Online Bayesian Transfer Learning for Sequential Data Modeling

Priyank Jaini

Machine Learning, Algorithms and Theory Lab

UNIVERSITY OF WATERLOO

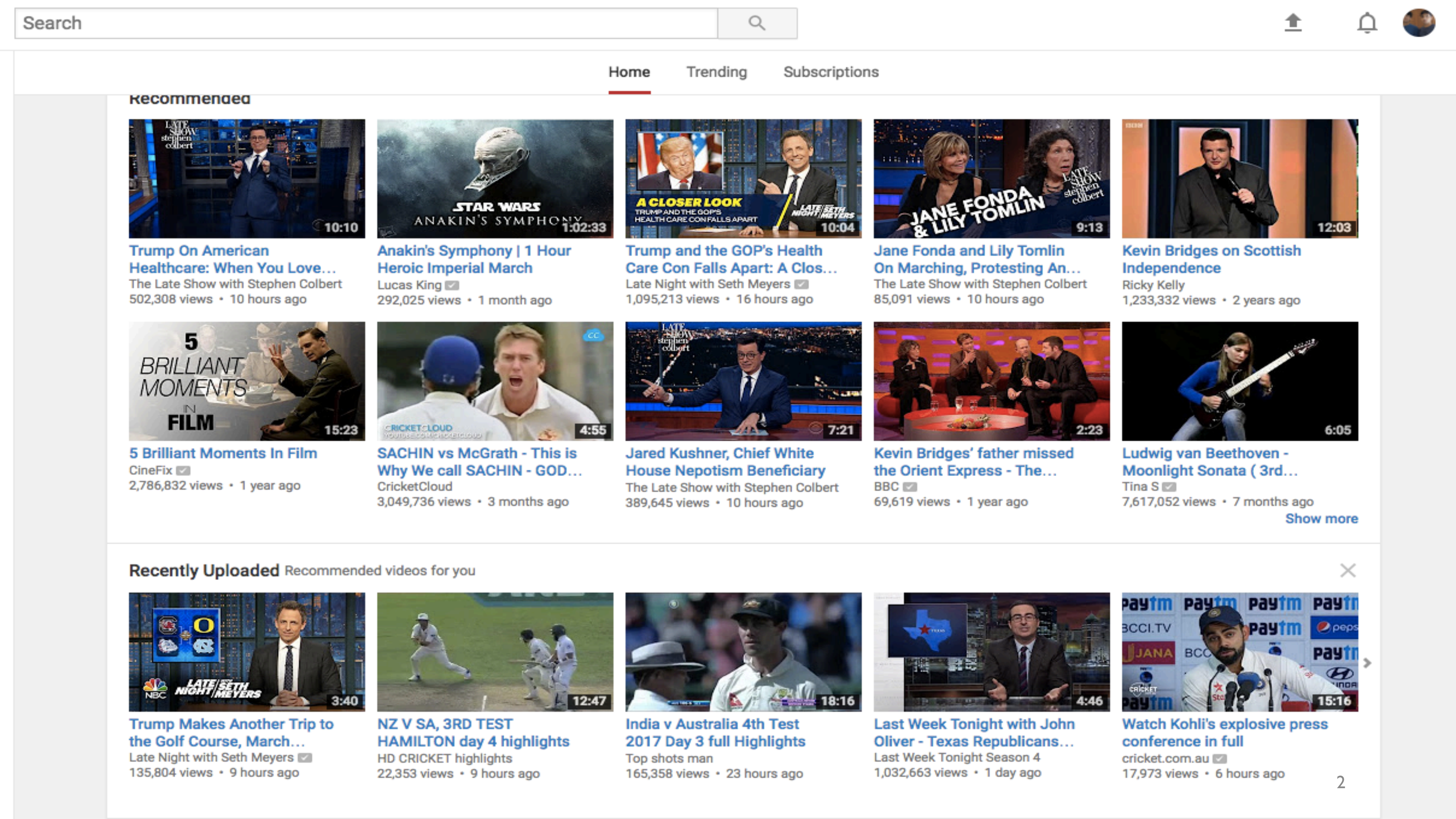
HUAWEI

University of Waterloo's CHRONIC DISEASE PREVENTION INITIATIVE at the PROPEL CENTRE FOR POPULATION HEALTH IMPACT™

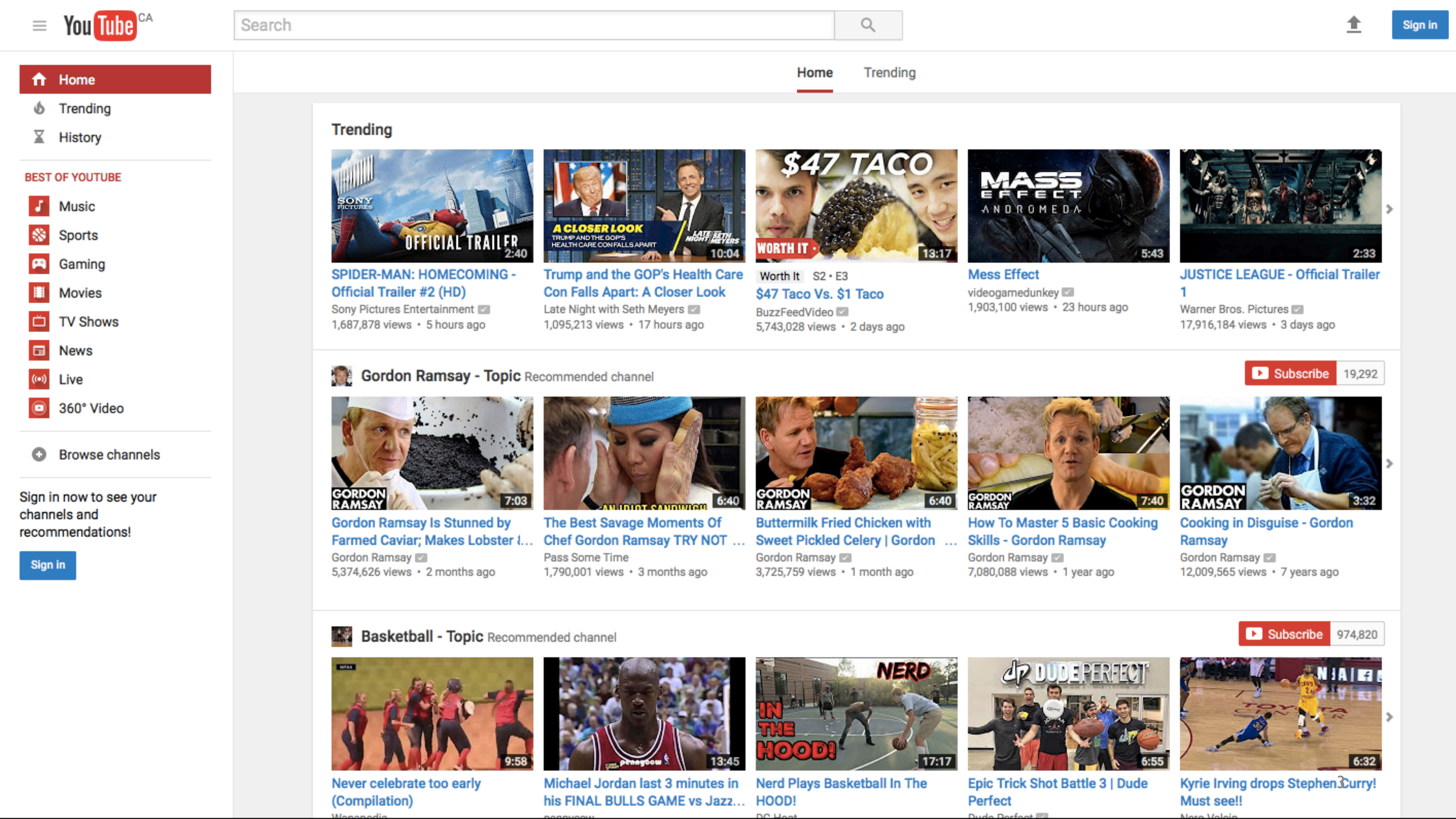
CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

