

ΛMiDST TOOLBOX

Scalable Probabilistic Machine Learning

Andrés R. Masegosa

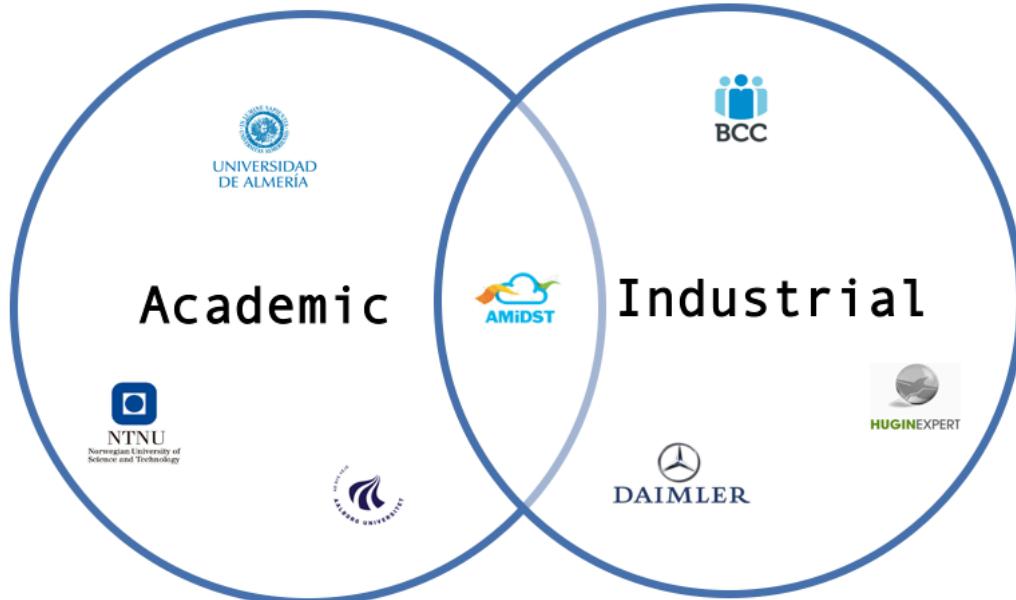
University of Almeria
andres.masegosa@ual.es

About us



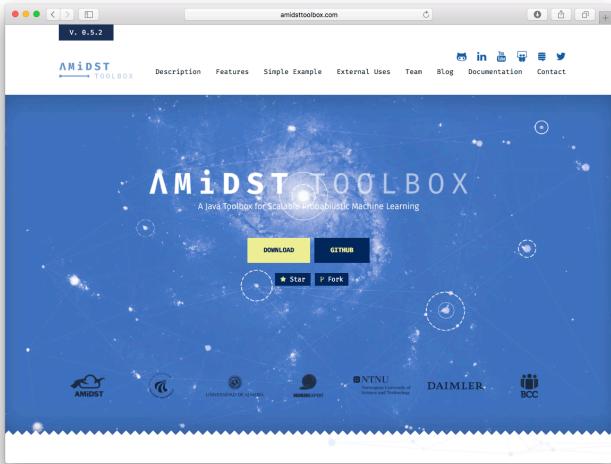
THE AMIDST CONSORTIUM

AMIDST
TOOLBOX

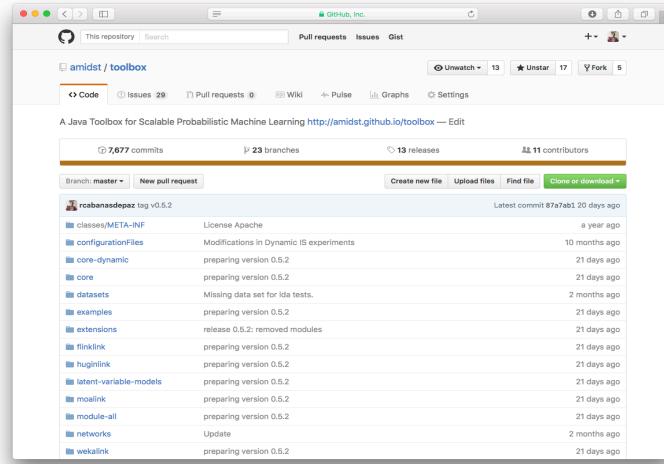


AMIDST TOOLBOX

AMIDST
TOOLBOX



www.amidsttoolbox.com



github.com/amidst/toolbox

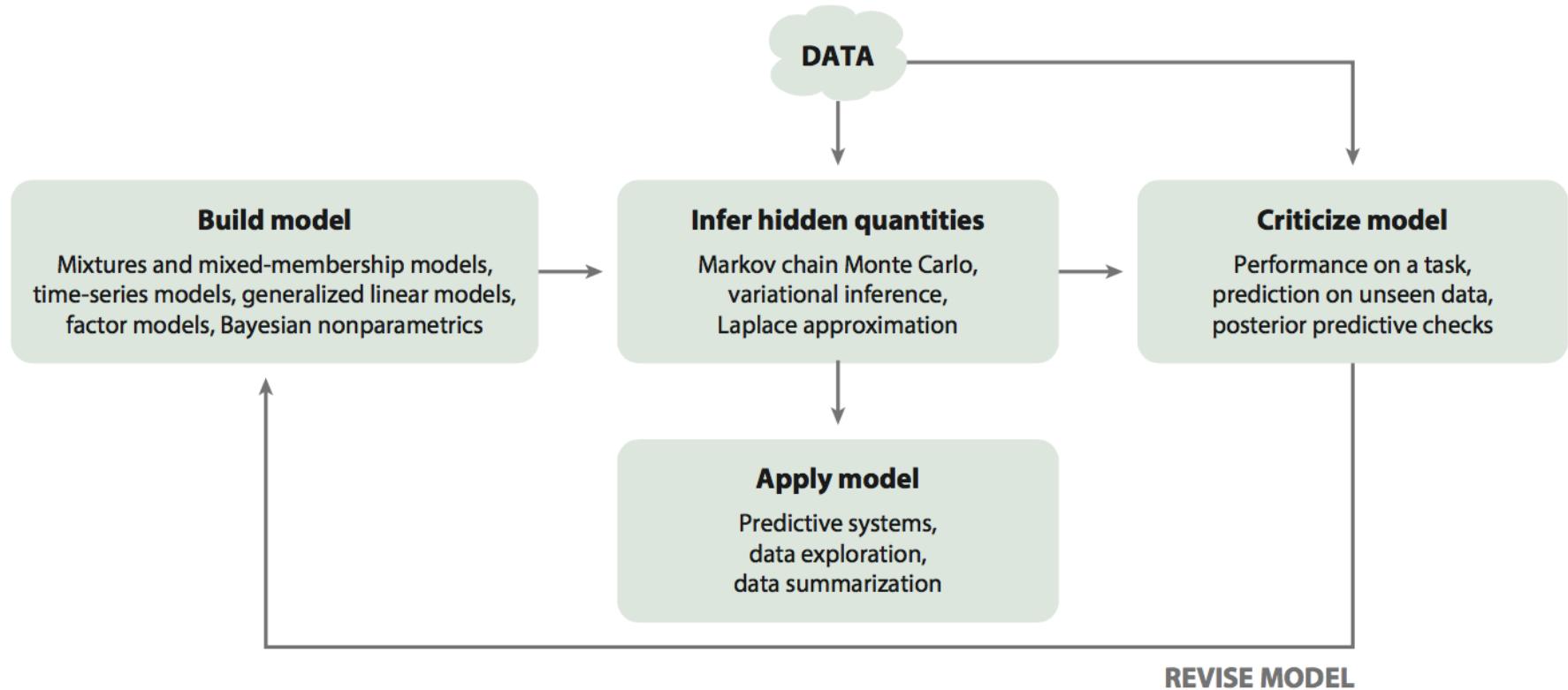


Apache
License 2.0

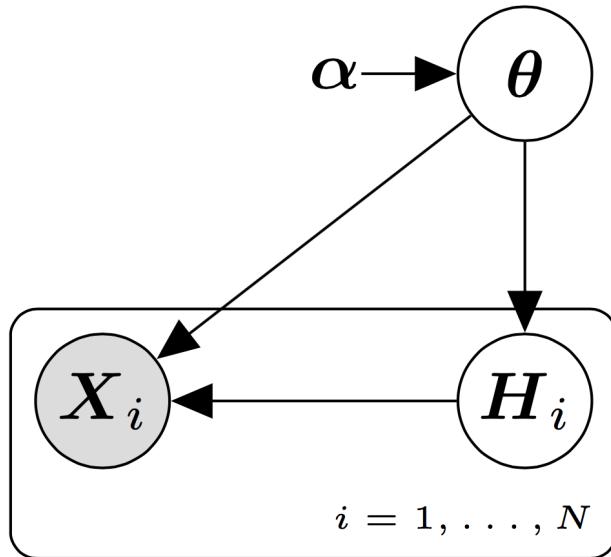


Probabilistic Machine Learning





Blei, David M. "Build, compute, critique, repeat: Data analysis with latent variable models." *Annual Review of Statistics and Its Application* 1 (2014): 203-232.



