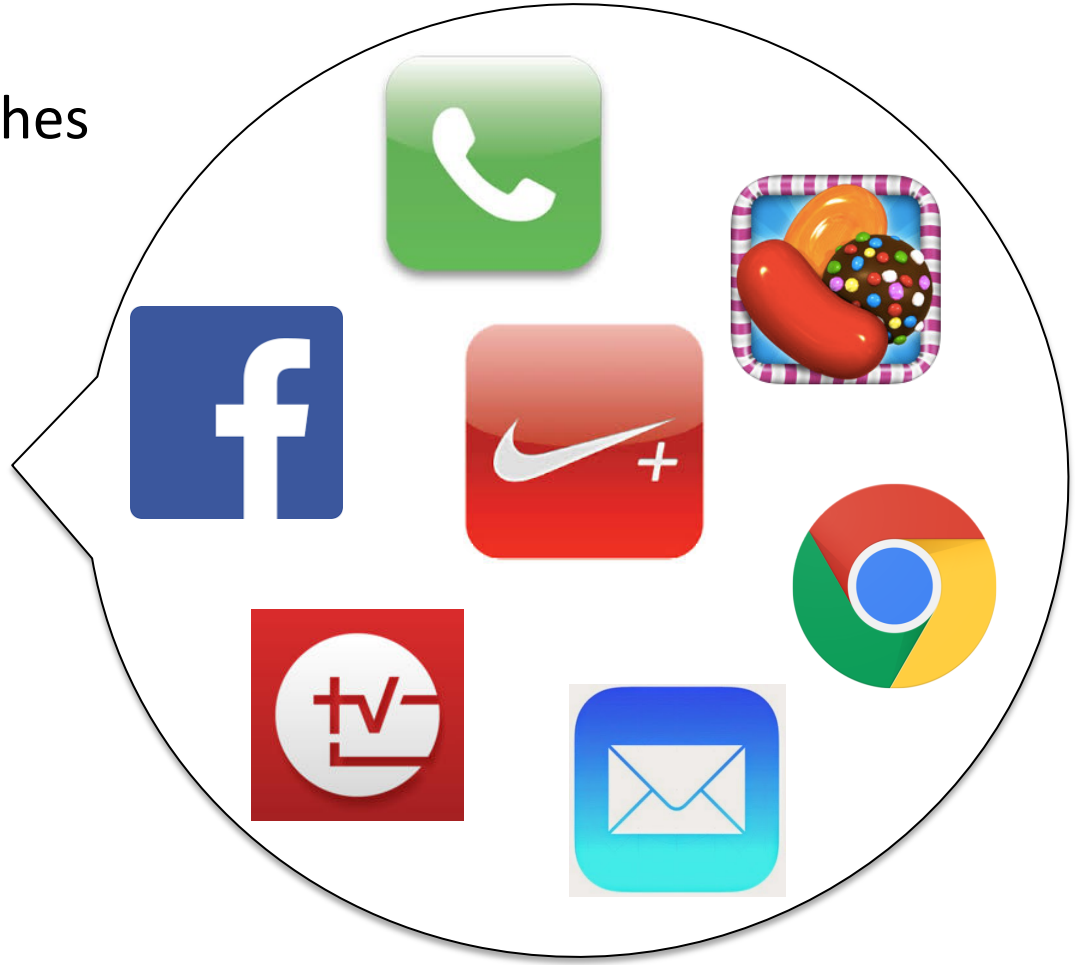
Problem Statement (1)

- Smartphone logs is a diverse and rich dataset

Application Launches





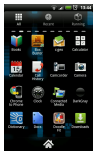
Problem Statement (1)

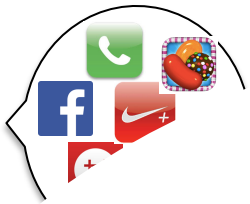
- Smartphone logs is a diverse and rich dataset

Notifications

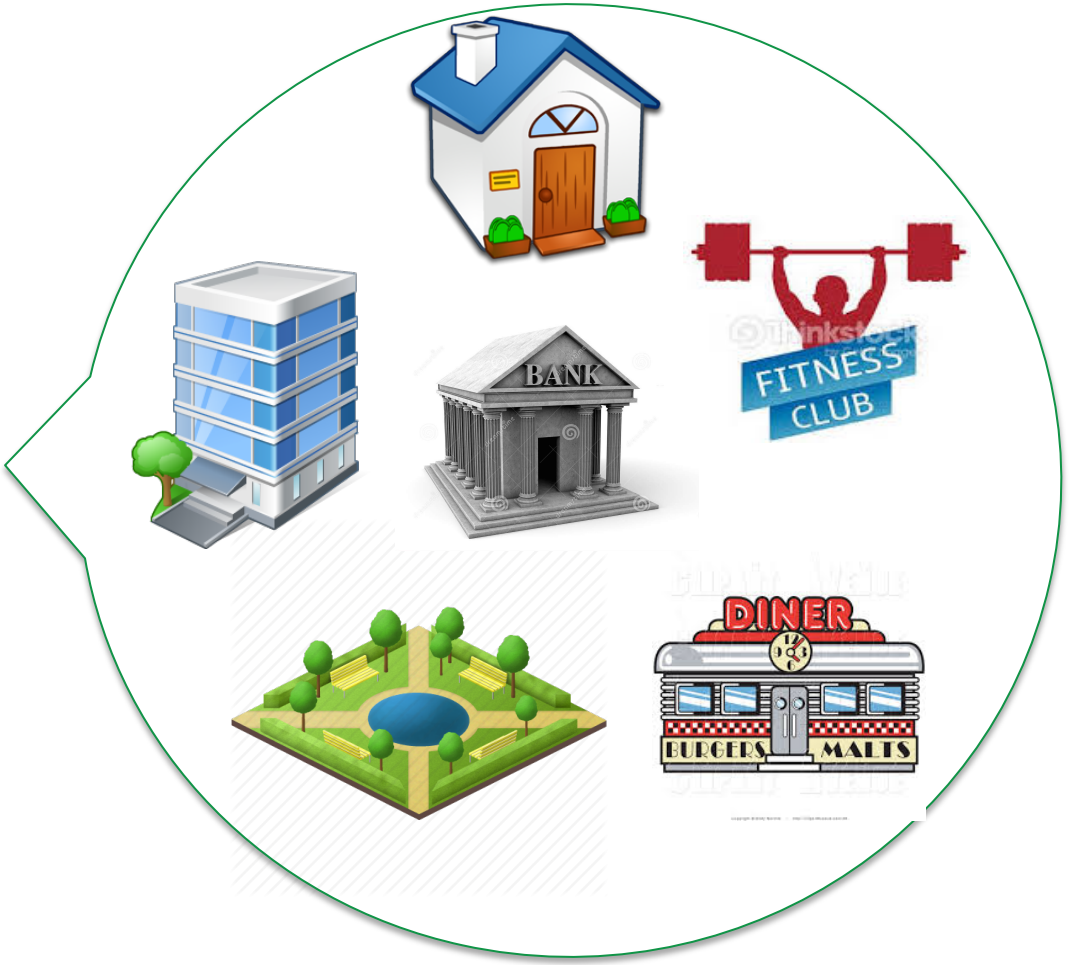
Problem Statement (1)