NSERC CRSNG

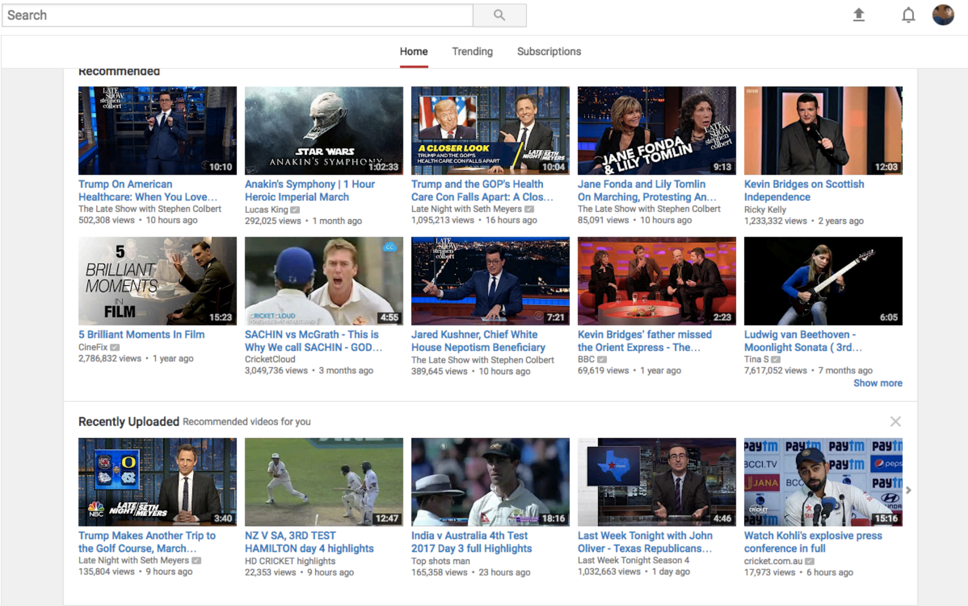
Network for Aging Research

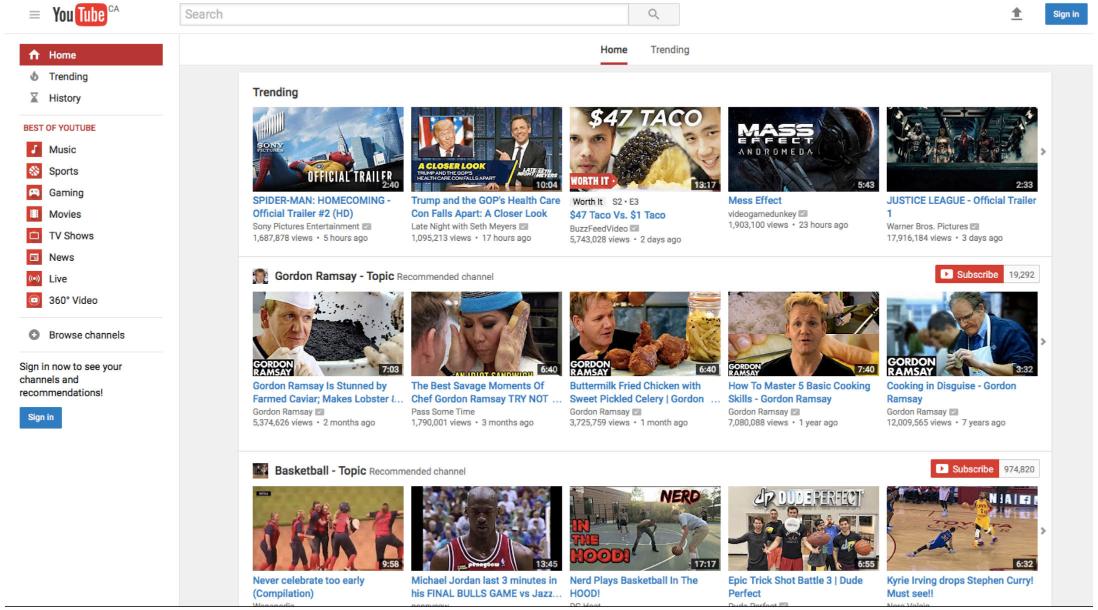


Search
Home
Trending
Subscriptions
Recommended
Trump On American Healthcare: When You Love…
The Late Show with Stephen Colbert
502,308 views • 10 hours ago
Anakin's Symphony | 1 Hour Heroic Imperial March
Lucas King
292,025 views • 1 month ago
Trump and the GOP's Health Care Con Falls Apart: A Clos…
Late Night with Seth Meyers
1,095,213 views • 16 hours ago
Jane Fonda and Lily Tomlin On Marching, Protesting An…
The Late Show with Stephen Colbert
85,091 views • 10 hours ago
Kevin Bridges on Scottish Independence
Ricky Kelly
1,233,332 views • 2 years ago
5 Brilliant Moments In Film
CineFix
2,786,832 views • 1 year ago
SACHIN vs McGrath - This is Why We call SACHIN - GOD…
CricketCloud
3,049,736 views • 3 months ago
Jared Kushner, Chief White House Nepotism Beneficiary
The Late Show with Stephen Colbert
389,645 views • 10 hours ago
Kevin Bridges' father missed the Orient Express - The…
BBC
69,619 views • 1 year ago
Ludwig van Beethoven - Moonlight Sonata (3rd…
Tina S
7,617,052 views • 7 months ago
Show more
Recently Uploaded Recommended videos for you
Trump Makes Another Trip to the Golf Course, March…
Late Night with Seth Meyers
135,804 views • 9 hours ago
NZ V SA, 3RD TEST HAMILTON day 4 highlights
HD CRICKET highlights
22,353 views • 9 hours ago
India v Australia 4th Test 2017 Day 3 full Highlights
Top shots man
165,358 views • 23 hours ago
Last Week Tonight with John Oliver - Texas Republicans…
Last Week Tonight Season 4
1,032,663 views • 1 day ago
Watch Kohli's explosive press conference in full
cricket.com.au
17,973 views • 6 hours ago



YouTube CA
Search
Sign in
Home
Trending
History
BEST OF YOUTUBE
Music
Sports
Gaming
Movies
TV Shows
News
Live
360° Video
Browse channels
Sign in now to see your channels and recommendations!
Sign in
Home
Trending
Trending
OFFICIAL TRAILER
2:40
SPIDER-MAN: HOMECOMING - Official Trailer #2 (HD)
Sony Pictures Entertainment
1,687,878 views • 5 hours ago
A CLOSER LOOK
TRUMP AND THE GOP'S HEALTH CARE CON FALLS APART
10:04
Trump and the GOP's Health Care Con Falls Apart: A Closer Look
Late Night with Seth Meyers
1,095,213 views • 17 hours ago
$47 TACO
WORTH IT
13:17
Worth It S2 • E3
$47 Taco Vs. $1 Taco
BuzzFeedVideo
5,743,028 views • 2 days ago
MASS EFFECT ANDROMEDA
5:43
Mess Effect
videogamedunkey
1,903,100 views • 23 hours ago
2:33
JUSTICE LEAGUE - Official Trailer 1
Warner Bros. Pictures
17,916,184 views • 3 days ago
Gordon Ramsay - Topic Recommended channel
Subscribe 19,292
GORDON RAMSAY
7:03
Gordon Ramsay Is Stunned by Farmed Caviar; Makes Lobster &...
Gordon Ramsay
5,374,626 views • 2 months ago
6:40
The Best Savage Moments Of Chef Gordon Ramsay TRY NOT ...
Pass Some Time
1,790,001 views • 3 months ago
GORDON RAMSAY
6:40
Buttermilk Fried Chicken with Sweet Pickled Celery | Gordon ...
Gordon Ramsay
3,725,759 views • 1 month ago
GORDON RAMSAY
7:40
How To Master 5 Basic Cooking Skills - Gordon Ramsay
Gordon Ramsay
7,080,088 views • 1 year ago
GORDON RAMSAY
3:32
Cooking in Disguise - Gordon Ramsay
Gordon Ramsay
12,009,565 views • 7 years ago
Basketball - Topic Recommended channel
Subscribe 974,820
9:58
Never celebrate too early (Compilation)
13:45
Michael Jordan last 3 minutes in his FINAL BULLS GAME vs Jazz...
NERD
IN THE HOOD!
17:17
Nerd Plays Basketball In The HOOD!
DUDE PERFECT
6:55
Epic Trick Shot Battle 3 | Dude Perfect
6:32
Kyrie Irving drops Stephen Curry! Must see!!