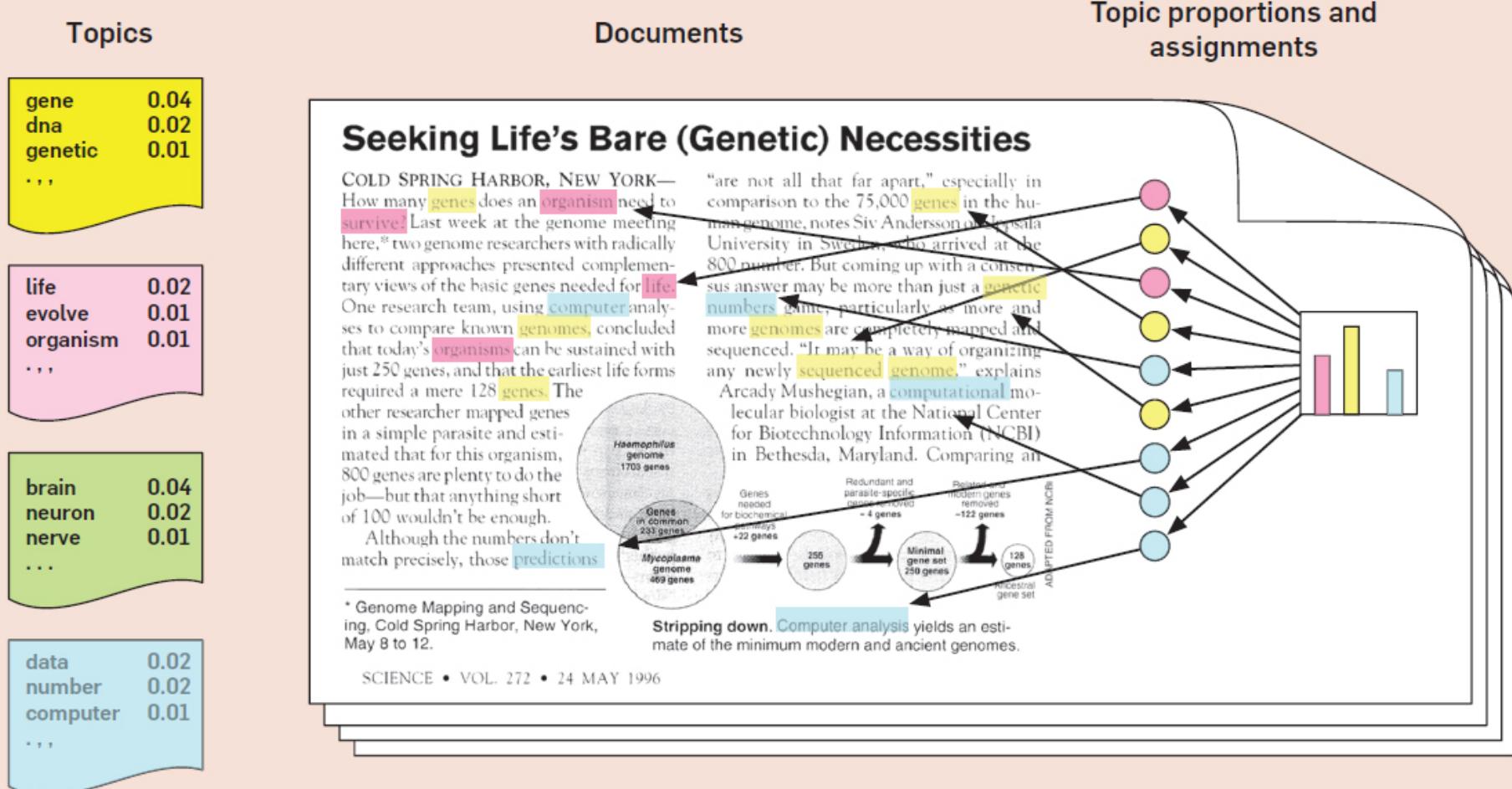
$$p(\theta, \mathbf{H}|D)$$

Latent Variable Models

Modeling non-observable mechanisms.
[Conjugate Exponential Family]

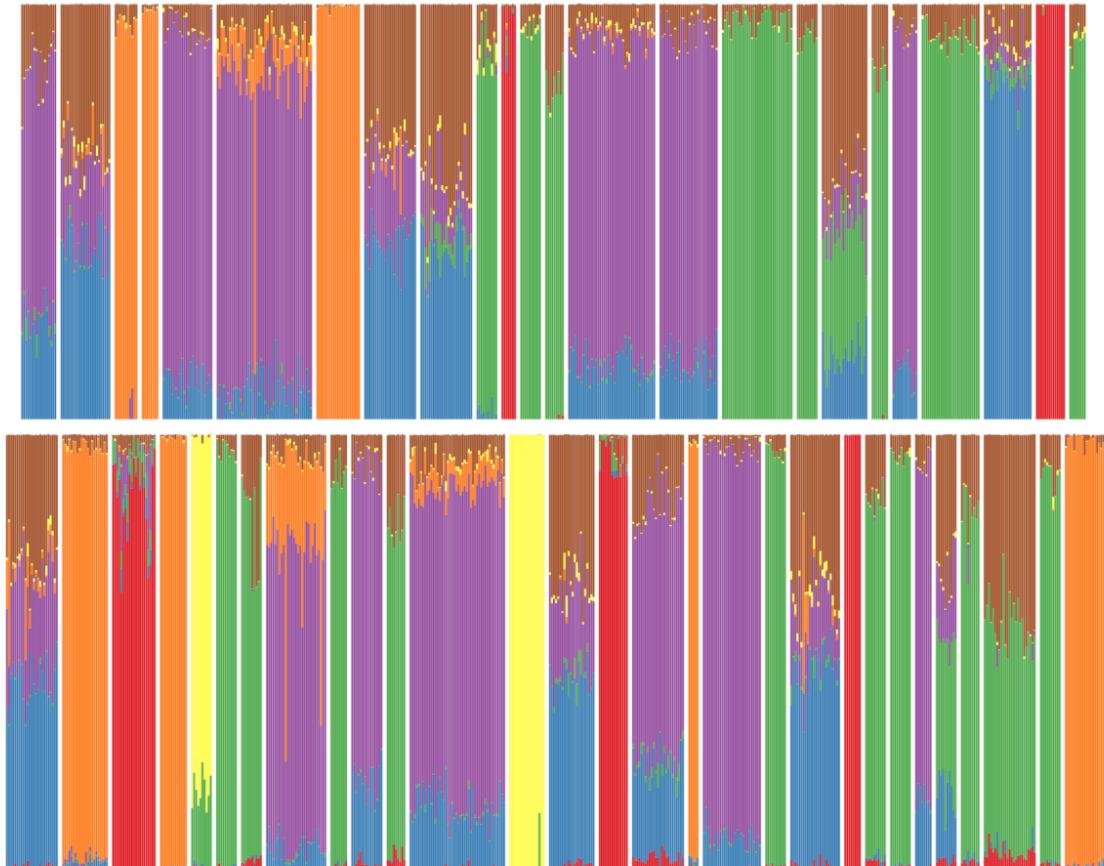
TEXT MODELING

AMIDST
TOOLBOX



David Blei, Probabilistic Topic Models, Communications of the ACM, Vol. 55 No. 4, Pages 77-84

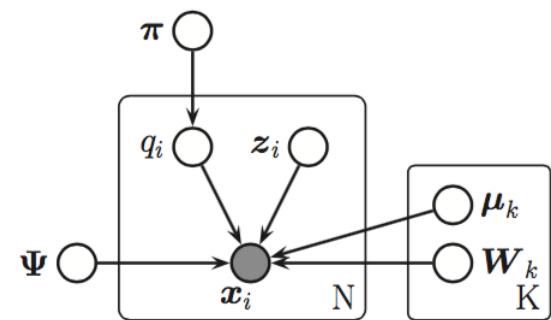




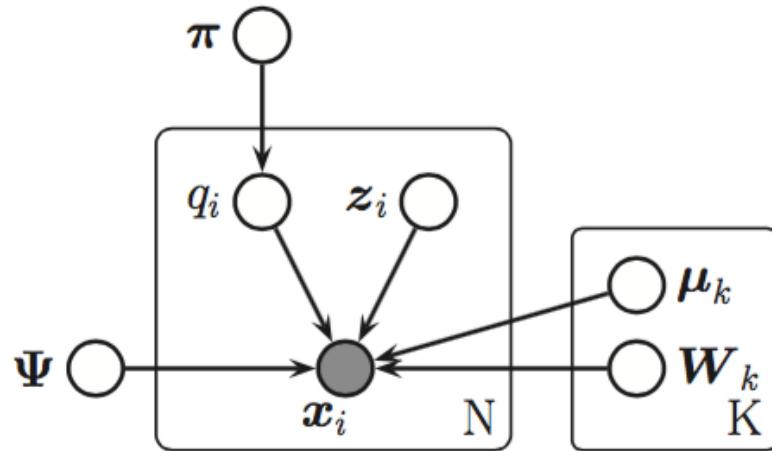
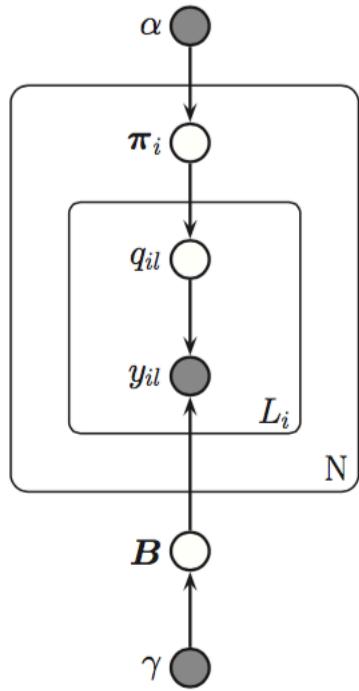
$$\begin{aligned}\beta_{k,\ell} &\sim \text{Beta}(a, b) \\ \theta_i &\sim \text{Dirichlet}(c) \\ x_{i,l} &\sim \text{Binomial}(2, \sum_k \theta_{i,k} \beta_{k,\ell})\end{aligned}$$

Gopalan, Prem, et al. Scaling probabilistic models of genetic variation to millions of humans.
Nature Research, 2016.