- Smartphone logs is a diverse and rich dataset



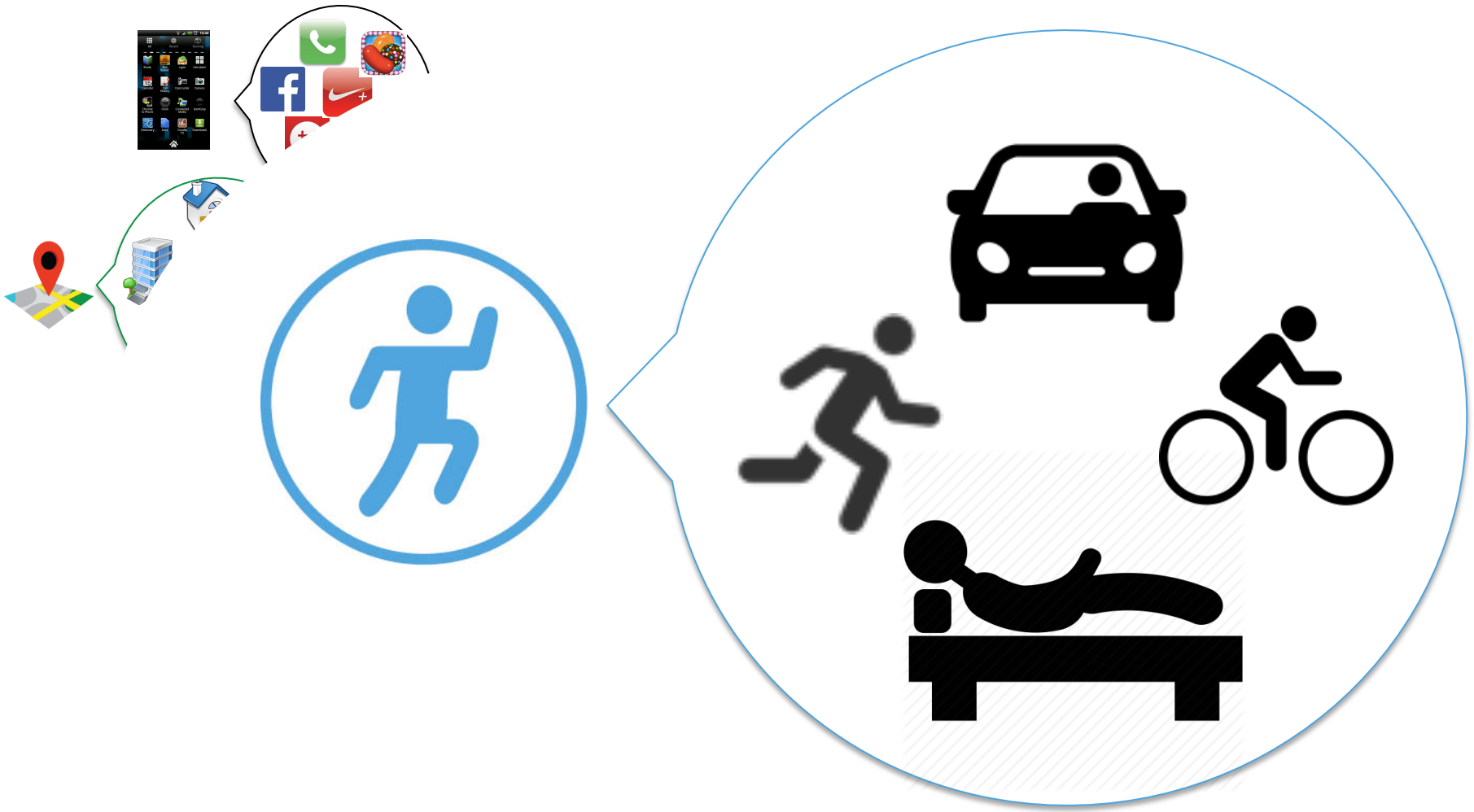


Locations



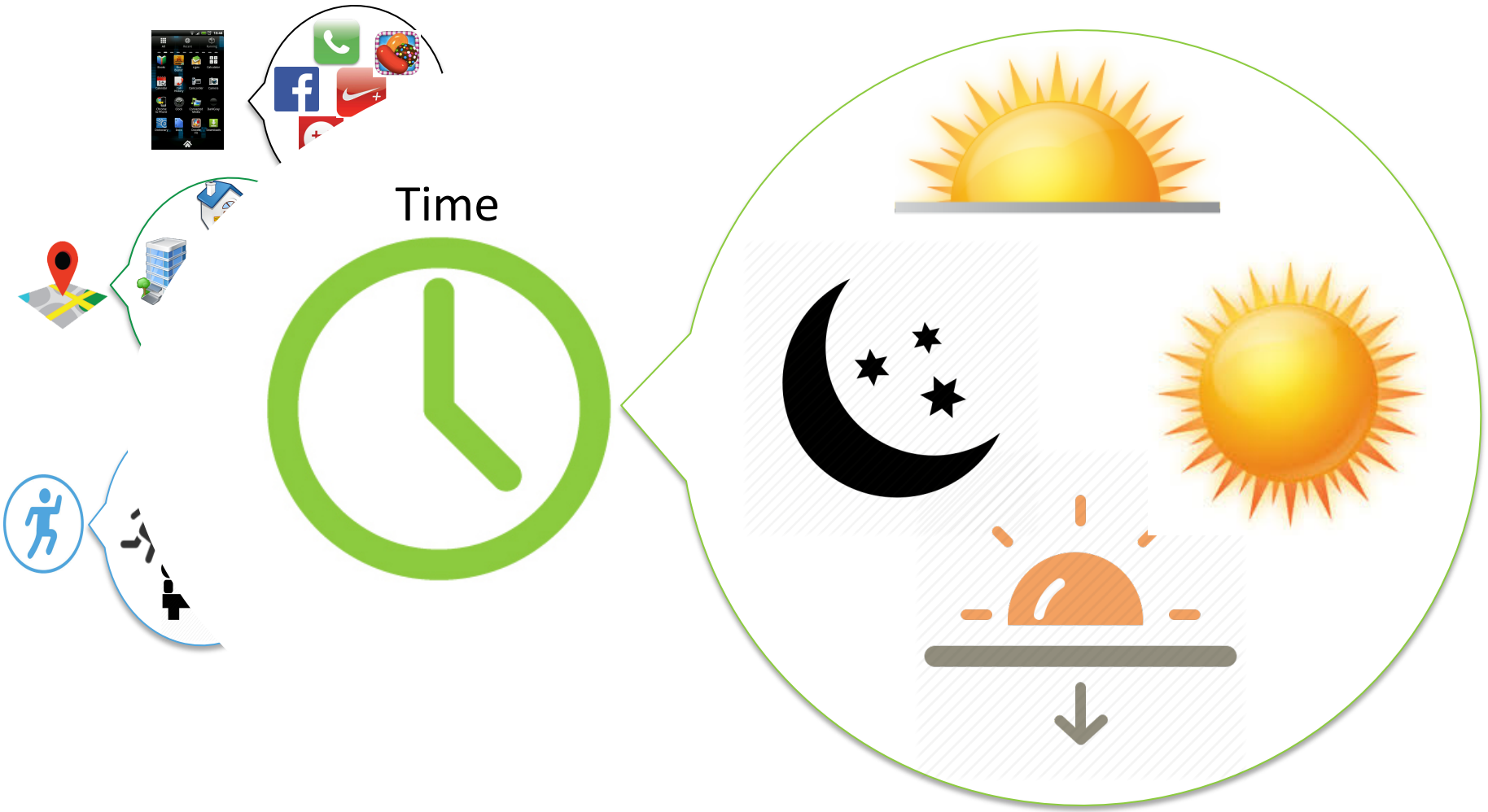
Problem Statement (1)

- Smartphone logs is a diverse and rich dataset



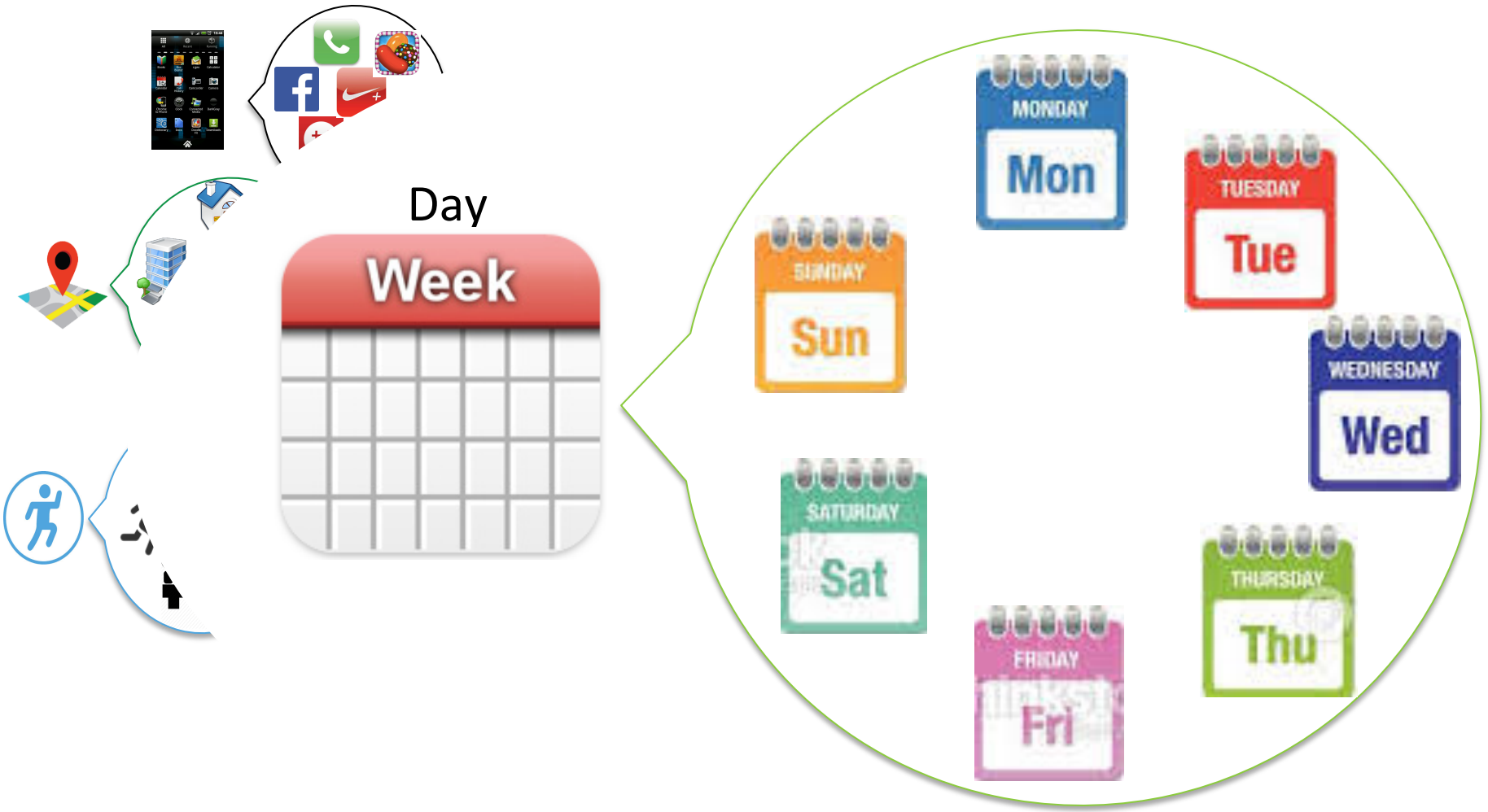
Problem Statement (1)

- Smartphone logs is a diverse and rich dataset



Problem Statement (1)

- Smartphone logs is a diverse and rich dataset



Problem Statement (1)

- Smartphone logs is a diverse and rich dataset

Problem Statement (1)