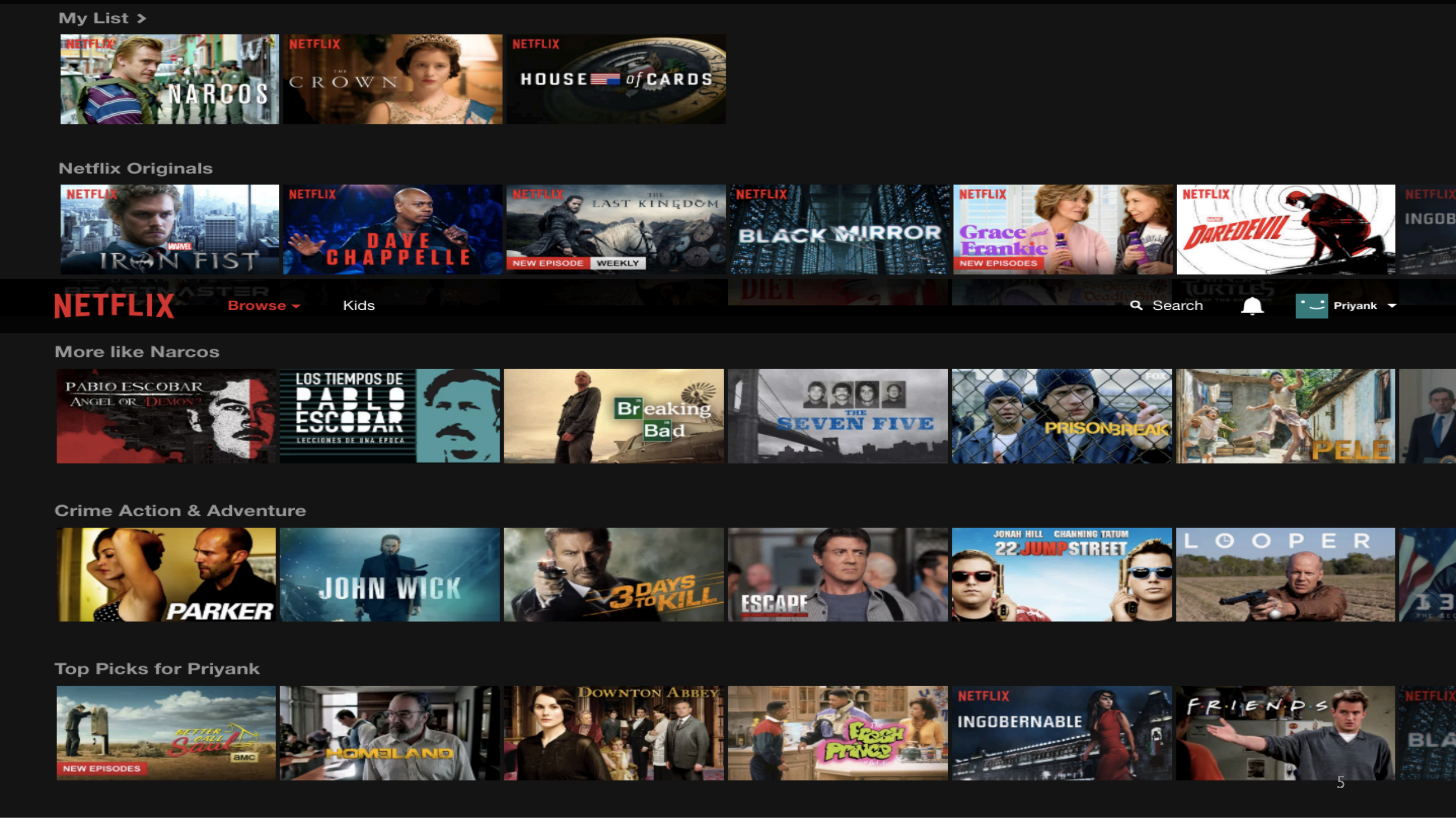


- Data of personal preferences (years)
- Well established model (enough data)

- Data (non-existent)
- Model ?

How do we predict preferences with limited data ?

Population of individuals

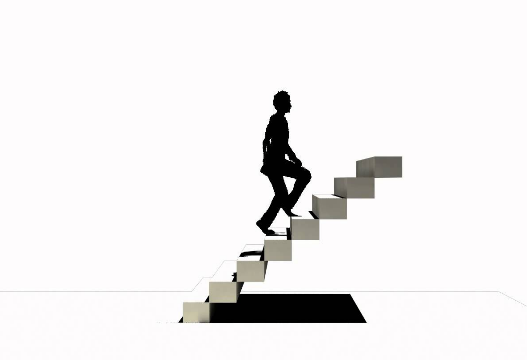

My List >
NETFLIX
NARCOS
THE CROWN
HOUSE of CARDS
Netflix Originals
IRON FIST
DAVE CHAPPELLE
THE LAST KINGDOM
NEW EPISODE
WEEKLY
BLACK MIRROR
Grace and Frankie
NEW EPISODES
DAREDEVIL
NETFLIX
Browse
Kids
Search
Priyank
More like Narcos
PABLO ESCOBAR ANGEL OR DEMON?
LOS TIEMPOS DE PABLO ESCOBAR
LECCIONES DE UNA ÉPOCA
Breaking Bad
THE SEVEN FIVE
PRISON BREAK
PELÉ
Crime Action & Adventure
PARKER
JOHN WICK
3 DAYS TO KILL
ESCAPE
JONAH HILL CHANNING TATUM
22 JUMP STREET
LOOPER
Top Picks for Priyank
BETTER CALL Saul
NEW EPISODES
amc
HOMELAND
DOWNTON ABBEY
Fresh Prince
INGOBERNABLE
FRIENDS

Activity Recognition


Sensors





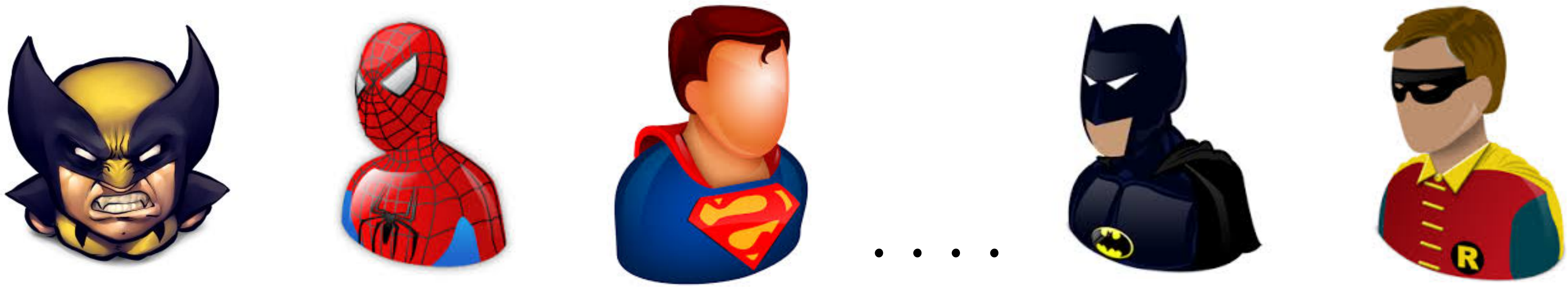



Activity Recognition

- Update model after each observation
- Real-time feedback/analysis
- Population of individuals can be used
- Inter-population variability

Different individuals may have different gait patterns

Inter-Population Variability



. . . .

Model M

Model based on similar individuals

Predict

Inaccurate

Model learned from population does not capture the features of an individual

Idea - Transfer Learning

M_1 M_2 M_3 $M_{(K-1)}$ M_K




. . . .

w_1 w_2 w_3 $w_{(K-1)}$ w_K

$$M = \sum w_i M_i$$

- For each new observation the weights **w** are updated
- Predictions are made using these updated weights

Contributions

Online Bayesian Transfer Learning Algorithm

Step 1 : Source Domain
Online learning HMM models for source individuals

Learning Gaussian Mixture emission distribution using Bayesian Moment Matching

Step 2 : Target Domain
Online learning & prediction for target individual

Updating model weights using Bayesian Moment Matching
&
Classification using MAP

Activity Recognition

Sleep Stage Classification

Network Flow Prediction

Comparison to BMM, oEM and RNN

How can we learn mixture models robustly from streaming data ?

Learning Algorithms

- **Robust** : Tensor Decomposition(Tao, Li et al, 2005),Spectral Learning(Kamvar et al, 2003); offline

- **Online** :
  - Assumed Density Filtering (Maybeck 1982; Lauritzen 1992; Opper & Winther 1999); not robust
  - Expectation Propagation (Minka 2001); does not converge
  - Stochastic Gradient Descent (Zhang 2004)
  - online Expectation Maximization (Cappe 2012)

  SGD and oEM : local optimum and cannot be distributed

Learning Algorithms

- **Exact Bayesian Learning** : Dirichlet Mixtures(Ghosal et al 1999), Gaussian Mixtures(Lijoi et al, 2005), Non-parametric Problems (Barron et al, 1999), (Freedman, 1999)

Exact Bayesian Learning

Online

Distributed

Consistent

In theory; practical problems!