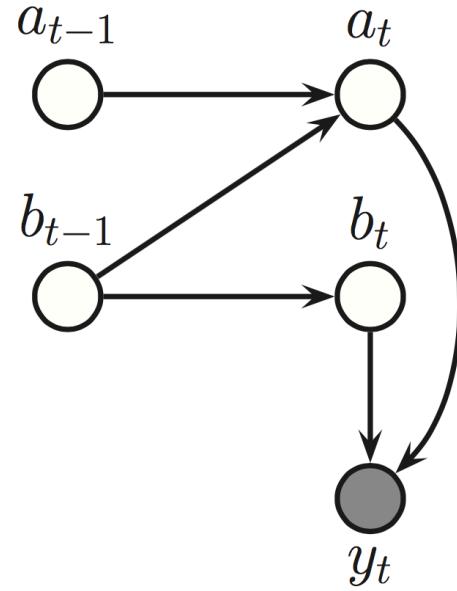
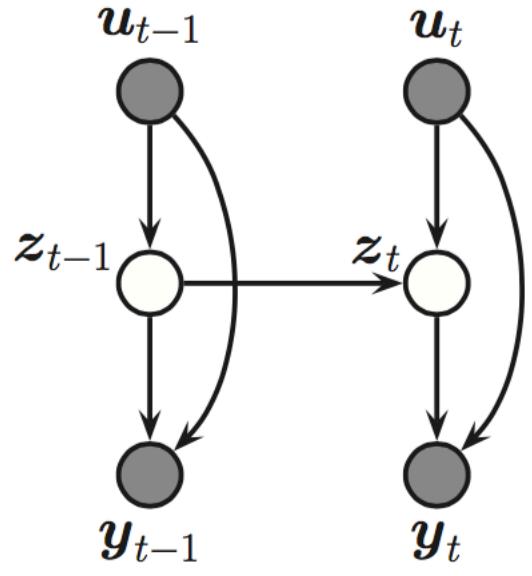


Trun et al. Automatic Differentiation Variational Inference. JMLR, 2016.



Examples of LVMs

Gaussian Mixture Models, Principal Component Analysis, Factor Analyzers, Latent Dirichlet Allocation, etc.



Dynamic/Temporal Models

Hidden Markov Models, Linear Dynamical Systems, State Space Models, Input-Output HMM, etc.



$$P(\theta | \mathbf{D})$$

Bayesian Learning

$$\{\theta : \theta \in \Theta\}$$

[Set of Models]

Data →

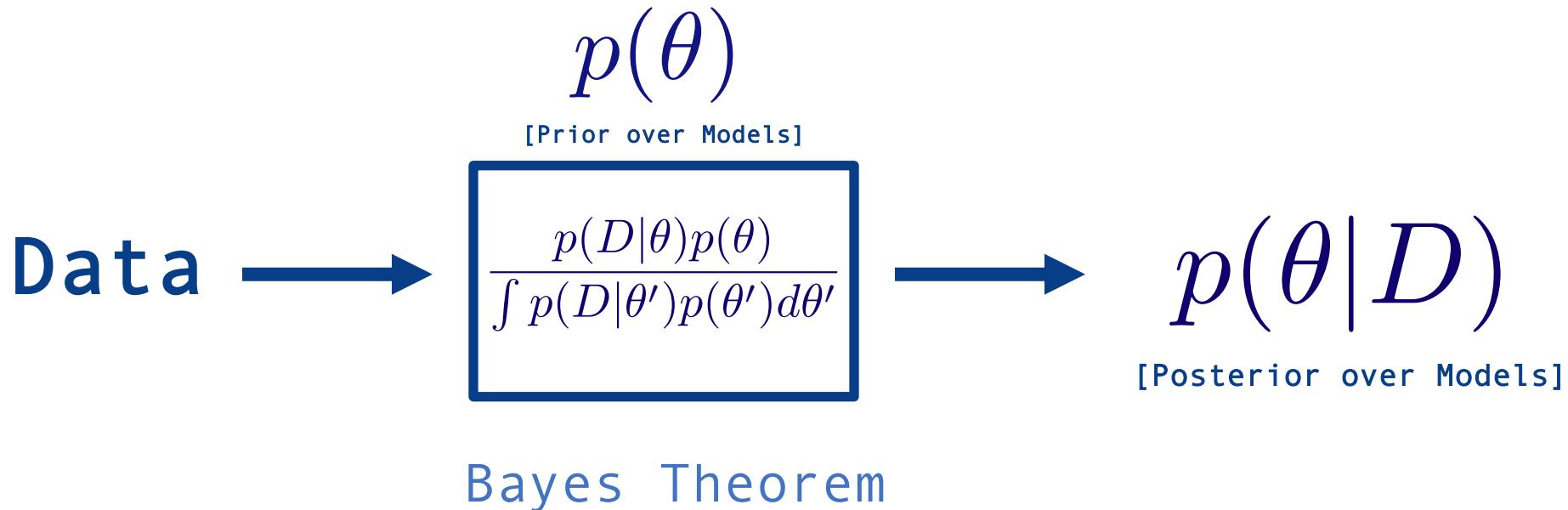
$$\arg \min_{\theta \in \Theta} \ell(D, \theta)$$



$$\theta^*$$

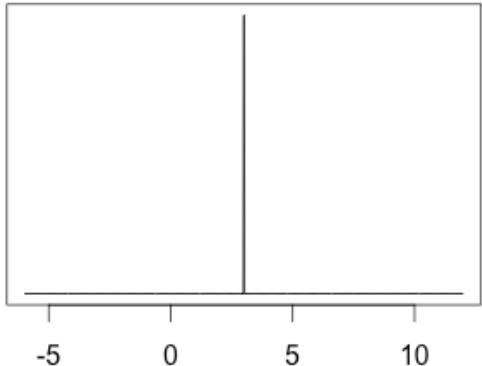
[Optimal Model]

Loss Minimization
(Stochastic Gradient Descent)

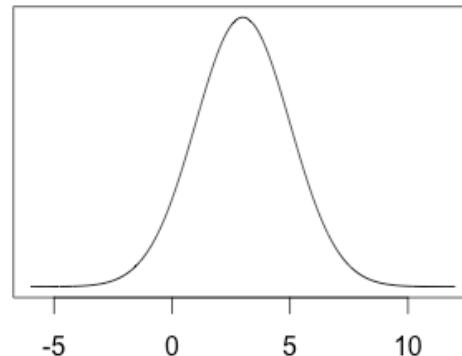


MODEL UNCERTAINTY

AMIDST
TOOLBOX



VS



$$\theta^*$$

[Point Estimate]

$$p(\theta|D)$$

[Bayesian Estimate]



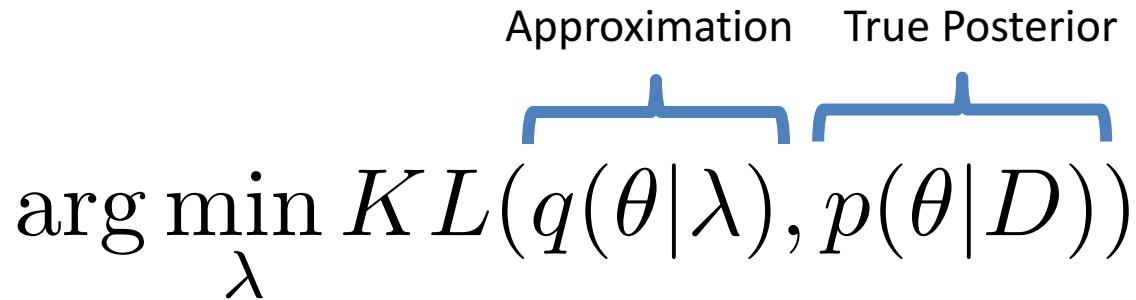
$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta')p(\theta')d\theta'}$$

How to compute it?



$$\arg \min_{\lambda} KL(q(\theta|\lambda), p(\theta|D))$$

Approximation True Posterior

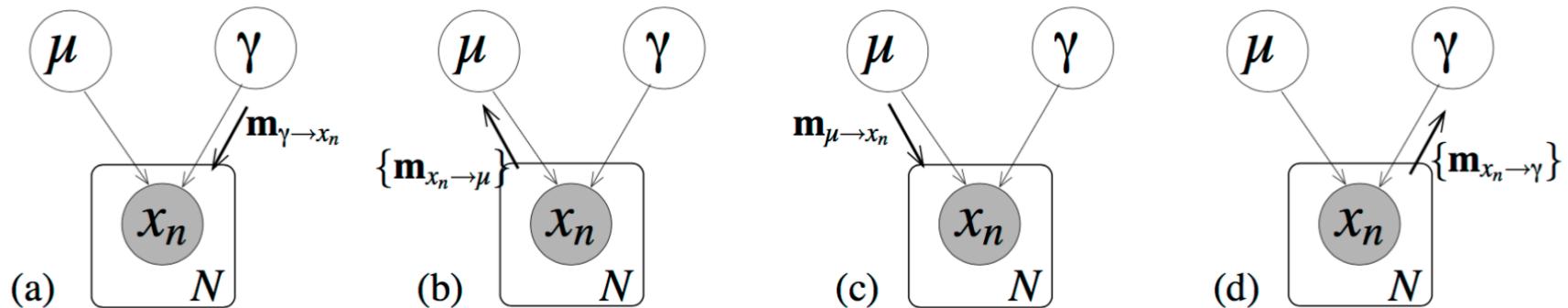


Variational Methods

- Pros: The inference problem is casted as an optimization problem.
- Pros: Deterministic approximation.
- Cons: Manual derivation of variational updating equations.

Hoffman, Matthew D., et al. "Stochastic variational inference." *Journal of Machine Learning Research* 14.1 (2013): 1303-1347.