- Use smartphone logs to discover user behaviors and habits

Bob habits

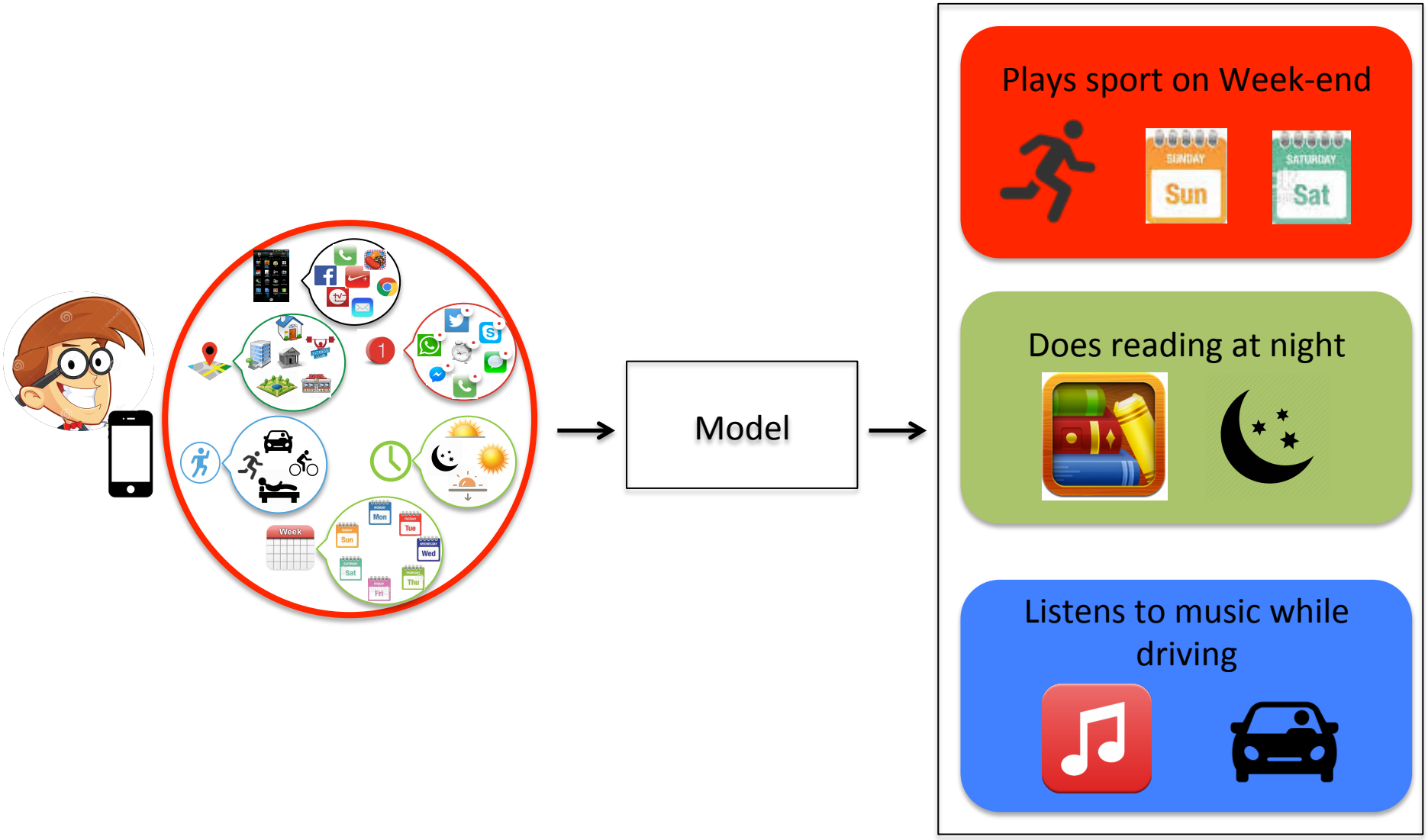
Plays sport on Week-end

Does reading at night

Listens to music while driving

Problem Statement (1)

- Use smartphone logs to discover user behaviors and habits



Problem Statement (2)

- **Goal:** Use smartphone logs to discover user behaviors and habits in an unsupervised manner
  - Is it possible?
  - How precise can it be?

- The smartphone logs of a unique user are considered at one time

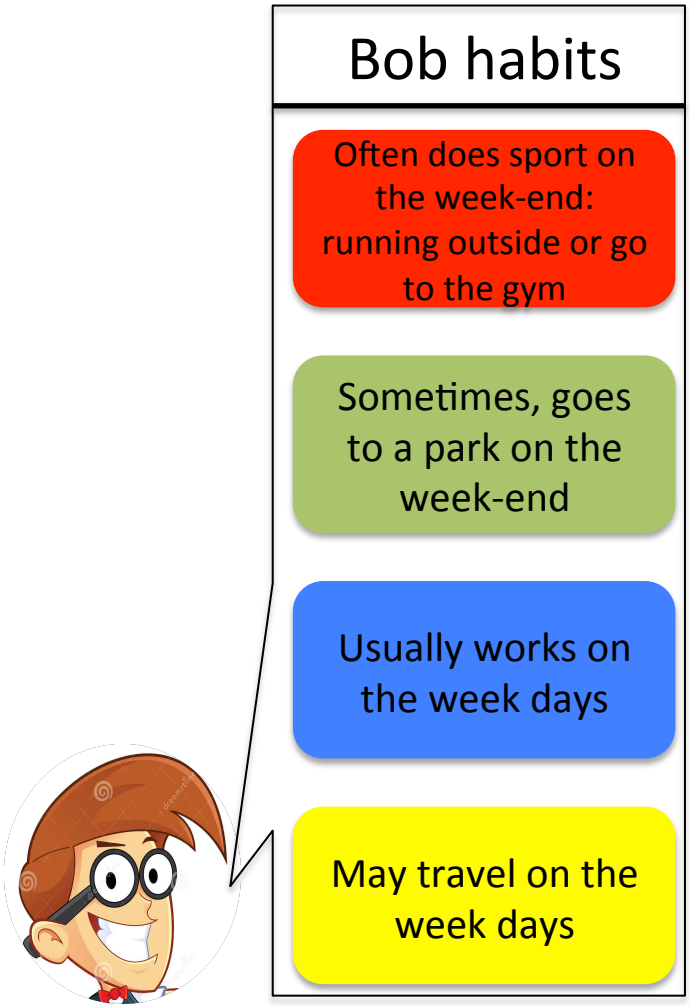
- No prior knowledge used
  - Ex: A user is likely to sleep at night (What if Bob is a watchman)
  - Ex: A user is likely to work during the week days (What if Bob is a farmer)

- **Terms:**
  - behavior = habit
  - Smartphone logs = user logs = dataset

Headlines

1. Latent Multimodal Representation (LMR)
2. Dirichlet Latent Multimodal Representation (DLMR)
3. Evaluation Metrics
4. Other Baseline models
5. Results
6. Conclusion

What is a good representation of a Behavior with smartphone logs?



What is a good representation of a Behavior with smartphone logs?

Often does sport on the week-end: running outside or go to the gym	
Gps_gym	0.6
Gps_other_places	0.4
Activity_running	0.8
Activity_still	0.2
App_nikeplus	0.7
App_noapp	0.3
Notif_nikeplus	0.8
Notif_others	0.2
Day_saturday	0.4
Day_sunday	0.6

- A Behavior is represented by a set of distributions
- Each distribution represents one different feature
- The probabilities of the distribution of feature f indicates the likelihood of each of the realizations of f to appear in the concerned behavior

Models the multimodality of the smartphone data

What is a good representation of a Behavior with smartphone logs?

Sometimes, goes to a park on the week-end	
Gps_park	0.9
Gps_other_places	0.1
Activity_running	0.1
Activity_still	0.9
App_news	0.1
App_noapp	0.9
Notif_email	0.2
Notif_others	0.8
Day_saturday	0.6
Day_sunday	0.4

- A Behavior is represented by a set of distributions
- Each distribution represents one different feature
- The probabilities of the distribution of feature f indicates the likelihood of each of the realizations of f to appear in the concerned behavior

Models the multimodality of the smartphone data

Representing smartphone logs as a corpus

One Record Of Bob	
Gps_gym Gps_park	2 1
Activity_still	1
App_nikeplus App_news	2 1
Notif_nikeplus Notif_others	5 2
Day_saturday	1
Time_morining	1

- Represent smartphone logs as a list of records
- Each record represents the events that happened during 1 hour of the period of observation of the user
- A record contains the number of times each realization were observed during the concerned period of observation

Latent Multimodal Representation (LMR)

One Record Of Bob	
Gps_gym	2
Gps_park	1
Activity_still	1
App_nikeplus	3
App_news	1
Notif_nikeplus	5
Notif_others	2
Day_saturday	1
Time_morining	1

Latent Multimodal Representation (LMR)

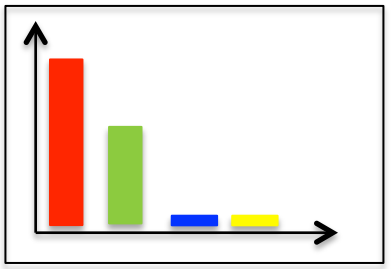
Gps_gym	0.6
Gps_other_places	0.4
⋮	
Day_sunday	0.6