Bayesian Learning – Mixture models

Data : $x_{1:n}$ where $x_i \sim \sum_{j=1}^{M} w_j \, N(x_i; \mu_j, \Sigma_j)$

$$P_n(\Theta) = \Pr(\Theta | x^{1:n})$$
$$\propto P_{n-1}(\Theta) \Pr(x_n | \Theta)$$
$$\propto P_{n-1}(\Theta) \Pr(x_n | \Theta)$$
$$\propto P_{n-1}(\Theta) \sum_{j=1}^{M} w_j \, N(x_i; \mu_j, \Sigma_j)$$

Intractable!!!

Solution : Bayesian Moment Matching Algorithm

Bayesian Learning

Thomas Bayes
(c. 1700-1761)

- Uses Bayes' Theorem

$$P(\Theta|x) = \frac{P(\Theta)P(x|\Theta)}{P(x)}$$

Prior Belief → New Information → Bayes' Theorem → Updated Belief

$P(\Theta)$ $\quad x \quad$ $\frac{P(\Theta)P(x|\Theta)}{P(x)}$ $\quad P(\Theta|x)$

Method of Moments

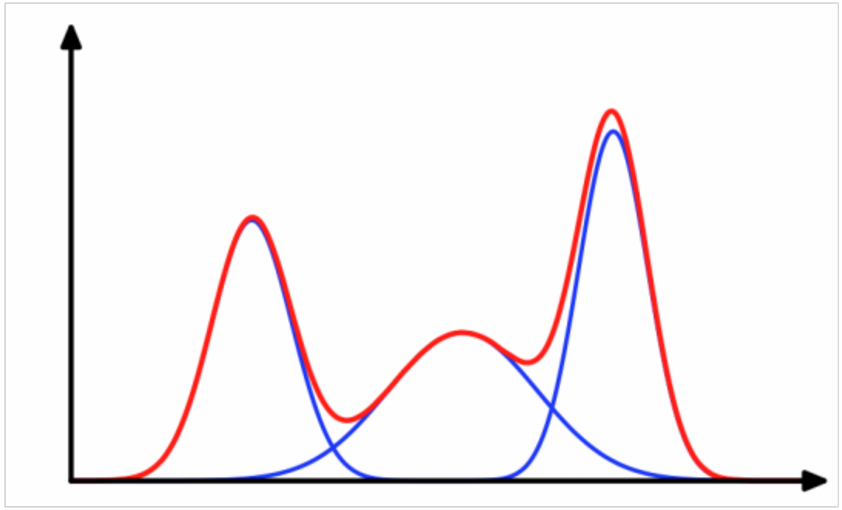
Karl Pearson
(c. 1837-1936)

- Probability distributions defined by set of parameters
- Parameters can be estimated by a set of moments

$$X \sim N(X; \mu, \sigma^2)$$
$$E[X] = \mu$$
$$E[(X-\mu)^2] = \sigma^2$$

Gaussian Mixture Models

$$x_i \sim \sum_{j=1}^{M} w_j \, N(x_i; \mu_j, \Sigma_j)$$



Parameters : weights, means and precisions (inverse covariance matrices)

Bayesian Moment Matching for Gaussian Mixture Models

$$P(\Theta|x) = \frac{\overbrace{P(\Theta)}^{\text{Prior}}\overbrace{P(x|\Theta)}^{\text{Likelihood}}}{P(x)}$$

Parameters

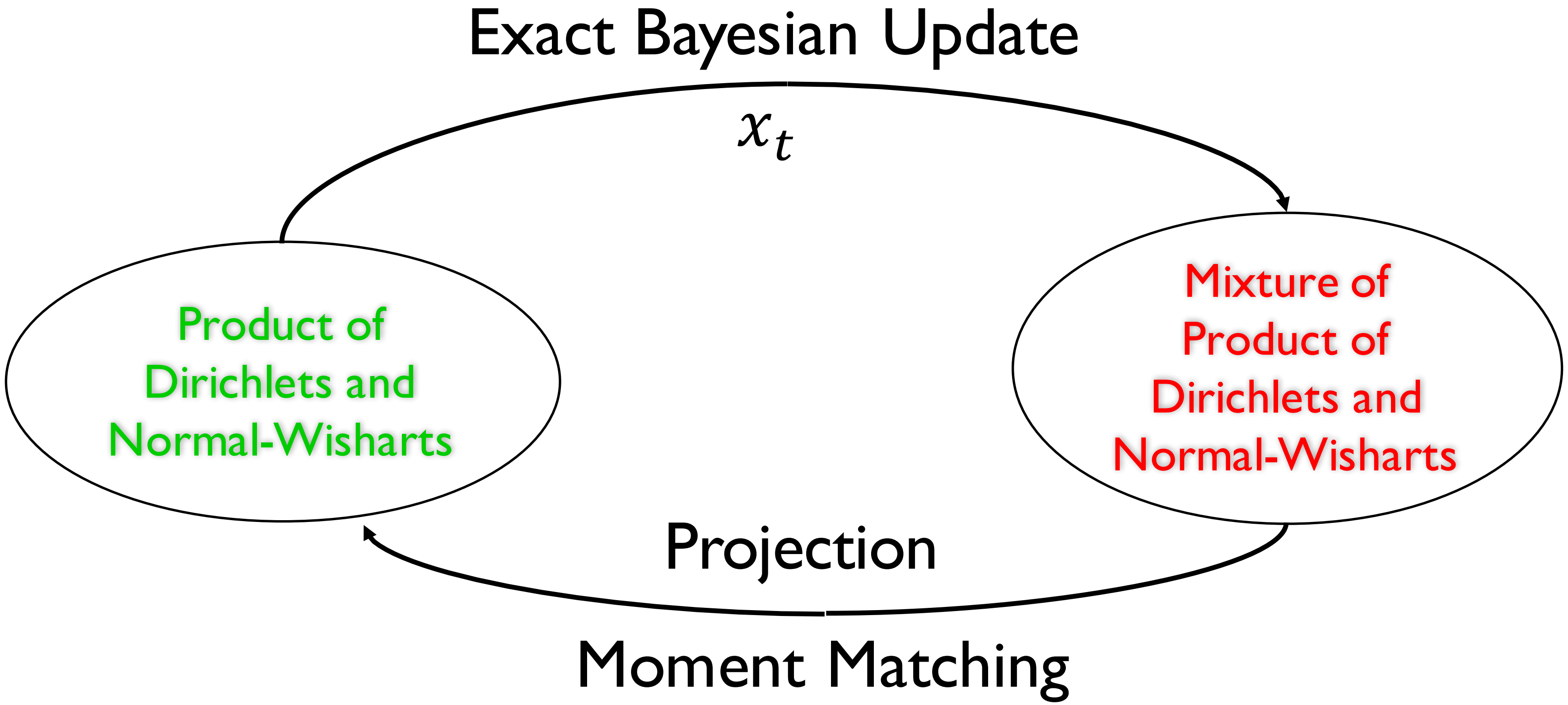
Parameters : weights, means and precisions (inverse covariance matrices)

Prior : $P(\boldsymbol{w}, \mu, \boldsymbol{\Lambda})$; product of Dirichlets and Normal-Wisharts

Likelihood :

$P(\boldsymbol{x}\,; \boldsymbol{w}, \mu, \boldsymbol{\Lambda}) = \sum_{j=1}^{M} w_j\, N(\boldsymbol{x}; \mu_j, \boldsymbol{\Lambda}_j^{-1})$

Bayesian Moment Matching Algorithm



Sufficient Moments

Dirichlet : $Dir(w_1, w_2 \ldots w_M; \alpha_1, \alpha_2 \ldots, \alpha_M)$

$$E[w_i] = \frac{\alpha_i}{\sum_j \alpha_j}; \qquad E\left[w_i^2\right] = \frac{\alpha_i(\alpha_i+1)}{(\sum_j \alpha_j)(1+\sum_j \alpha_j)}$$