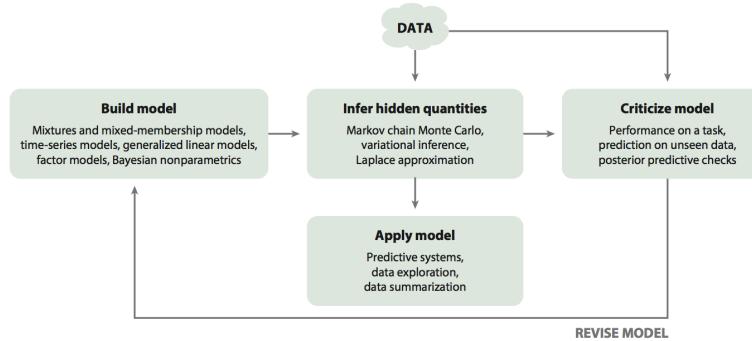
VMP: Automatic Variational Inference

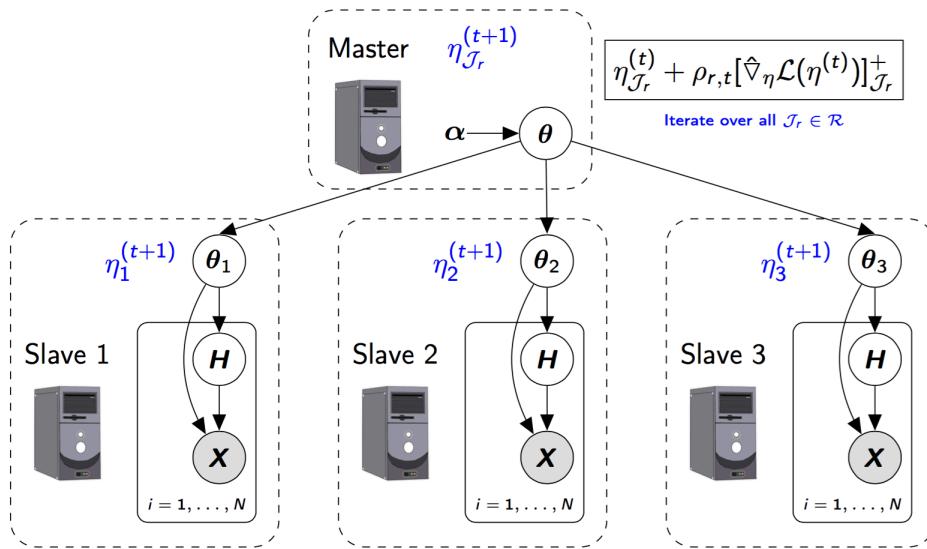
- Conjugate Exponential Family Models.
- Updating Equations can be expressed in terms of messages.
- Moments and Natural parameters are sent around.

Winn, John, and Christopher M. Bishop. "Variational message passing." *Journal of Machine Learning Research* 6.Apr (2005): 661-694.



$P(\theta|\text{BigData})?$



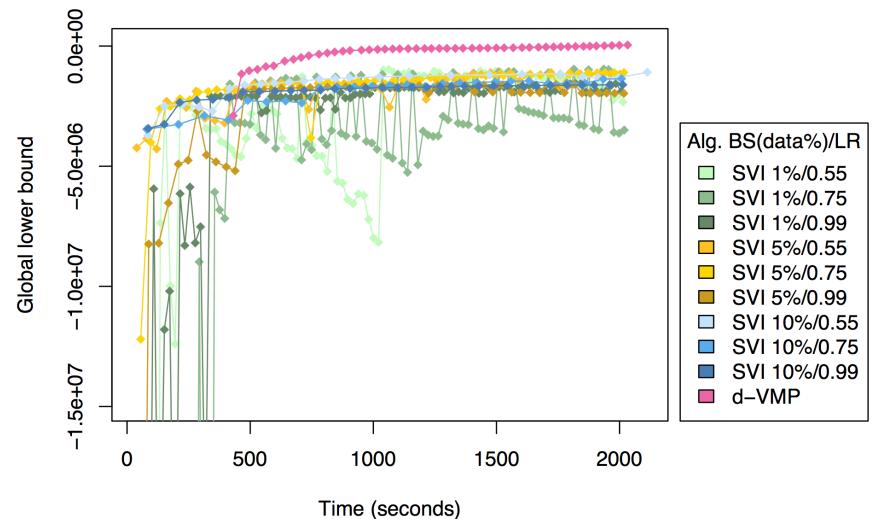
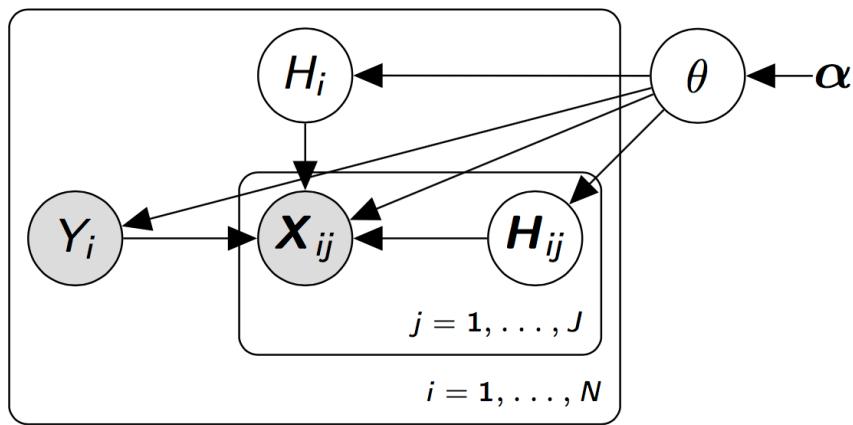


Masegosa, Andrés R., et al. "d-VMP: Distributed Variational Message Passing." *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*. 2016.

d-VMP Algorithm

Trick 1: Cast VMP as projected natural gradient ascent algorithm.

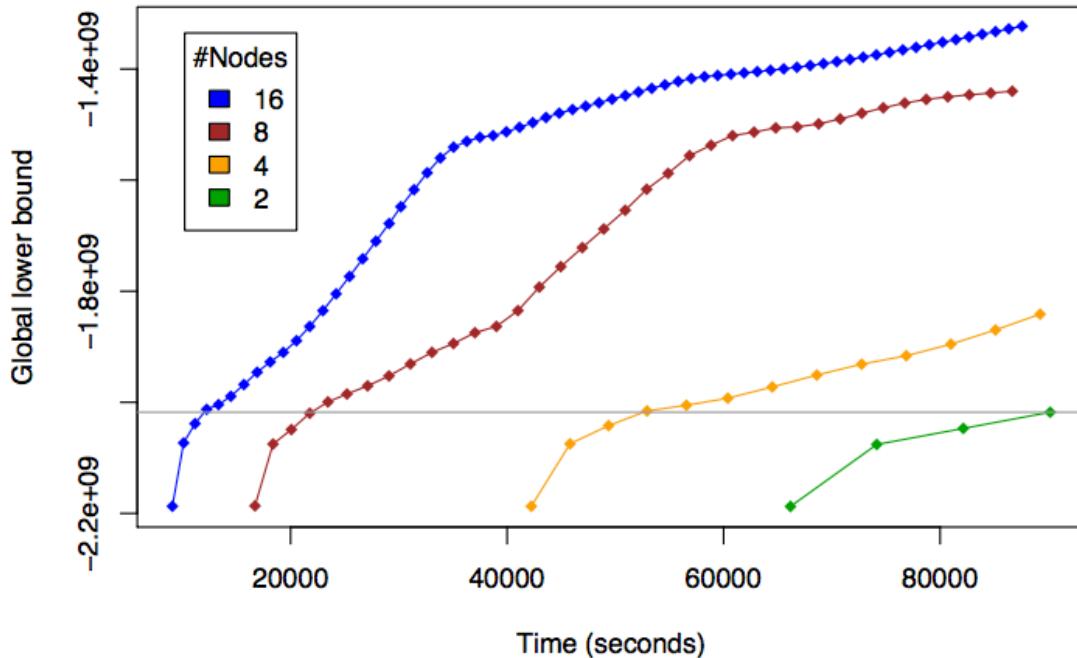
Trick 2: Exploit modern big data management tools such as Apache Flink.



Masegosa, Andrés R., et al. "d-VMP: Distributed Variational Message Passing." *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*. 2016.

d-VMP converges quicker than SVI

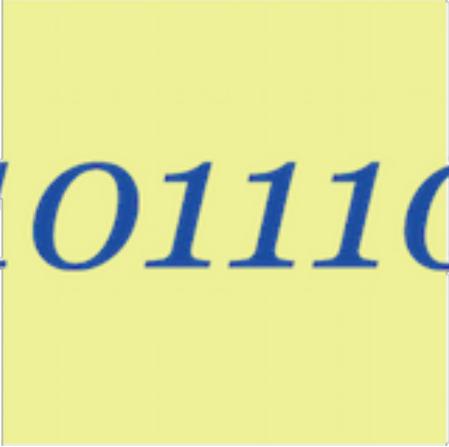
No hyper-parameter tuning



Masegosa, Andrés R., et al. "d-VMP: Distributed Variational Message Passing." *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*. 2016.

One billion node latent variable model

Experiment with Apache Flink on a AWS cluster.