Gps_park	0.9
Gps_other_places	0.1
⋮	
Day_sunday	0.4

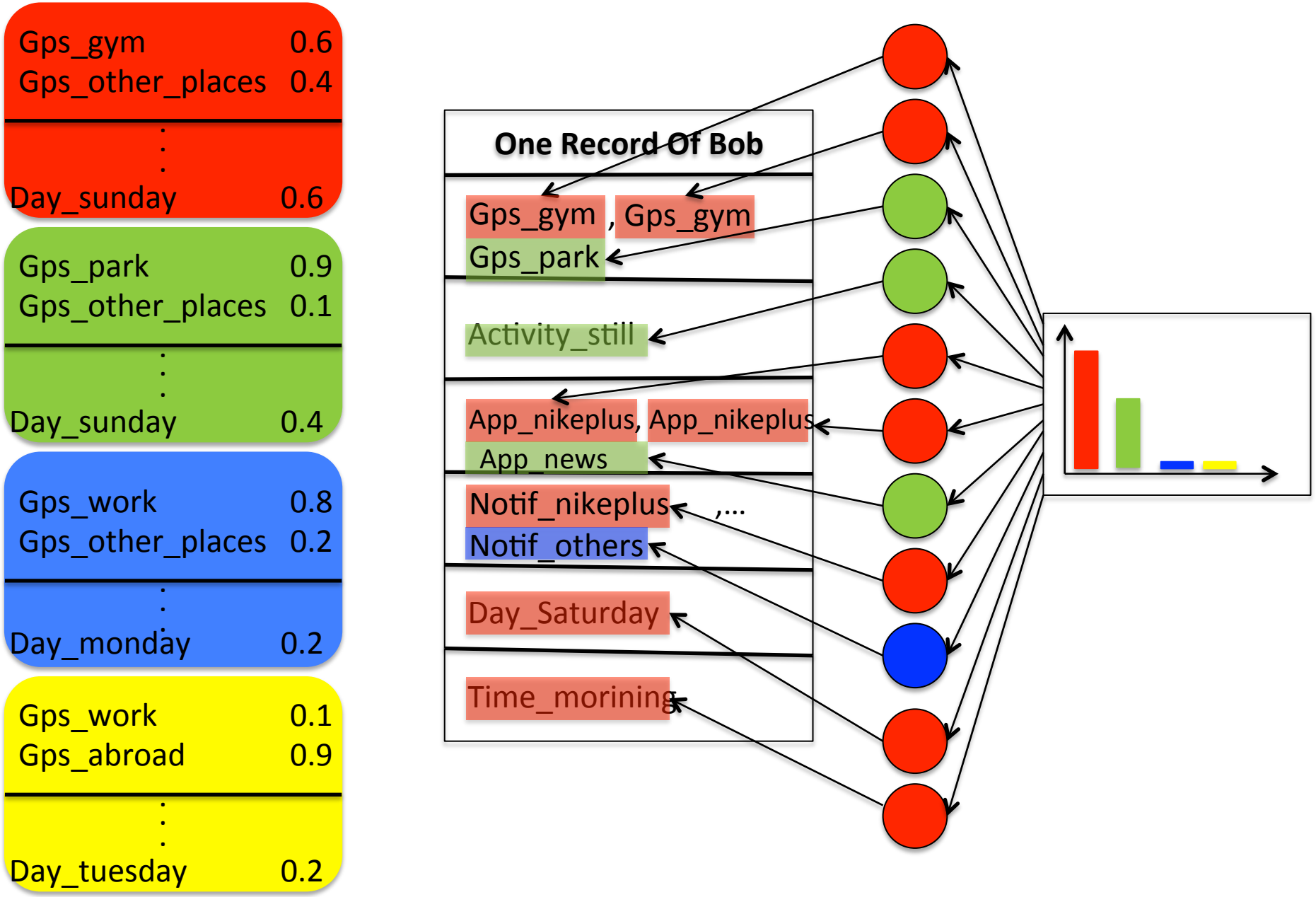
Gps_work	0.8
Gps_other_places	0.2
⋮	
Day_monday	0.2

Gps_work	0.1
Gps_abroad	0.9
⋮	
Day_tuesday	0.2

One Record Of Bob	
Gps_gym	2
Gps_park	1
Activity_still	1
App_nikeplus	3
App_news	1
Notif_nikeplus	5
Notif_others	2
Day_saturday	1
Time_morining	1



Latent Multimodal Representation (LMR)



Latent Multimodal Representation (LMR)

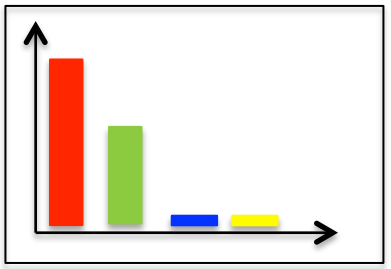
Gps_gym ?
Gps_other_places ?
:
Day_sunday ?

Gps_park ?
Gps_other_places ?
:
Day_sunday ?

Gps_work ?
Gps_other_places ?
:
Day_monday ?

Gps_work ?
Gps_abroad ?
:
Day_Tuesday ?

One Record Of Bob
Gps_gym , Gps_gym Gps_park
Activity_still
App_nikeplus, App_nikeplus App_news
Notif_nikeplus,… Notif_others
Day_Saturday
Time_morining



- The distributions of features for each behavior are unknown

Latent Multimodal Representation (LMR)

Gps_gym ?
Gps_other_places ?
:
Day_sunday ?

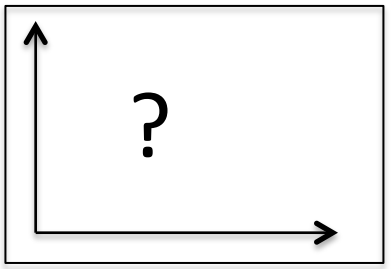
Gps_park ?
Gps_other_places ?
:
Day_sunday ?

Gps_work ?
Gps_other_places ?
:
Day_monday ?

Gps_work ?
Gps_abroad ?
:
Day_Tuesday ?

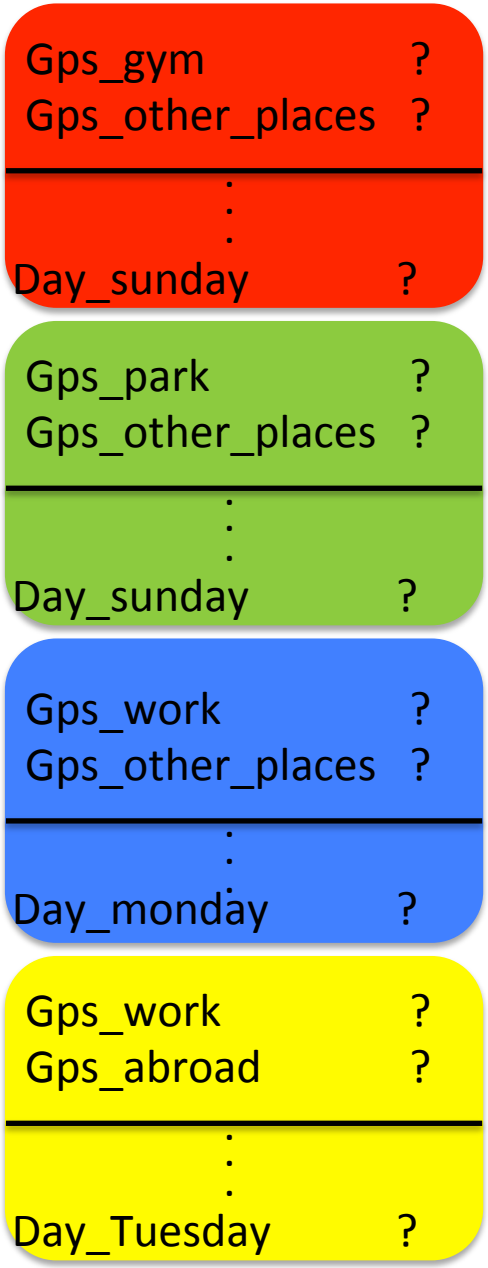
- **We choose the parameters (i.e probabilities) that maximize the probability of the observed data**

- The distribution of behaviors for each record is unknown



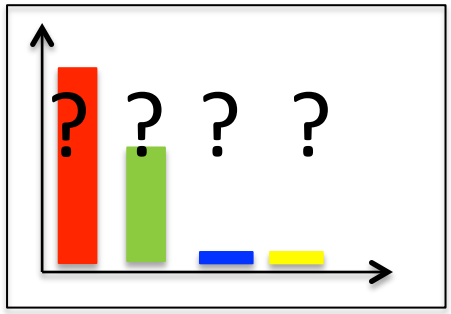
- The distributions of features for each behavior are unknown

LMR: some properties



- For each Behavior, learns one distribution for each feature
- K= the number of behaviors, V = number of possible realizations (over all the features), #parms_to_learn = K.V

LMR: some properties



- For each Behavior, learns one distribution for each feature (K.V params)
- For each record, learns one distribution over the possible behaviors
- M=number of records in the corpus, #parms_to_learn = K.M

LMR: some properties

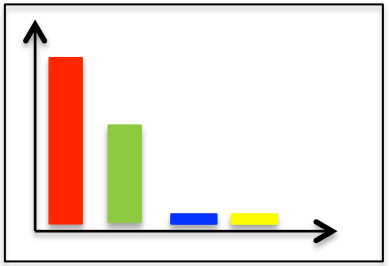
- For each Behavior, learns one distribution for each feature (K.V params)
- For each record, learns one distribution over the possible behaviors (K.M)
- Number of total parameters to learn (K.V +K.M) grows linearly with K and with V