Normal-Wishart : $NW(\mu, \Lambda; \mu_0, \kappa, W, v)$

$\Lambda \sim Wi(W, v)$ and $\mu|\Lambda \sim N_d(\mu_0, (\kappa\Lambda)^{-1})$

$$E[\mu] = \mu_0$$

$$E[(\mu - \mu_0)(\mu - \mu_0)^T] = \frac{\kappa+1}{\kappa(v-d-1)} W^{-1}$$

$$E[\Lambda] = vW$$

$$Var(\Lambda_{ij}) = v(W_{ij}^2 + W_{ii}W_{jj})$$

Overall Algorithm

- Bayesian Step
  - Compute posterior $P_t(\Theta)$based on observation x_t
- Sufficient Moments
  - Compute set of sufficient moments S for $P_t(\Theta)$
- Moment Matching
  - System of linear equations
  - Linear complexity in the number of components

Bayesian Moment Matching

- **Discrete Data** : Omar (2015, PhD Thesis) for Dirichlets; Rashwan, Zhao & Poupart (AISTATS'16) for SPNs; Hsu & Poupart (NIPS'16) for Topic Modelling

- **Continuous Data** : Jaini & Poupart, 2016 (*arxiv*); Jaini, Rashwan et al, (PGM'16) for SPNs; Poupart, Chen, Jaini et al (NetworksML'16)

- **Sequence Data and Transfer Learning** : Jaini, Poupart et al, (ICLR'17)

Make Bayesian Learning Great Again

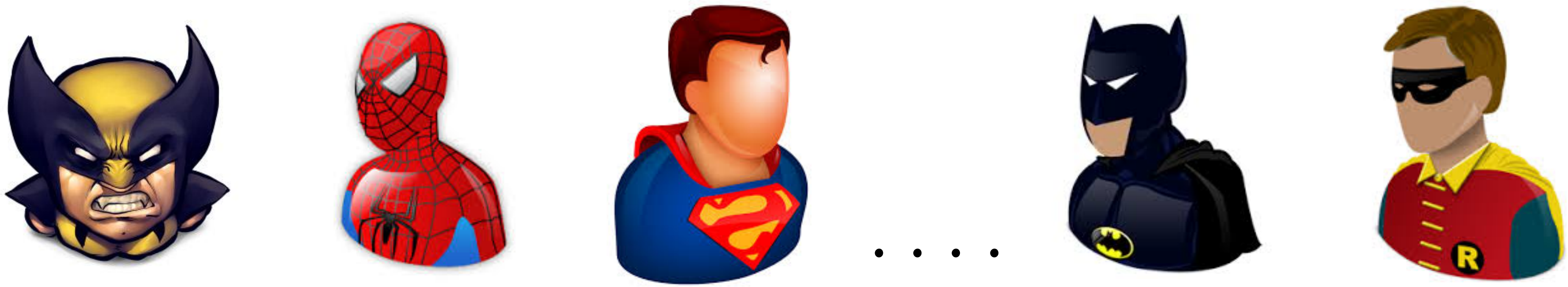
Bayesian Moment Matching Algorithm

- Uses Bayes' Theorem + Method of Moments
- Analytic solutions to Moment matching (unlike EP, ADF)
- One pass over data

Bayesian Moment Matching → Online
Bayesian Moment Matching → Distributed
Bayesian Moment Matching → Consistent ?

Online Bayesian Transfer Learning for Sequential Data Modeling

Problem Formulation



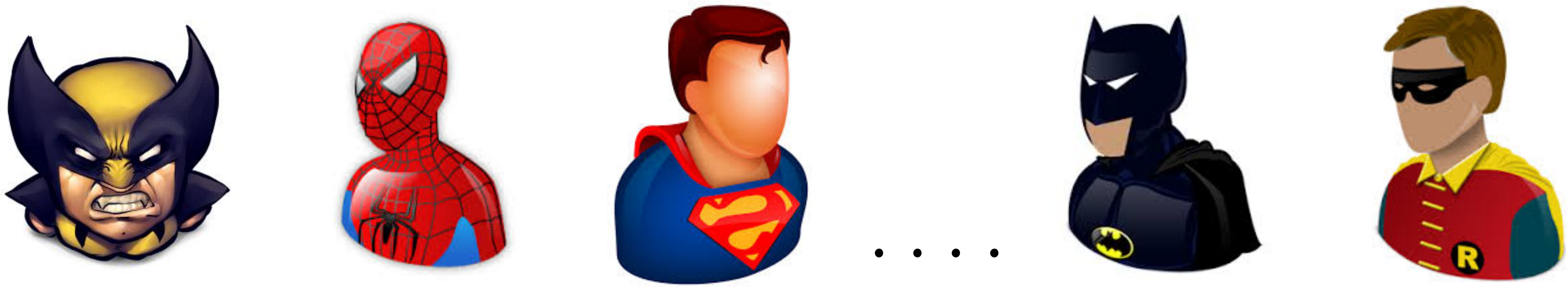
Source Domain

Observed Data

Target Domain

Predict based on observed data

Source Domain - Parameter Learning



M_1 M_2 M_3 . . . . $M_{(K-1)}$ M_K

- Learn an HMM model M_k over each source individual k
- M_k consists of
  - a transition matrix = $\Pr(Y_t = u \mid Y_{t-1} = v) = \varphi_{uv}$
  - an emission distribution = $\Pr(X_t \mid Y_t, \theta)$
- Estimate the transition matrix φ and emission parameters θ

M_k

$Data = (X_1, X_2, \ldots X_t \ldots X_N)$ Observed Variables

$Y_1 \quad Y_t \quad Y_N$ Latent/ Unobserved Variables

$$\Pr(\theta, \varphi, Y_t = j \mid X_{1:t}, Y_{t-1} = i)$$

$$\propto \underbrace{\Pr(X_t | Y_t = j)}_{\substack{\text{Emission} \\ \text{Distribution} \\ \theta}} \underbrace{\Pr(Y_t = j | Y_{t-1} = i)}_{\substack{\text{Transition} \\ \text{Distribution} \\ \varphi}} \underbrace{\Pr(\theta, \varphi, Y_{t-1} = i \mid X_{1:t-1})}_{\text{Prior}}$$

How to choose the Prior?

$$\varphi = \begin{bmatrix} \boxed{\varphi_{11} \quad \cdots \quad \varphi_{1M}} \\ \vdots \quad \ddots \quad \vdots \\ \varphi_{M1} \quad \cdots \quad \varphi_{MM} \end{bmatrix}$$

φ_{ij} = probability to go from state *i* to *j*

- A Dirichlet over each row of φ

- $\Pr(\varphi)$ is a product of Dirichlets; one for each row

$$\Pr(\varphi) = \prod_{m=1}^{M} Dir(\boldsymbol{\varphi}_m \mid \boldsymbol{\alpha}_m)$$

How to choose the Prior?

$$\Pr(X_t | Y_t = j) = \sum_{h=1}^{H} w_{h,j} \, N(X_t; \mu_{h,j}, \Lambda_{h,j}^{-1})$$

- For each *j* we have a product of Dirichlet and Normal-Wishart
- The complete prior is

$$\Pr(\theta) = \prod_{l=1}^{L} \boxed{Dir(w_l \,; \beta_l) \prod_{h=1}^{H} NW\left(\mu_{h,l}, \Lambda_{h,l} \,;\, m_{h,l}, \kappa_{h,l}, W_{h,l}, \nu_{h,l}\right)}$$