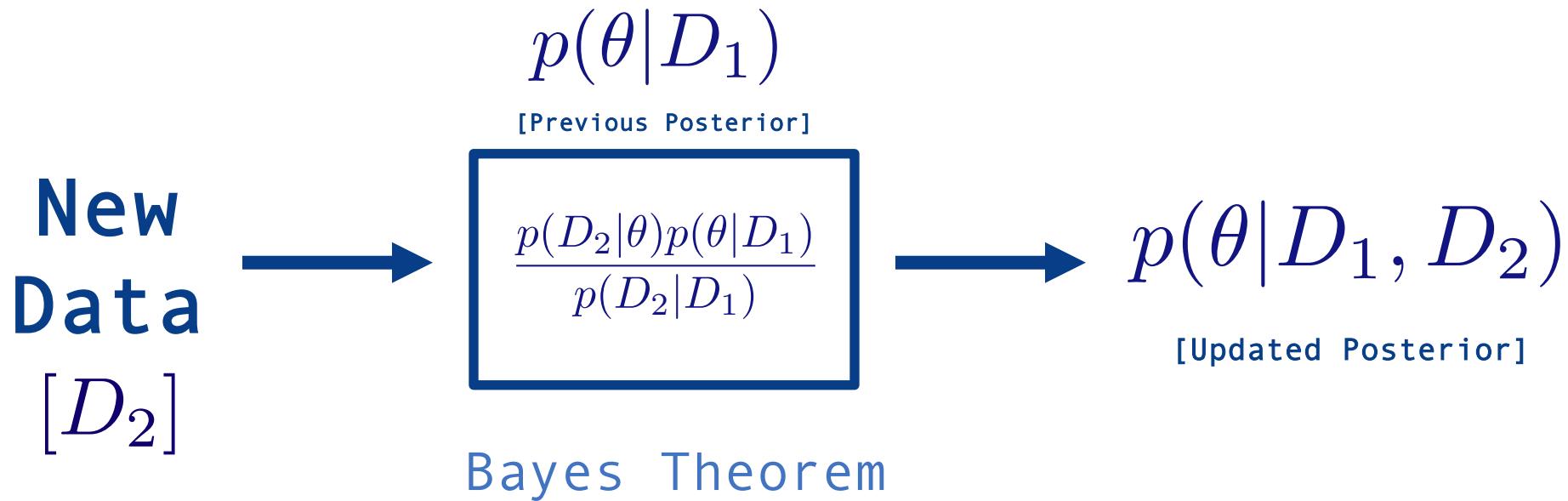
A large binary sequence "01011100" is displayed in blue text. It is overlaid on a yellow rectangular background that has rounded corners and a slight shadow, giving it the appearance of a card or a frame.

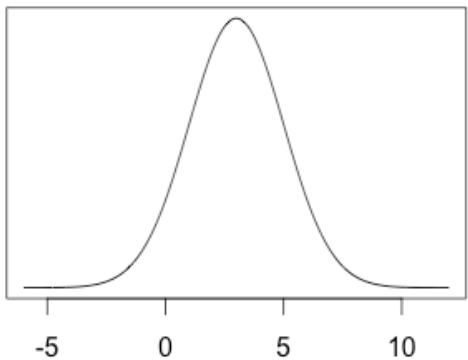
01011100

Data Streams

Update your model when new data is available.

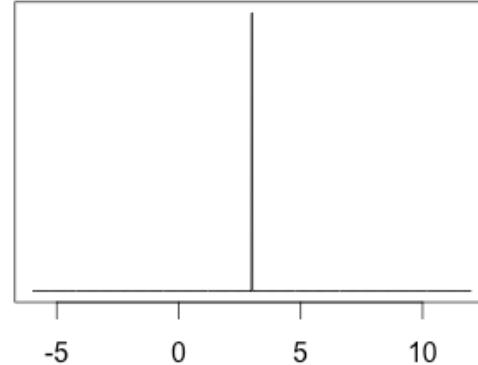






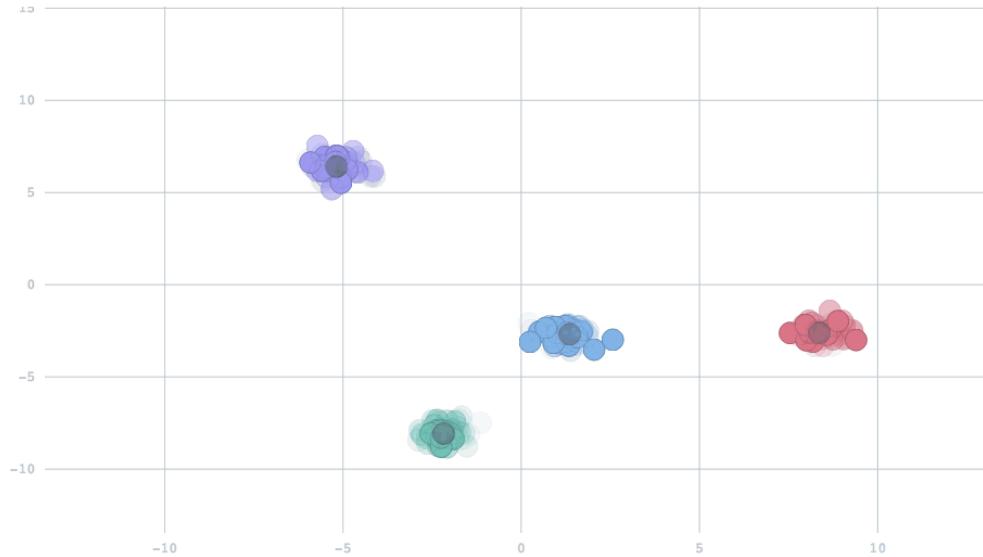
$$p(\theta | \mathbf{D}_{1:n})$$

[Capture Uncertainty]



$$p(\theta | \mathbf{D}_{1:\infty})$$

[Point Estimate]

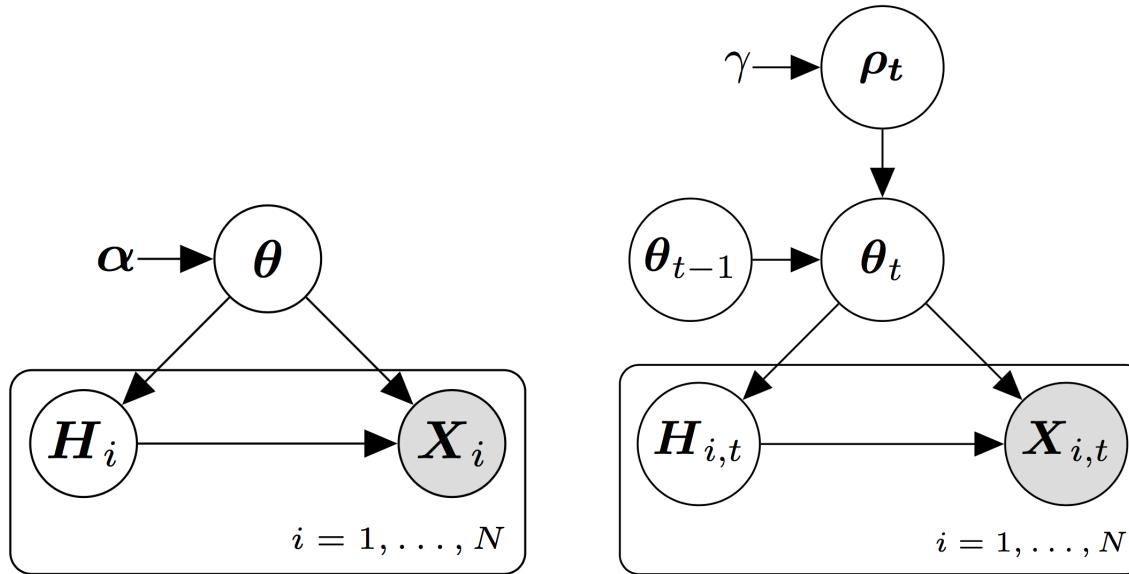


Freeman J. Introducing streaming k-means in Apache Spark 1.2.
<https://databricks.com/blog/2015/01/28/introducing-streaming-k-means-in-spark-1-2.html>

Concept Drift

Your data changes over time.



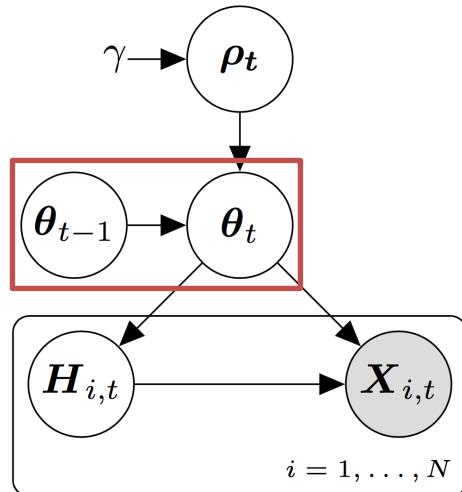


Masegosa, A. , et al. "Bayesian Models of Data Streams with Hierarchical Power Priors"
International Conference on Machine Learning. Sydney (Australia). 2017.

A Bayesian Model for Concept Drift

Non iid or exchangeability assumption.
Model parameters transition over time.