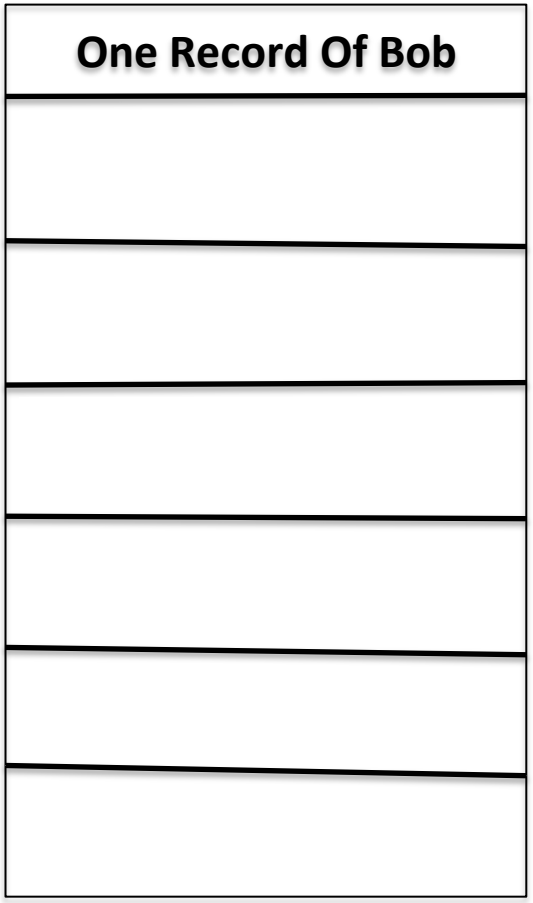
Risk of overfitting when K or V increases

Dirichlet Latent Multimodal Representation (DLMR)

- For each record, Use a Dirichlet distribution of parameter $\vec{\alpha}$ to generate behavior distribution

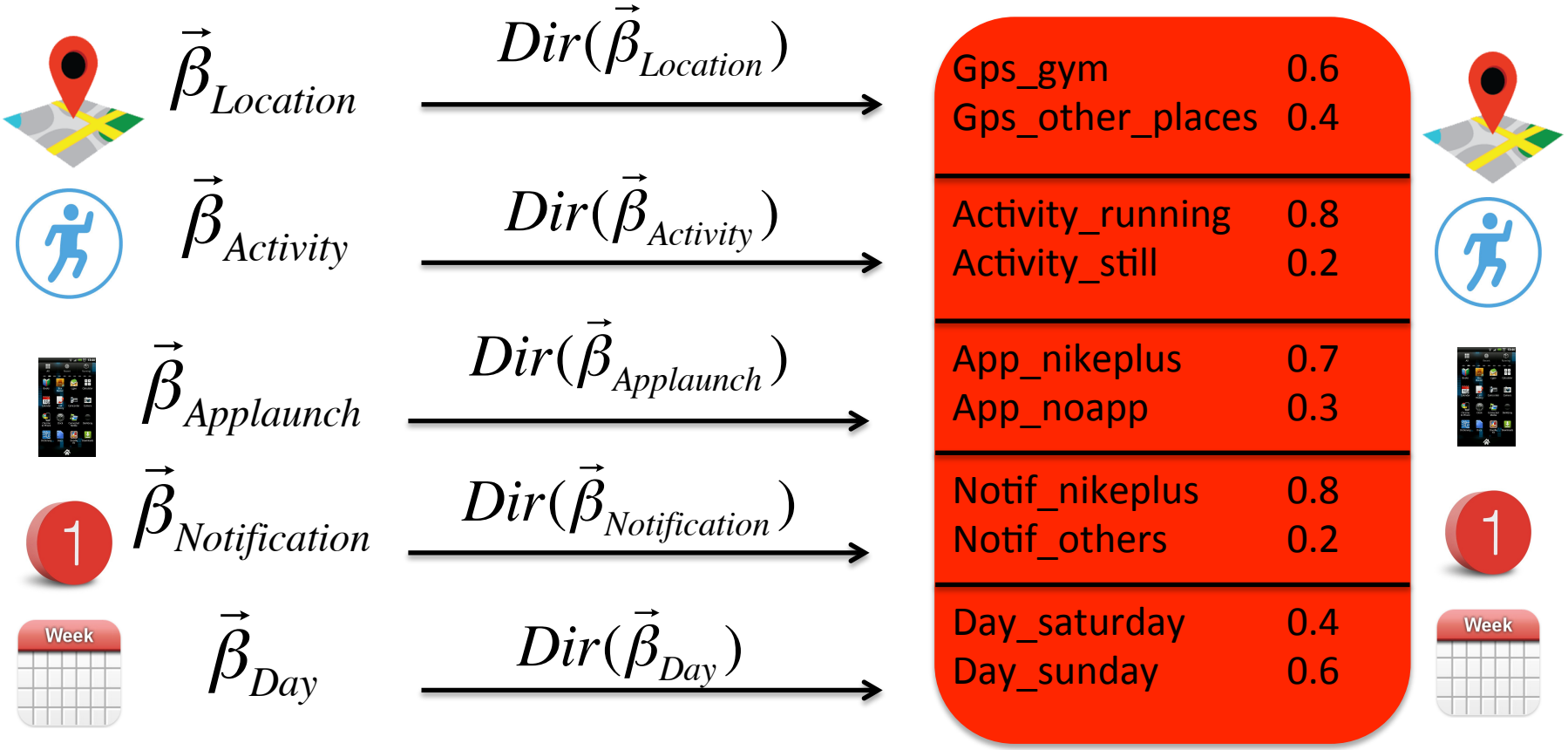
$$\vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} \xrightarrow{Dir(\vec{\alpha})}$$





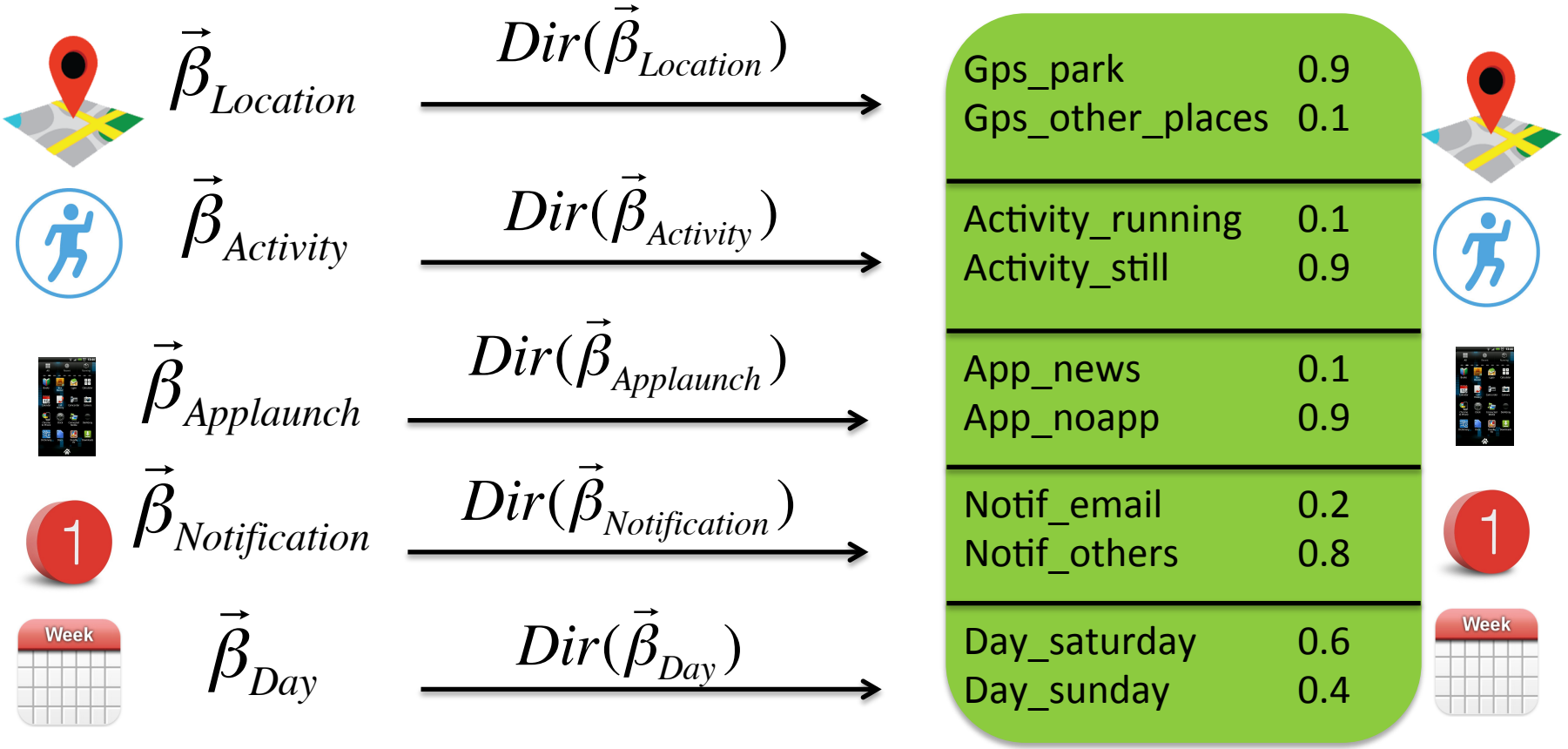
Dirichlet Latent Multimodal Representation (DLMR)

- To generate a behavior do the following : For each feature f, Use a Dirichlet distribution of parameter $\vec{\beta}_f$ to generate the distribution of realizations of feature f