$$\Pr(\theta,\varphi,Y_t=j\mid X_t,Y_{t-1}=i)$$

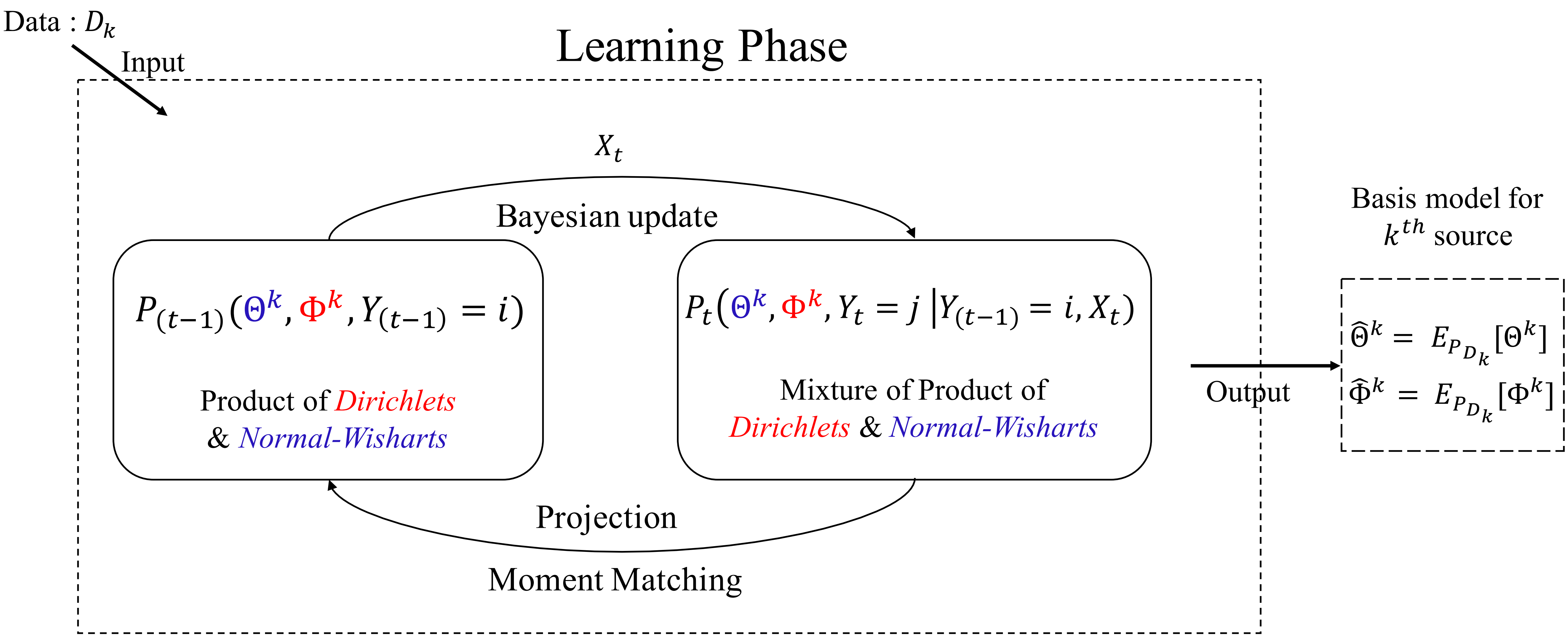
$$\propto \Pr(X_t|Y_t=j)\,\Pr(Y_t=j|Y_{t-1}=i)\,\underbrace{\Pr(\theta,\varphi,Y_{t-1}=i\mid X_{1:t-1})}$$

M_k

$$\sum_{h=1}^{H} w_{h,j}\, N(X_t;\mu_{h,j},\boldsymbol{\Lambda}_{h,j}^{-1}) \qquad \boldsymbol{\varphi}_{ij}$$

$$\prod_{m=1}^{M} Dir(\boldsymbol{\varphi}_m\,|\boldsymbol{\alpha}_m)\prod_{l=1}^{L} Dir(w_l\,;\beta_l)\prod_{h=1}^{H} NW\big(\mu_{h,l},\Lambda_{h,l}\,;\,m_{h,l},\kappa_{h,l},W_{h,l},\nu_{h,l}\big)$$

Mixture of terms in the posterior : use BMM for update


Data : D_k
Input
Learning Phase
X_t
Bayesian update
$P_{(t-1)}(\Theta^k, \Phi^k, Y_{(t-1)} = i)$
Product of Dirichlets
& Normal-Wisharts
$P_t(\Theta^k, \Phi^k, Y_t = j \mid Y_{(t-1)} = i, X_t)$
Mixture of Product of
Dirichlets & Normal-Wisharts
Projection
Moment Matching
Output
Basis model for
k^{th} source
$\hat{\Theta}^k = E_{P_{D_k}}[\Theta^k]$
$\hat{\Phi}^k = E_{P_{D_k}}[\Phi^k]$

Target Domain – Learning and Prediction

M_1

w_1

M_2

w_2

M_3

w_3

. . . .

$M_{(K-1)}$

$w_{(K-1)}$

M_K

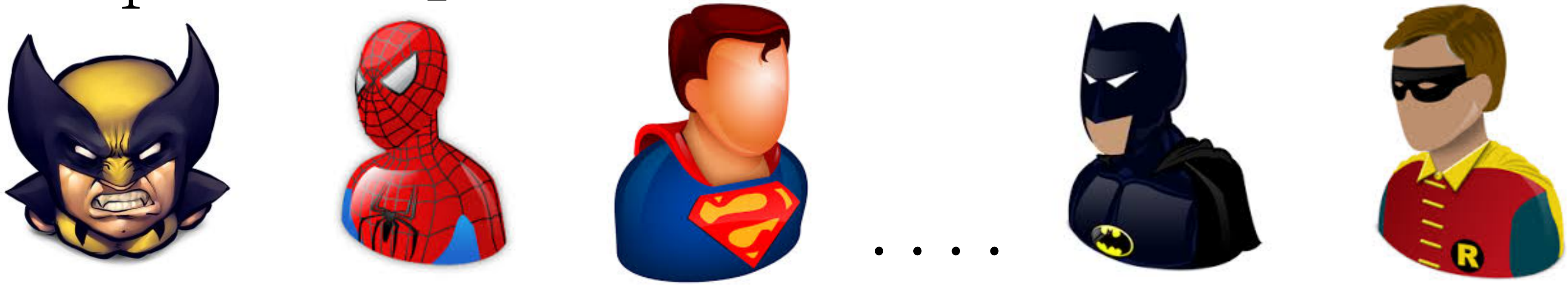
w_K

M

$$M = \sum w_i M_i$$

Target Domain – Learning and Prediction

M_1 M_2 M_3 $M_{(K-1)}$ M_K



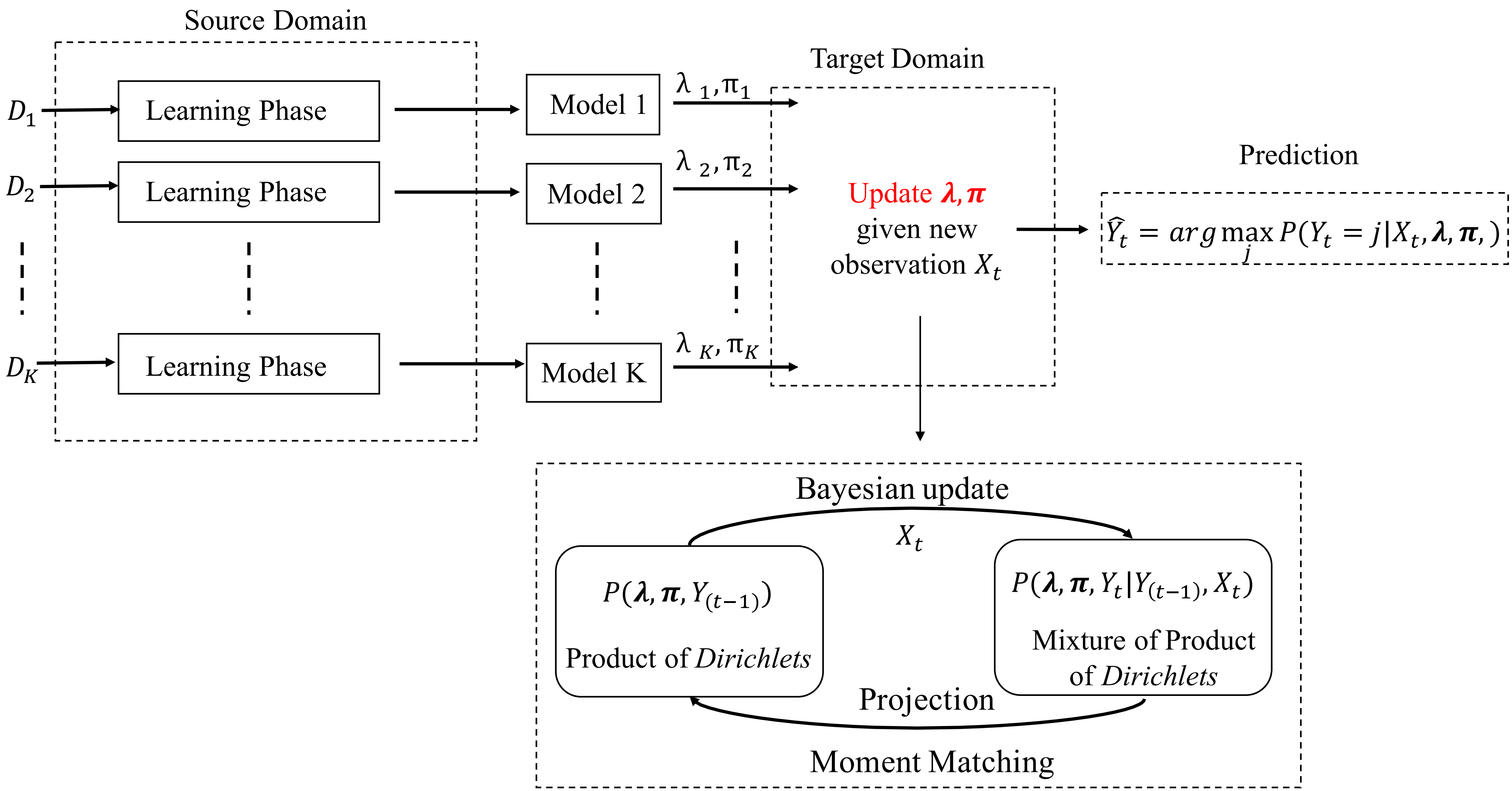
w_1 w_2 w_3 . . . . $w_{(K-1)}$ w_K

M

$$w_i = (\lambda_i, \pi_i)$$

$$\Pr(Y_t = j | Y_{t-1} = i) = \textstyle\sum_{k=1}^{k} \lambda_k \Pr\left(Y_t^k = j \middle| Y_{t-1}^k = i\right)$$

$$\Pr(X_t | Y_t = j) = \sum_{k=1}^{k} \pi_k \Pr\left(X_t^k \middle| Y_t^k = j\right)$$