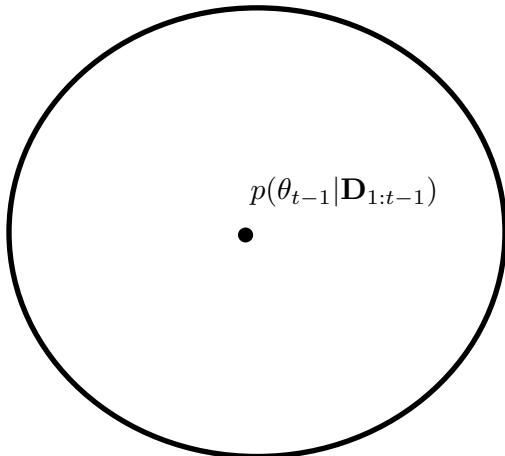


$$p(\theta_t | \mathbf{D}_{1:t-1}) = \int p(\theta_t | \theta_{t-1}) p(\theta_{t-1} | \mathbf{D}_{1:t-1}) d\theta_{t-1}$$

Explicit Transition Models

Assumption of single transition models, domain knowledge, etc
Outside of the exponential family

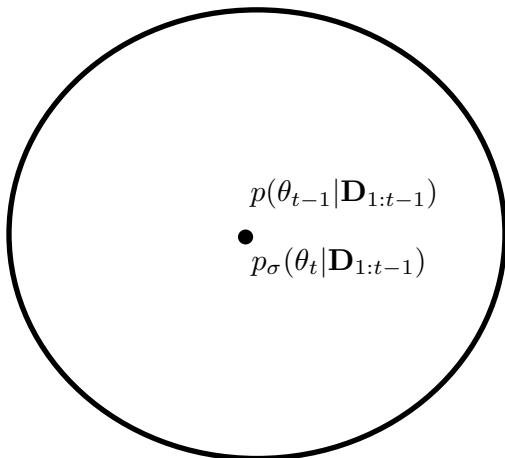




Implicit Transition Models

Maximum Entropy criteria

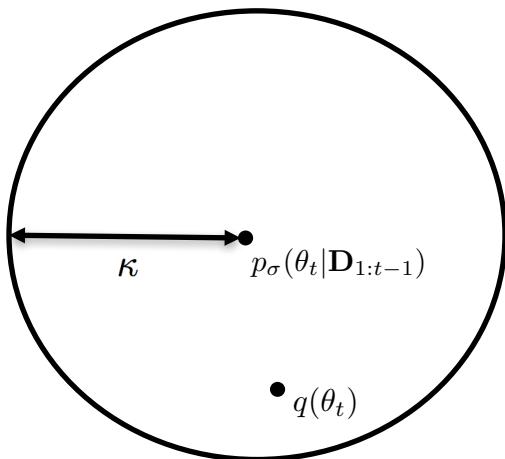




$$p_\delta(\theta_t | \mathbf{D}_{1:t-1}) = \int \delta(\theta_t - \theta_{t-1}) p(\theta_{t-1} | \mathbf{D}_{1:t-1}) d\theta_{t-1}$$

Implicit Transition Models

Maximum Entropy criteria

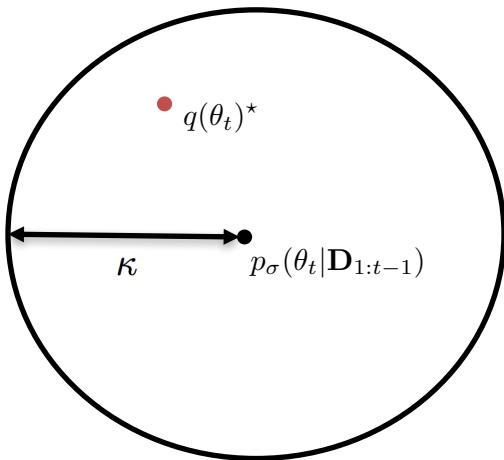


$$KL(q(\theta_t), p_\sigma(\theta_t | \mathbf{D}_{t-1})) \leq \kappa$$

Implicit Transition Models

Maximum Entropy criteria





$$q(\theta_t)^\star = \arg \max_{q(\theta_t)} H(q(\theta_t))$$
$$KL(q(\theta_t), p_\sigma(\theta_t | \mathbf{D}_{t-1})) \leq \kappa$$

Implicit Transition Models

Maximum Entropy criteria



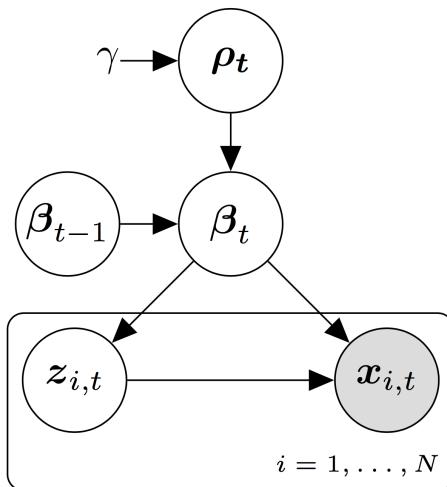
$$q^*(\theta_t) \propto p_u(\theta_t)^{(1-\rho)} p_\sigma(\theta_t | \mathbf{D}_{1:t-1})^\rho$$

$$\rho \in [0, 1]$$

Implicit Transition Models

$\rho = 0$ implies absolute forgetting of past data.
 $\rho = 1$ implies no forgetting at all past data.





$$\rho_t \sim \text{TruncatedExponential}(\gamma)$$

Masegosa, A. , et al. "Bayesian Models of Data Streams with Hierarchical Power Priors"
International Conference on Machine Learning. Sydney (Australia). 2017.

Adaptive Forgetting Mechanism

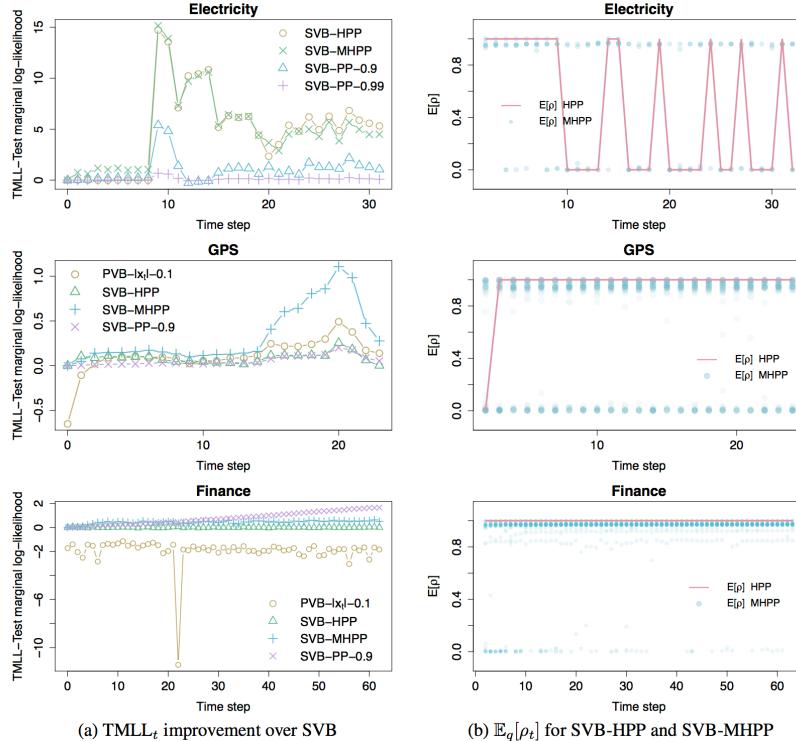
ρ is time-indexed.

Adaptive forgetting mechanism.

Closed-form (automatic) variational updating equations.

BAYESIAN MODELS OF DATA STREAMS

AMIDST
TOOLBOX



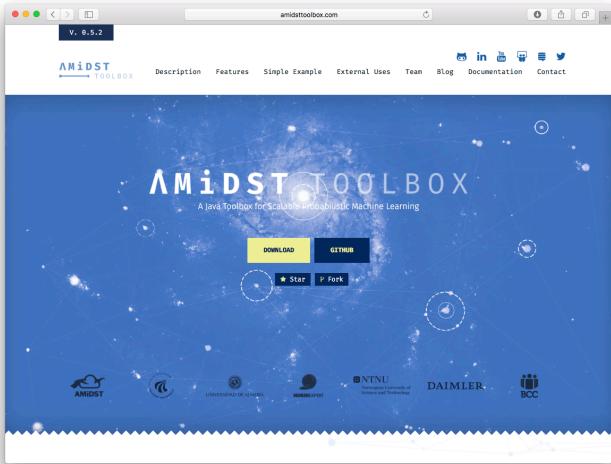
State-of-the-art performance

Trade, GPS and Financial data.
Different Latent Variable Models.