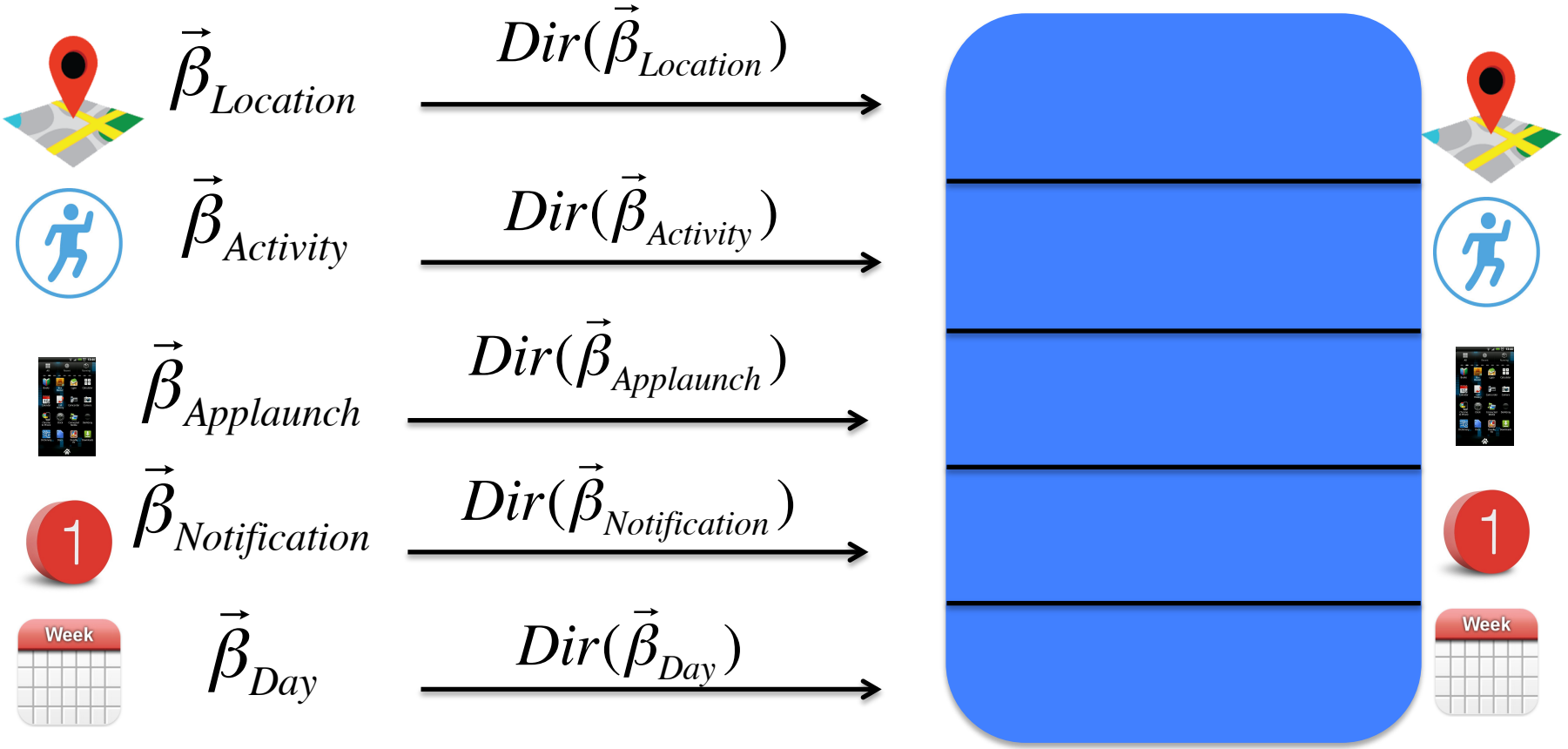
Dirichlet Latent Multimodal Representation (DLMR)

- To generate a behavior do the following : For each feature f, Use a Dirichlet distribution of parameter $\vec{\beta}_f$ to generate the distribution of realizations of feature f



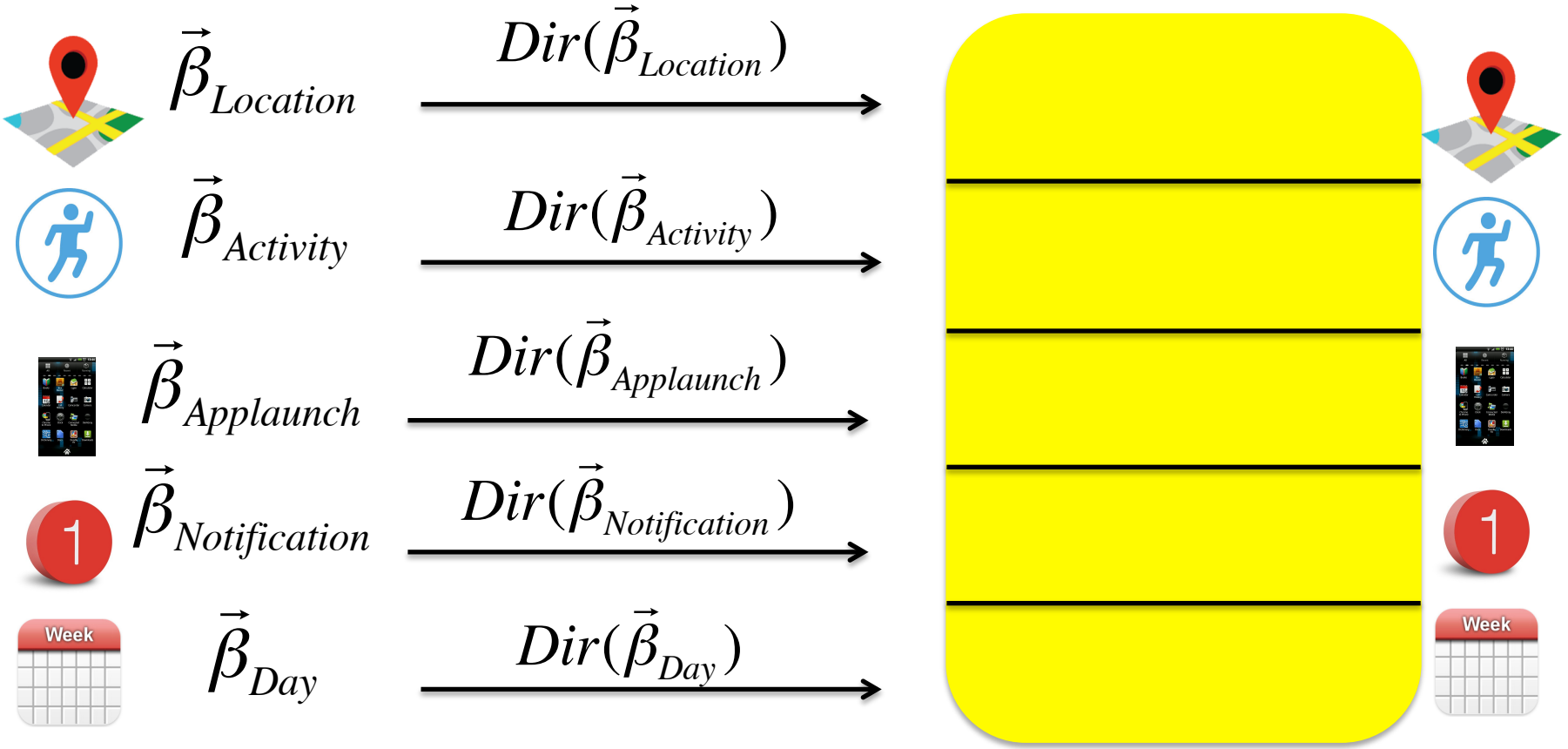
Dirichlet Latent Multimodal Representation (DLMR)

- To generate a behavior do the following : For each feature f, Use a Dirichlet distribution of parameter $\vec{\beta}_f$ to generate the distribution of realizations of feature f



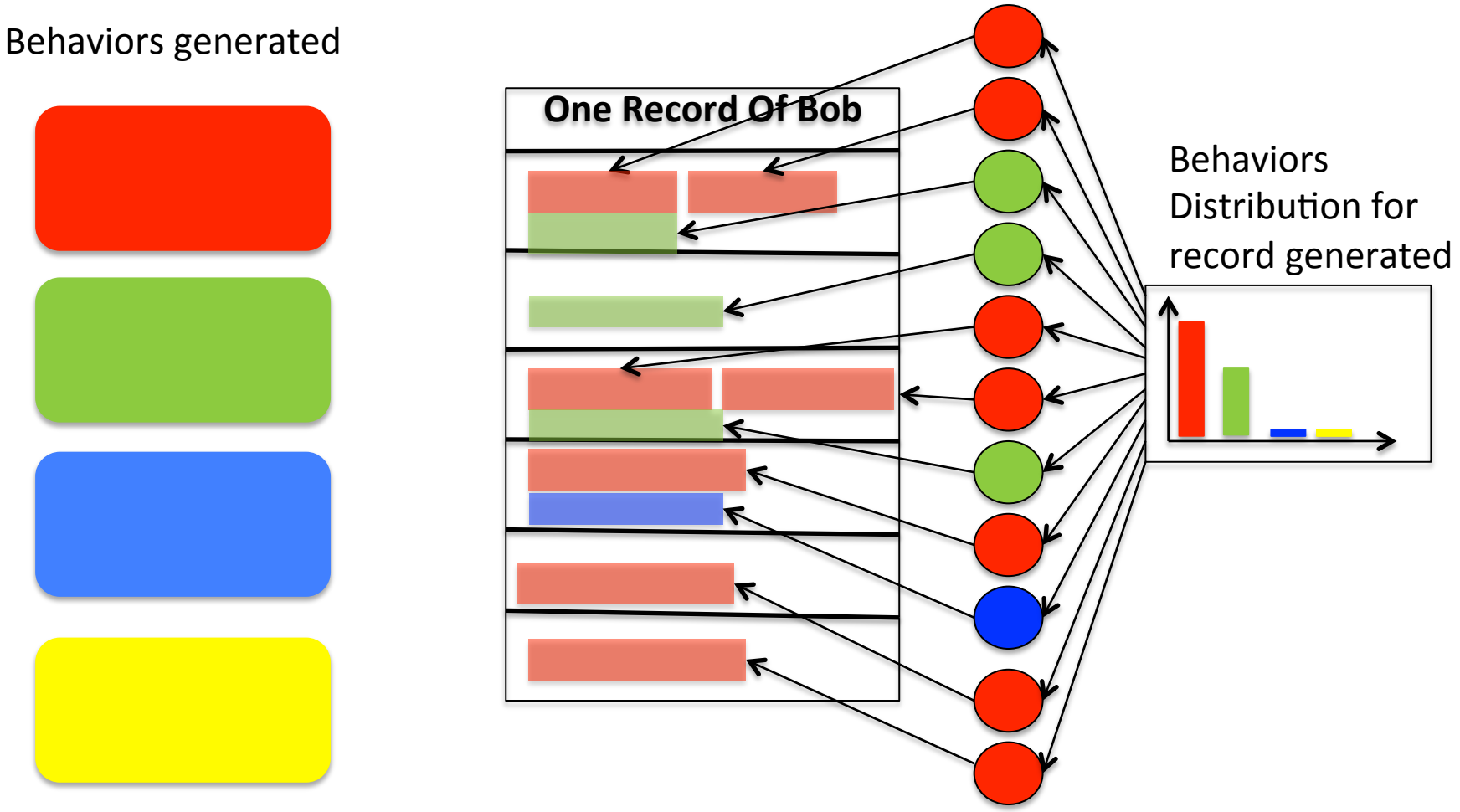
Dirichlet Latent Multimodal Representation (DLMR)