Source Domain
D_1
D_2
D_K
Learning Phase
Learning Phase
Learning Phase
Model 1
Model 2
Model K
λ_1, π_1
λ_2, π_2
λ_K, π_K
Target Domain
Update $\boldsymbol{\lambda}, \boldsymbol{\pi}$ given new observation X_t
Prediction
$\widehat{Y}_t = \arg\max_j P(Y_t = j | X_t, \boldsymbol{\lambda}, \boldsymbol{\pi},)$
Bayesian update
X_t
$P(\boldsymbol{\lambda}, \boldsymbol{\pi}, Y_{(t-1)})$
Product of *Dirichlets*
$P(\boldsymbol{\lambda}, \boldsymbol{\pi}, Y_t | Y_{(t-1)}, X_t)$
Mixture of Product of *Dirichlets*
Projection
Moment Matching

Experiments

- Three real-world applications:
  1. Activity Recognition
  2. Sleep Stage Classification
  3. Network Flow Prediction

- Online transfer learning algorithm for prediction

- Comparison to BMM, oEM and RNN

- We use leave-one-out cross validation method

University of Waterloo's CHRONIC DISEASE PREVENTION INITIATIVE at the PROPEL CENTRE FOR POPULATION HEALTH IMPACT™

Activity Recognition

Network for Aging Research

Pascal Poupart

James Tung

Laura Middleton

Pabla Carbajal

Kayla Regan

University of Waterloo's
CHRONIC DISEASE PREVENTION INITIATIVE at the
PROPEL CENTRE FOR POPULATION HEALTH IMPACT™

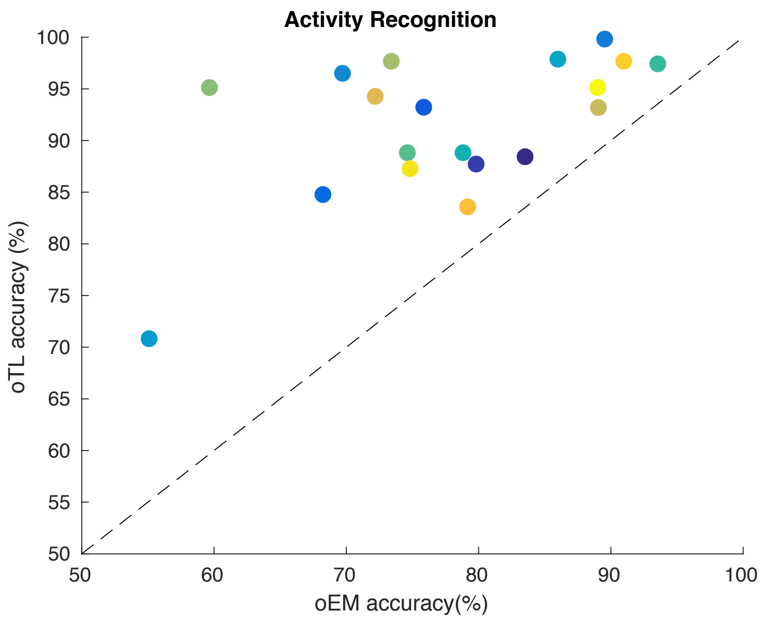
Activity Recognition

Network for Aging Research

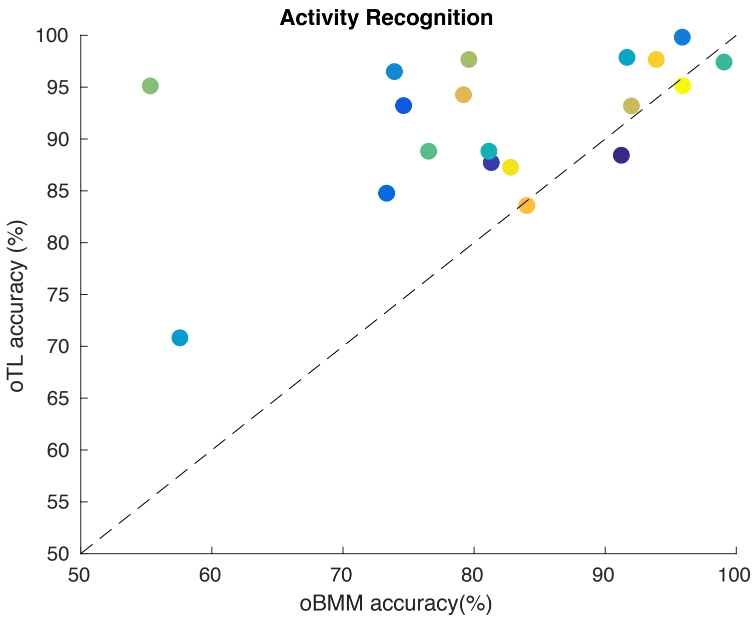
- Study to promote physical activity
- Labeled data collected from 19 participants using smartphones
- Activities include walking, standing, sitting, running and in a moving vehicle
- Aim – robust recognition algorithms for older adults or individuals with perturbed gait

Activity Recognition

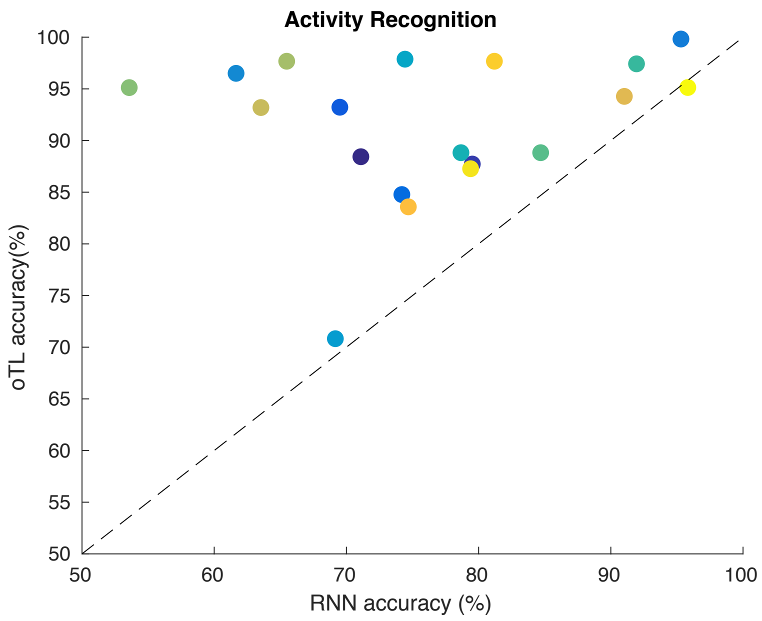
Network for Aging Research



Online EM



Online BMM



RNN

University of Waterloo's CHRONIC DISEASE PREVENTION INITIATIVE at the PROPEL CENTRE FOR POPULATION HEALTH IMPACT™