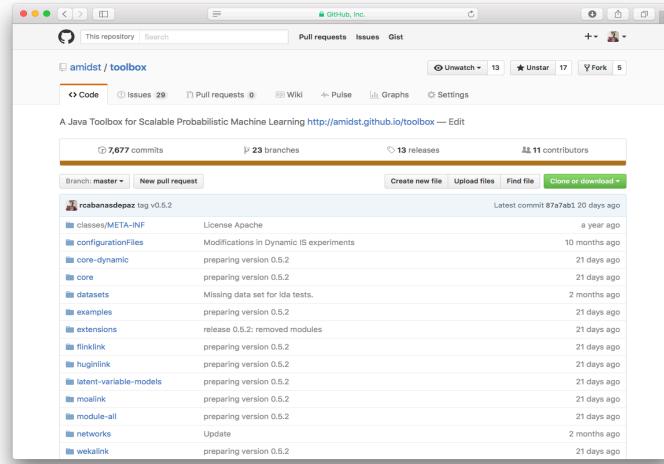


AMIDST TOOLBOX

AMIDST
TOOLBOX



www.amidsttoolbox.com



github.com/amidst/toolbox



Apache
License 2.0



Use Case

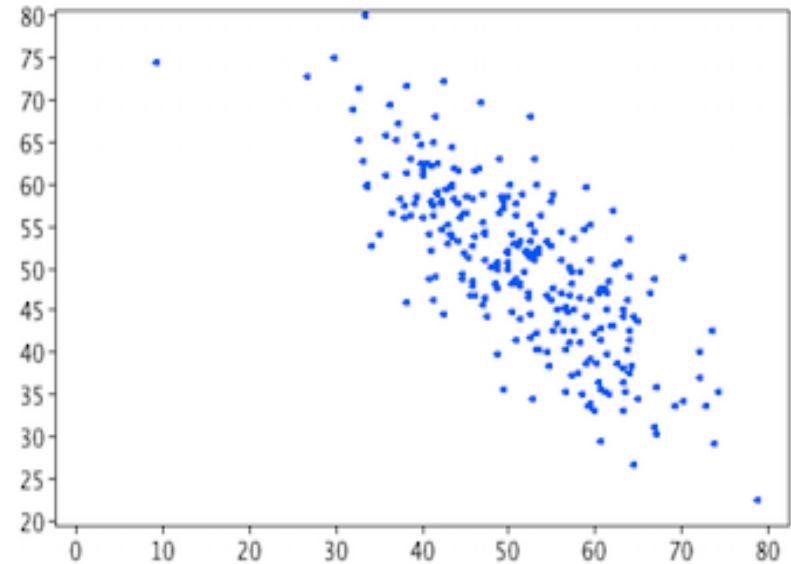
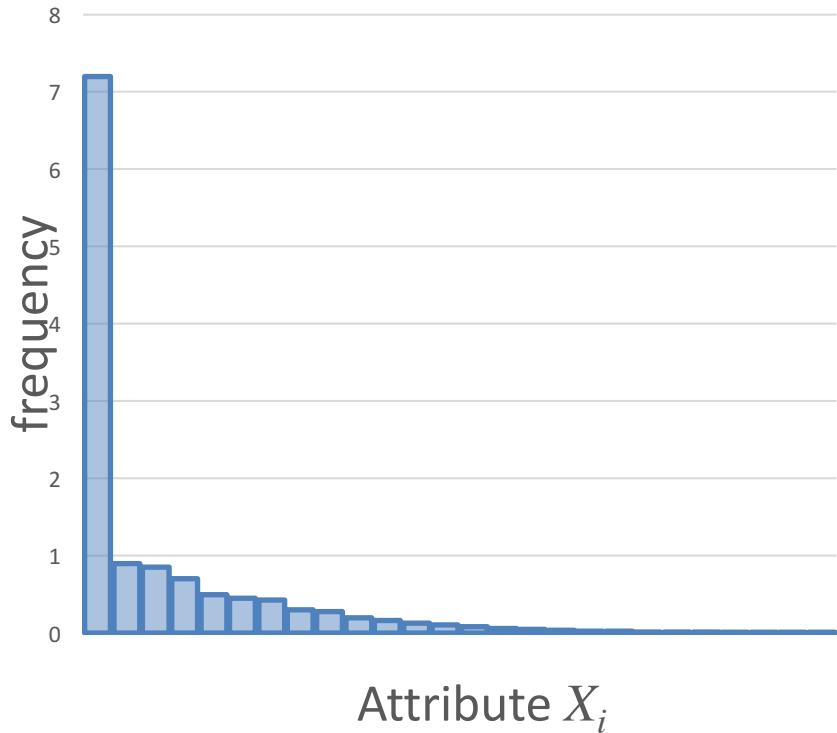




Predicting Defaulting Clients

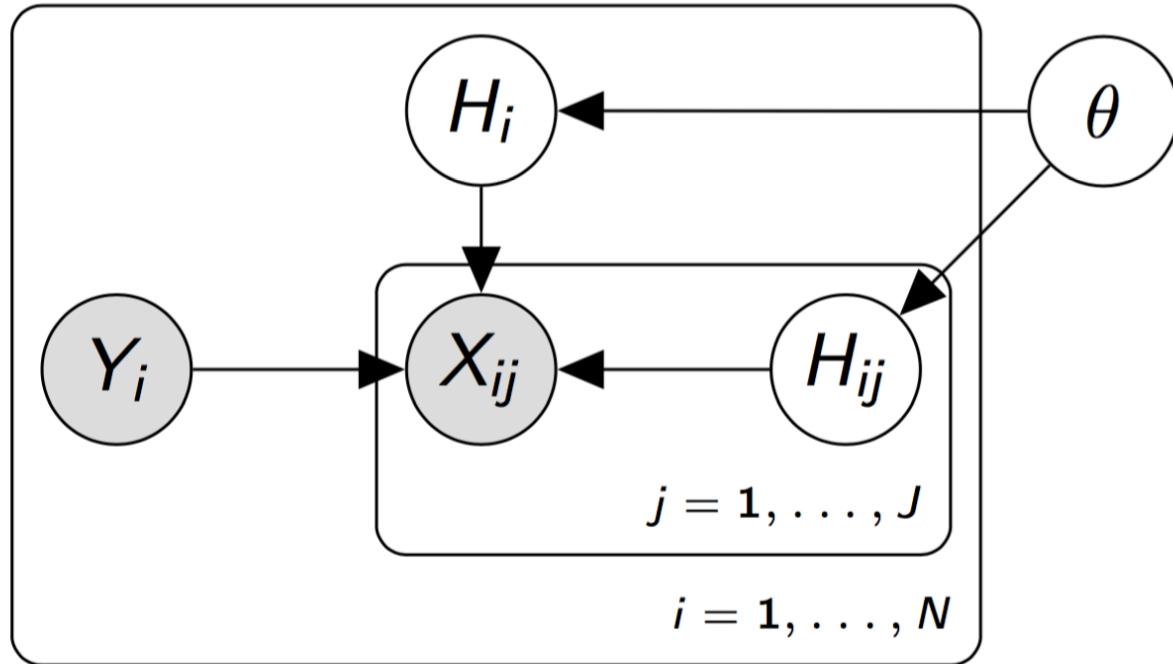
Predicts probability a customer will default within 2 years

RUNNING USE CASE



- Daily data for millions of clients
- Tons of missing data.
- Odd distributions.





Custom Gaussian Mixture Model

H_{ij} defines local mixture

H_i defines a global mixture.

```
//Set-up Flink session.  
final ExecutionEnvironment env = ExecutionEnvironment.getExecutionEnvironment();  
  
//Load the data stream  
String filename = "hdfs://dataFlink_month0.arff";  
DataFlink<DataInstance> data =  
    DataFlinkLoader.loadDataFromFolder(env, filename, false);  
  
//Build the model  
Model model = new CustomGaussianMixture(data.getAttributes());  
    .setClassIndex(2);
```



SCALABLE INFERENCE

RUNNING CODE EXAMPLE

AMIDST
TOOLBOX

```
//Set-up Flink session.  
final ExecutionEnvironment env = ExecutionEnvironment.getExecutionEnvironment();  
  
//Load the data stream  
String filename = "hdfs://dataFlink_month0.arff";  
DataFlink<DataInstance> data =  
    DataFlinkLoader.loadDataFromFolder(env, filename, false);  
  
//Build the model  
Model model = new CustomGaussianMixture(data.getAttributes())  
    .setClassIndex(2);  
  
//Learn the model  
model.updateModel(data);
```

DATA STREAMS

RUNNING CODE EXAMPLE

```
//Set-up Flink session.  
final ExecutionEnvironment env = ExecutionEnvironment.getExecutionEnvironment();  
  
//Load the data stream  
String filename = "hdfs://dataFlink_month0.arff";  
DataFlink<DataInstance> data =  
    DataFlinkLoader.loadDataFromFolder(env, filename, false);  
  
//Build the model  
Model model = new CustomGaussianMixture(data.getAttributes())  
    .setClassIndex(2);  
  
//Learn the model  
model.updateModel(data);  
  
//Update your model  
for(int i=1; i<12; i++) {  
    filename = "dataFlink_month"+i+".arff";  
    data = DataFlinkLoader.loadDataFromFolder(env, filename, false);  
    System.out.println(model.predict(data));  
    model.updateModel(data);  
}
```



Predicting Defaulting Clients

- Old BCC's models based on logistic regression got an AUC around 0.8
- AMIDST's models gets an AUC over 0.9
- Model will be in production soon.

.



Thanks for your attention

www

www.amidsttoolbox.com

@

contact@amidsttoolbox.com



[@AmidstToolbox](https://twitter.com/AmidstToolbox)