- To generate a behavior do the following : For each feature f, Use a Dirichlet distribution of parameter $\vec{\beta}_f$ to generate the distribution of realizations of feature f



DLMR: record generation

- When behaviors generated and behaviors distribution for records generated, DLMR behaves as LMR



DLMR: some properties

$\vec{\beta}_{Location}$

$\vec{\beta}_{Activity}$

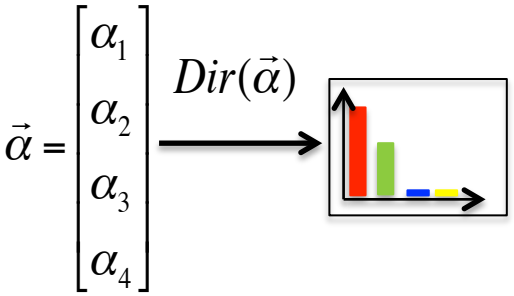
$\vec{\beta}_{Applaunch}$

$\vec{\beta}_{Notification}$

$\vec{\beta}_{Day}$

- Learns how to generate behaviors of Bob by estimating V parameters for $\vec{\beta}$

DLMR: some properties



- Learns how to generate behaviors of Bob by estimating V parameters for $\vec{\beta}$
- Learns how to generate behavior distribution for records by estimating K parameters for $\vec{\alpha}$

DLMR: some properties

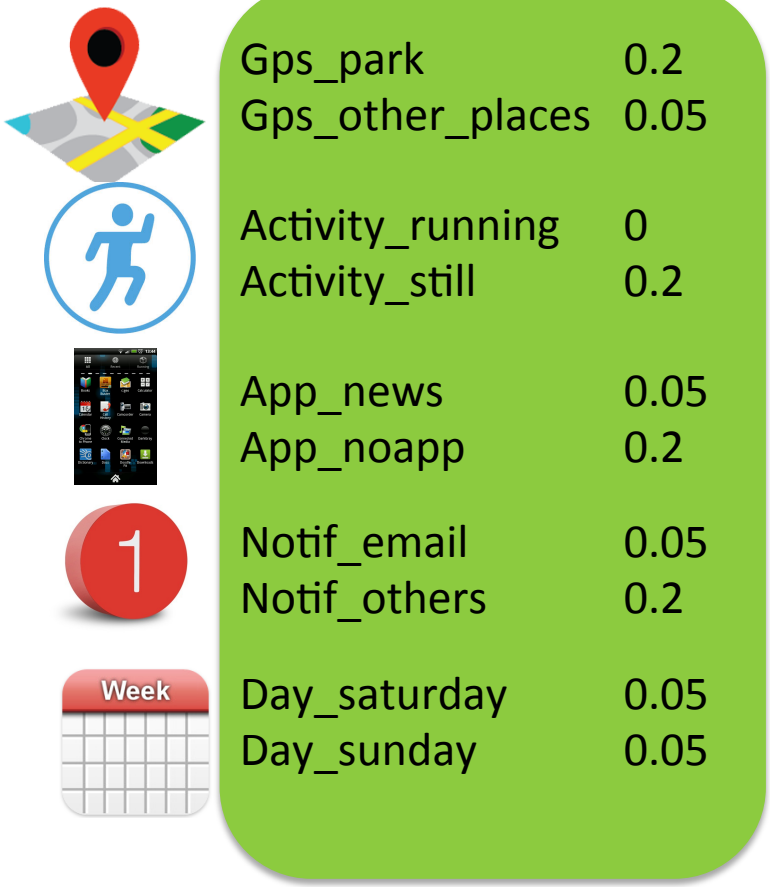
- Learns how to generate behaviors of Bob by estimating V parameters for $\vec{\beta}$
- Learns how to generate behavior distribution for records by estimating K parameters for $\vec{\alpha}$
- Needs to estimate K+V parameters

DLMR vs LMR

LMR	DLMR
K.V+K.M parameters to estimate	K+M parameters to estimate
Fits one behavior distribution to each record	Learns how to generate behavior distributions according to Bob
Learns K behaviors of Bob	Learns how to generate behaviors of Bob
⇨ Tries to find behaviors that describe the observed part of Bob life	⇨ Tries to understand how Bob behaves from the observed part of his life

Latent Dirichlet Allocation (LDA)

- LDA is a widely used model based on topic modeling
  - Corpus of Text representation
  - Recommender Systems
  - Image processing
- DLMR model is very similar to LDA
- However, LDA represents behavior as a unique distribution over all the possible realizations



DLMR vs LDA

LDA	DLMR
Model the behavior by combining the elements of different types in a same distribution Behaviors not intuitive to interpret	Model the behavior by representing the different types in separate distributions, then combine distributions to describe the behavior using the different types Behaviors very intuitive to interpret
⇨ Is not a good representation of multimodal data	⇨ Is a realistic representation of multimodal data

Evaluation Metrics(1): Idea

- **Goal:** Verify that behaviors extracted represent real user habits
  - Find a objective metric to evaluate the quality of the results
  - Find a objective metric to compare the performance of models

- **Idea:** If extracted behaviors able to describe future data ⇨ behaviors representative of user habits
  1. Train the model on train data
  2. Hide some information from the test records
  3. See How well the model is able to guess correctly the missing information from the truncated context described by the test record

Evaluation Metrics(2): Location Prediction

- Classify location into classes: "1st_frequent_loc", "2nd_frequent_loc", "others"
  - Hide location from test records
  - Use the model to guess a location
  - See if the location guessed belongs to the right class

- Metrics:

$$Accuracy = \frac{\#\,good\_guesses}{\#\,total\_records}$$

$$Avergae\_Accuracy = \frac{1}{\#\,classes} \sum_{c \in classes} \frac{\#\,good\_guesses\_of\_c}{\#\,total\_records\_of\_c}$$

Evaluation Metrics(3): Dataset

- Logs of 5 users observed during several months

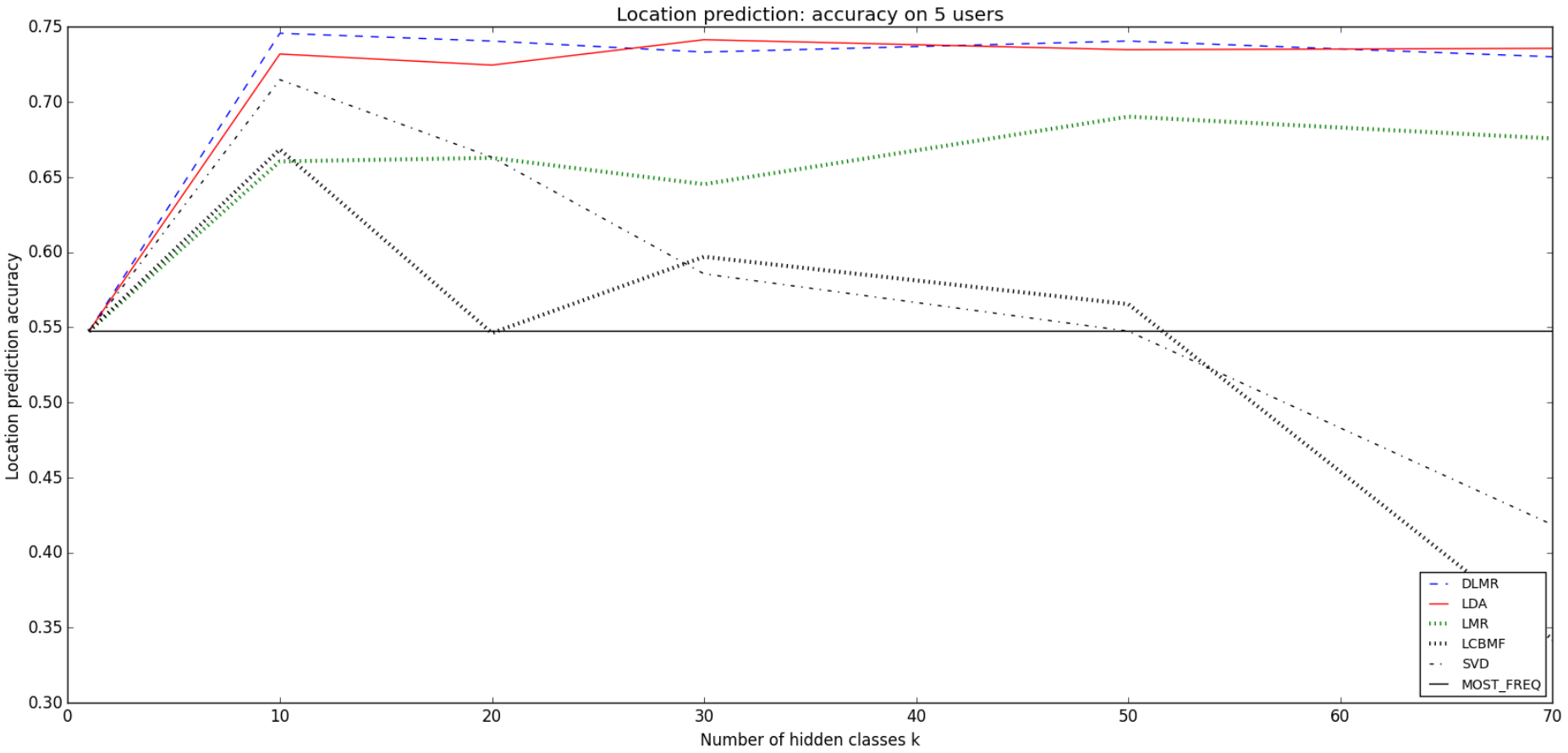
	User 1	User 2	User 3	User 4	User 5
Days of observation	300	231	229	249	224