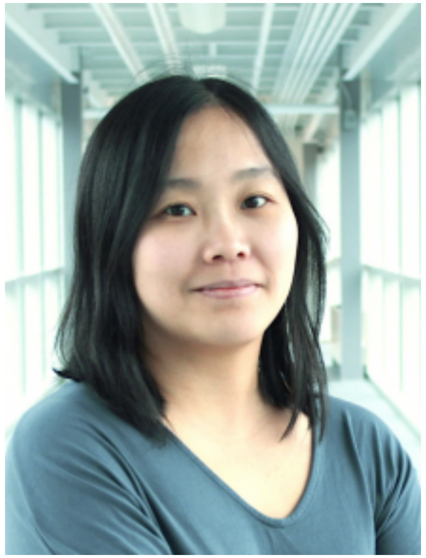
Activity Recognition

Network for Aging Research

TARGET DOMAIN	BASELINE	EM	RNN	TRANSFER LEARNING
PERSON 1	**91.29**	83.57	71.15	88.36 ↓
PERSON 2	81.37	79.87	79.58	**87.65**↑
PERSON 3	74.68	75.91	69.56	**93.15**↑
PERSON 4	73.39	68.29	74.25	**84.70**↑
PERSON 5	95.94	89.59	95.36	**99.75**↑
PERSON 6	73.98	69.77	61.71	**96.43**↑
PERSON 7	57.62	55.15	69.22	**70.75**↑
PERSON 8	91.72	86.05	74.49	**97.80**↑
PERSON 9	81.19	78.88	78.72	**88.75**↑
PERSON 10	**99.12**	93.60	92.00	97.35↓
PERSON 11	76.59	74.67	84.75	**88.75**↑
PERSON 12	55.36	59.71	53.63	**95.05**↑
PERSON 13	79.66	73.46	65.54	**97.60**↑
PERSON 14	92.06	89.11	63.59	**93.12**↑
PERSON 15	79.25	72.24	91.08	**94.20**↑
PERSON 16	**84.08**	79.23	74.74	83.51↓
PERSON 17	93.95	91.03	81.25	**97.60**↑
PERSON 18	82.84	74.88	79.45	**87.20**↑
PERSON 19	**95.97**	89.06	95.88	95.06↓

- All results are statistically significant
- Transfer Learning algorithm exhibited confusion b/w *standing – in a moving vehicle* and *sitting – in a moving vehicle* labels

Sleep Stage Classification



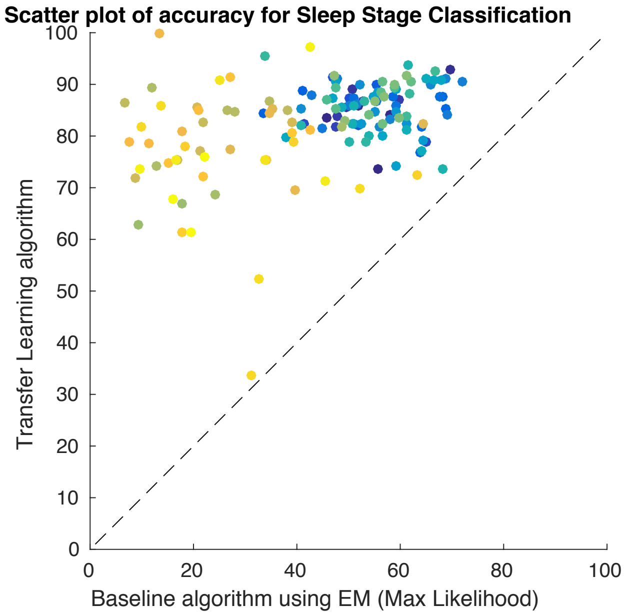
Edith Law

Mike Schäkermann

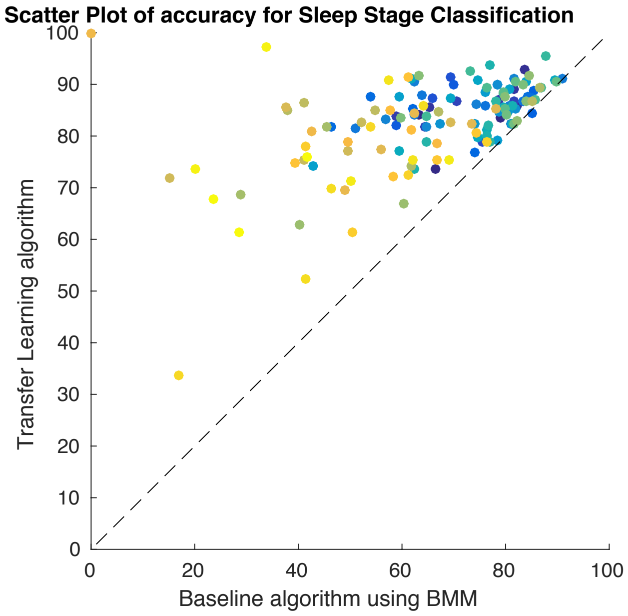
Sleep Stage Classification

- Study to analyze sleep patterns using EEG data
- Analysis of sleep patterns relevant in diagnosis of neurological disorders e.g. Parkinson
- Labeled data collected from 142 patients – 91 healthy and 51 with Parkinson's disease
- Sleep stages include wake, rapid eye movement, N1, N2 N3 and *unknown*

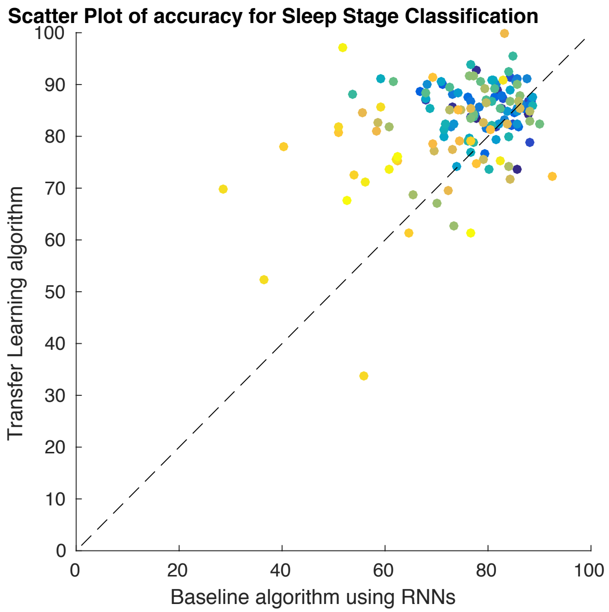
Sleep Stage Classification



Online EM



Online BMM

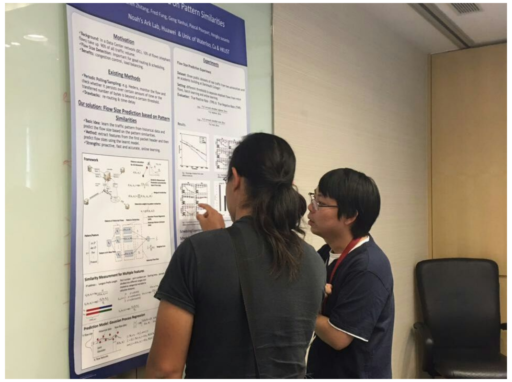


RNN

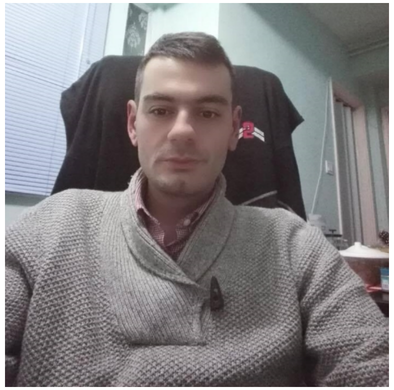
Transfer Learning performs better on 102 out of 142 patients compared to RNN

Network Flow Prediction

Pascal Poupart



Zhitang Chen



George Trimponias

Network Flow Prediction

- Prediction of future traffic → proactive network control
- Proactive network control helps in
  - Better network routing
  - priority scheduling
  - maximize rate control, min. transmission delay etc
- Used real traffic data from academic buildings with TCP flows
- Predict direction of flow b/w *Server* & *Client*

Network Flow Prediction

Target Domain	Baseline	EM	RNN	Transfer Learning
Source 1	72.00	54.90	**80.00**	71.02 ↓
Source 2	85.33	**89.10**	65.30	86.50↓
Source 3	80.33	81.90	**86.50**	83.33↑
Source 4	86.50	75.80	86.60	**87.17**↑
Source 5	**87.33**	82.80	81.70	86.00↓
Source 6	93.33	78.20	88.90	**93.50**↑
Source 7	95.17	90.70	93.50	**95.33**↑
Source 8	89.83	91.14	91.00	**91.63**↑
Source 9	76.67	75.68	**81.98**	78.83↑

Conclusion and Future Work

Contributions

- Online algorithm to tackle inter-population variability
- Online Bayeisan algorithm for sequential data with GMM emissions
- Application to three real world domains
- Comparision to other methods like RNN, oEM and BMM

Future Work

- Efficient choice of basis models
- Extension of online transfer learning technique to RNNs
- Theoretical properties of BMM – consistent?