- Features

Discovery of User behaviors models

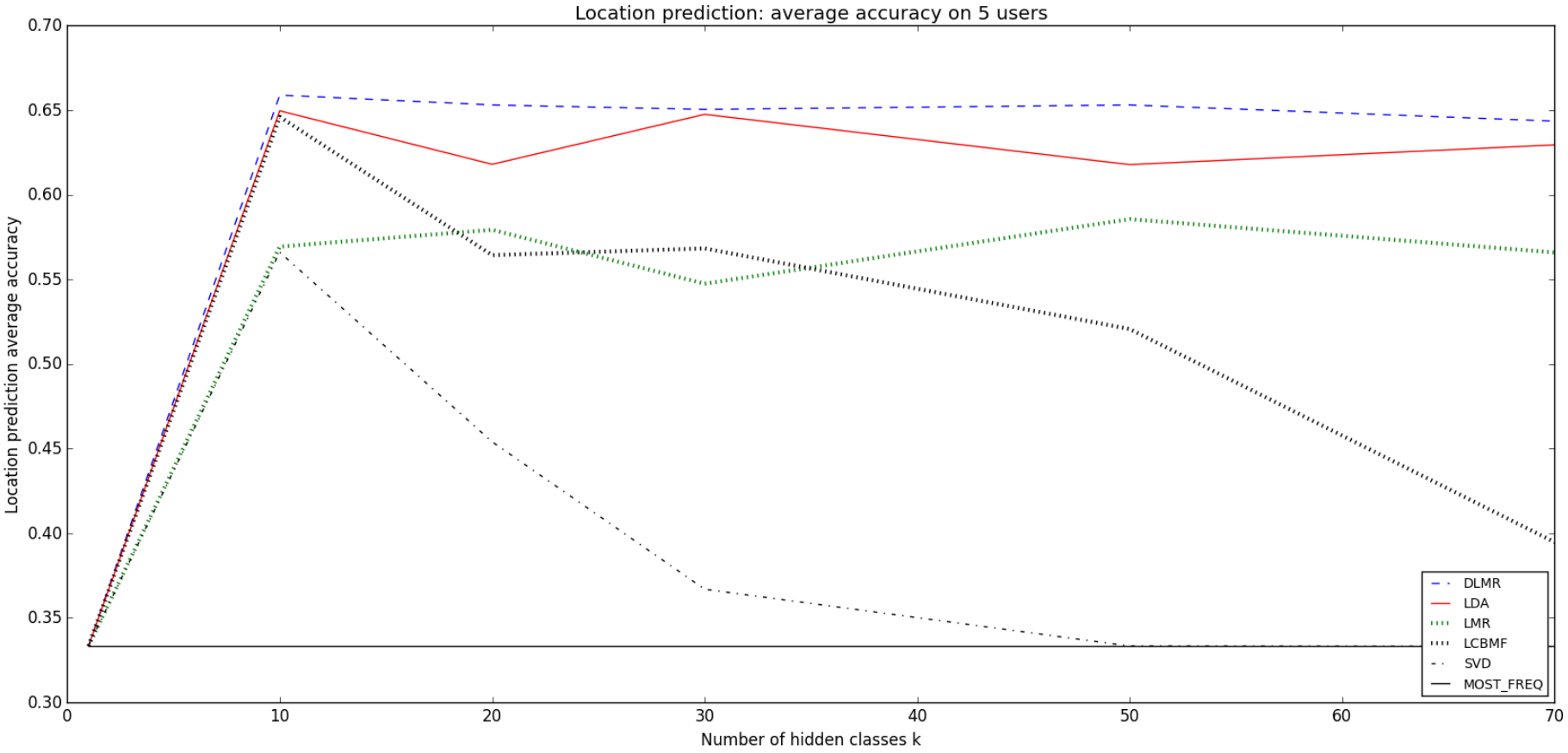
- Already presented LDA, LMR, DLMR (all probabilistic models)
- Use of some matrix factorization models
- Singular Value Decompostion (SVD)
- Linearly Constrained Bayesian Matrix Factorization (LCBMF)
  - State of the art matrix factorization model
  - Used for blind source separation

Results(1): Location prediction accuracy



- DLMR (and LDA) stable for big K
- DLMR and LDA largely outperforms all the other models
- DLMR (and LDA) makes a good prediction accuracy around 73%

Results(2): Location prediction average accuracy



- DLMR (and LDA) stable for big K
- DLMR better than LDA in average accuracy, i.e 2% better
- DLMR outperforms all the models in average accuracy

Results(3): Prediction results conclusions

1. $Accuracy_{DLMR} = Accuracy_{LDA}$, $Average\_Accuracy_{DLMR} > Average\_Accuracy_{LDA}$

⇒ DLMR is able to better catch the rare contexts than LDA

⇒ Sign for better generalization performances on unseen data

Why? : **By representing the multimodality of logs, DLMR assumes more realistic model of user habits than LDA**

2. DLMR largely outperforms LMR

Why? : **DLMR learns who Bob behaves, LMR finds behaviors that describe the observed sample of data**

3. DLMR has good prediction performances in Both considered scores

⇒ **Extracted behaviors describe well user's life** (able to predict future events)

Results(4): Examples of extracted behaviors

Feature : **Day**	
1. Wednesday	0.35
2. Thursday	0.20
3. Tuesday	0.15
4. Friday	0.13
5. Monday	0.14
6. Saturday	0.019
7. Sunday	0.001

Feature : **Hour**	
1. 9am-4pm	0.51
2. 4pm-0am	0.49
3. 0am-9am	0.01

Feature : **Location**	
1. Work	0.99
2. Others	≈ 0
3. Home	≈ 0
4. 3rd_frequent	≈ 0
5. 4nd_frequent	≈ 0

Feature : **Application Launch**	
1. Not_present	0.97
2. Tv Side View	≈ 0
3. Google Camera	≈ 0
4. Android Wear	≈ 0
5. Gmail	≈ 0

Feature : **Activity**	
1. Still	0.46
2. In Vehicle	0.21
3. Tilting	0.17
4. On Foot	≈ 0
5. In Bicycle	≈ 0

Feature : **Notification**	
1. Not_present	0.94
2. Gunosy	≈ 0
3. Gmail	≈ 0
4. News Pick	≈ 0
5. Nike +	≈ 0

- User2: works during the weed days

Feature : **Day**	
36. Sunday	0.40
37. Saturday	0.21
38. Monday	0.12
39. Friday	0.10
40. Wednesday	0.09
41. Thursday	0.04
42. Tuesday	0.04

Feature : **Hour**	
16. 9am-4pm	0.65
17. 4pm-0am	0.34
18. 0am-9am	0.01

Feature : **Location**	
26. Others	0.81
27. Home	0.12
28. 5th_frequent	0.02
29. 10th_frequent	0.01
30. 8rd_frequent	0.01

Feature : **Application Launch**	
26. Ingress	0.52
27. Mozilla Firefox	0.17
28. Gmail	0.14
29. Not_Present	0.04
30. Google Maps	0.03

Feature : **Activity**	
31. Tilting	0.68
32. Still	0.22
33. In Vehicle	0.07
34. On Foot	0.03
35. In Bicycle	≈ 0

Feature : **Notification**	
31. Not_present	0.43
32. Google_Agenda	0.14
33. Gmail	0.12
34. Inbox_Gmail	0.07
35. Ingress	0.07

- User2: plays an augmented reality game on the week end

Conclusion

- DLMR able to catch both general (i.e recurrent) behaviors and specific ones
- The Behaviors extracted are very intuitive to understand and interpret

- As LDA for unimodal data, DLMR can be applied to any multimodal data to accomplish many different tasks:
  - Profile Facebook users by integrating their messages, their likes, their friends, their interests, their events,…
  - Analyze a sensor equipped car dataset by integrating its different sensors at the same time
  - Compress a big multimodal dataset

Many Thanks!

Questions