AMIDST
→ TOOLBOX

Use Case II

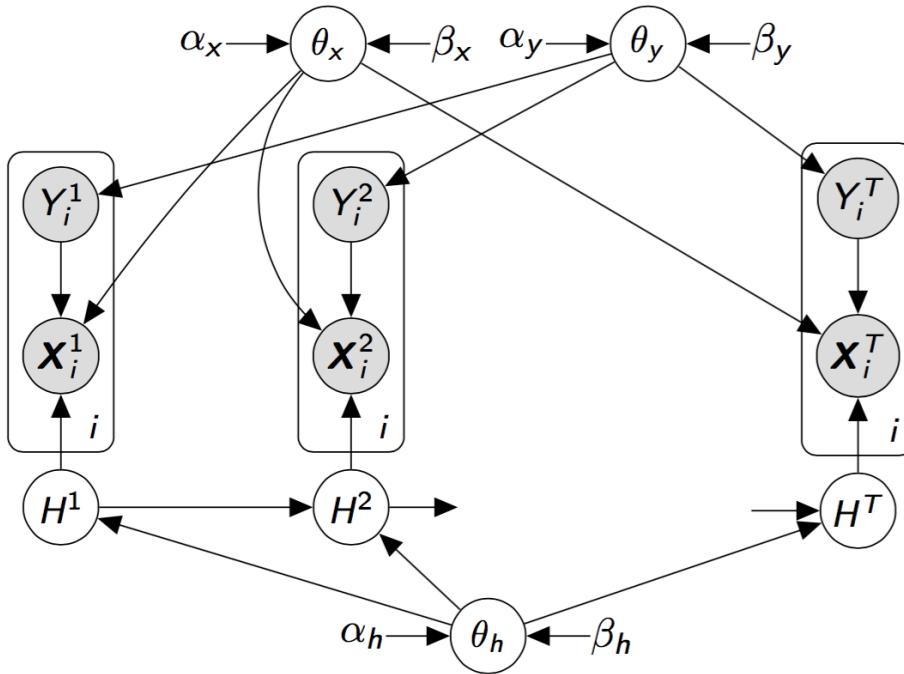




Tracking Concept Drift

Detects changes in customer profiles during Spanish financial crisis

CONCEPT DRIFT DETECTION

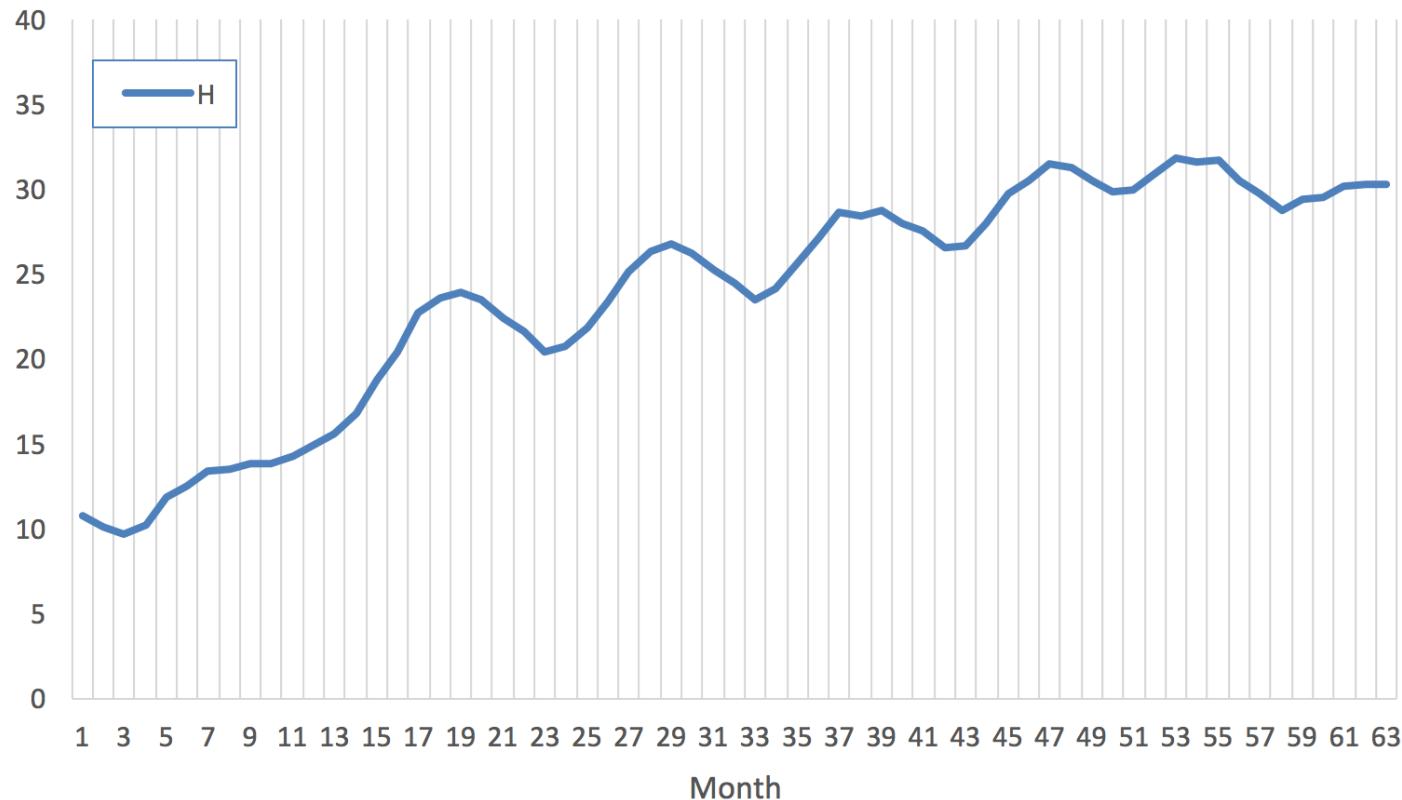


Hidden Variables are used to capture changes in customer profile



CONCEPT DRIFT DETECTION RESULTS

AMIDST
TOOLBOX



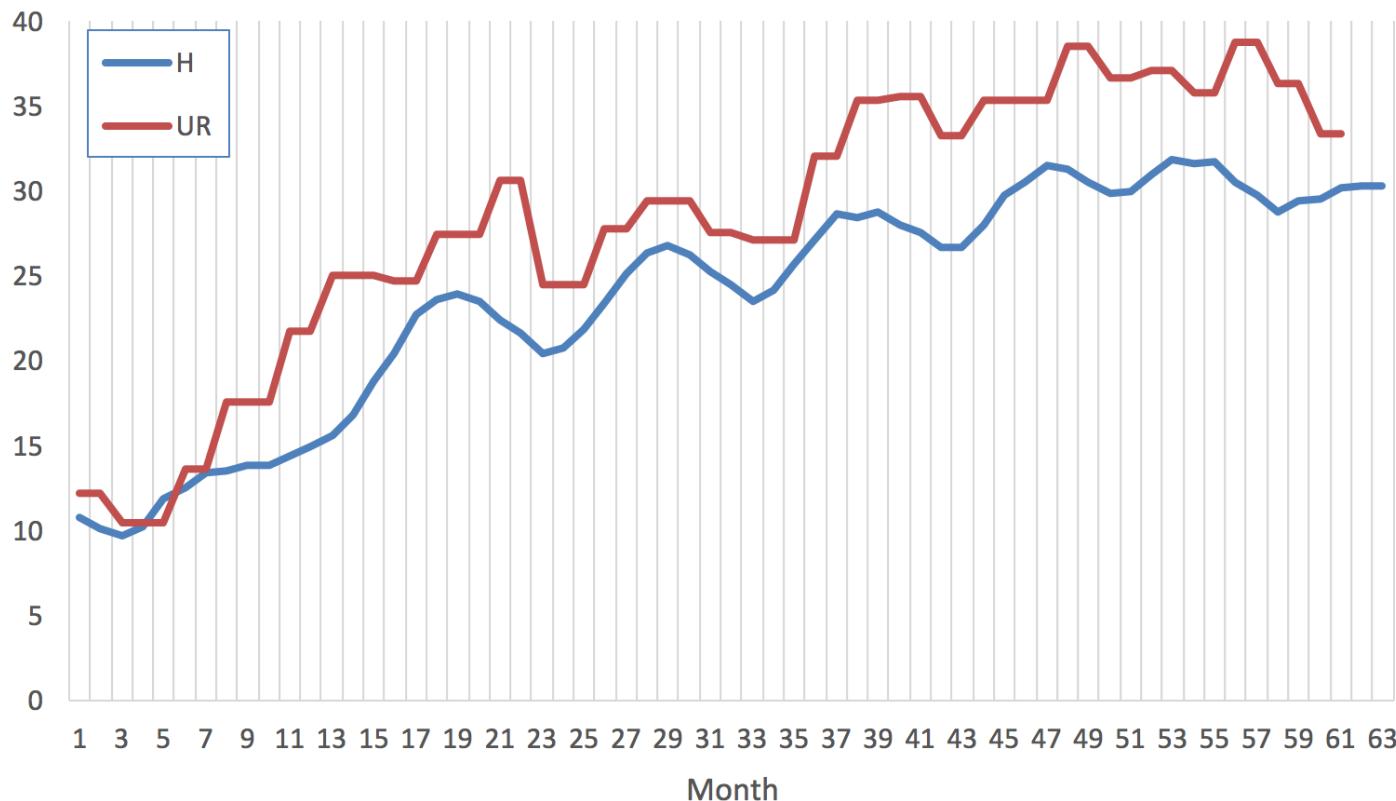
Hidden Variable Captures Concept Drift

Drift Pattern: Seasonal + Global trend



CONCEPT DRIFT DETECTION RESULTS

AMIDST
TOOLBOX



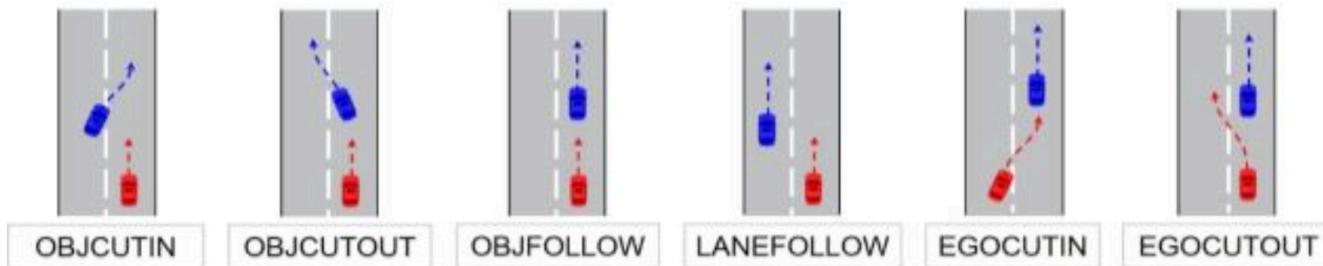
Unemployment Rate main driver of Concept Drift

Hidden Variable correlates with unemployment rate ($\rho = 0.961$)



Use Case III



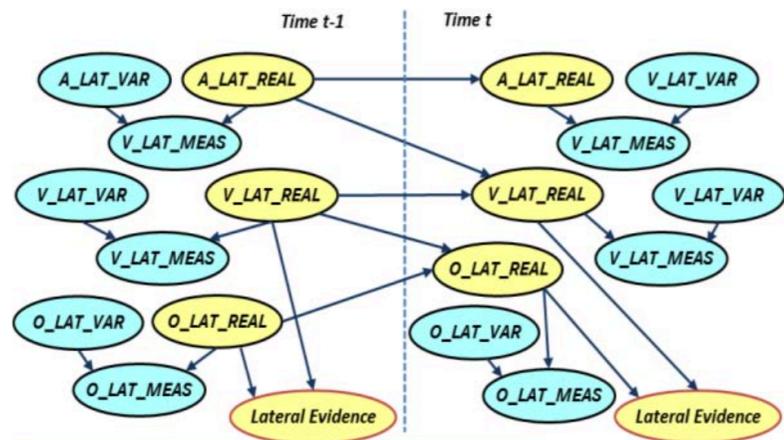
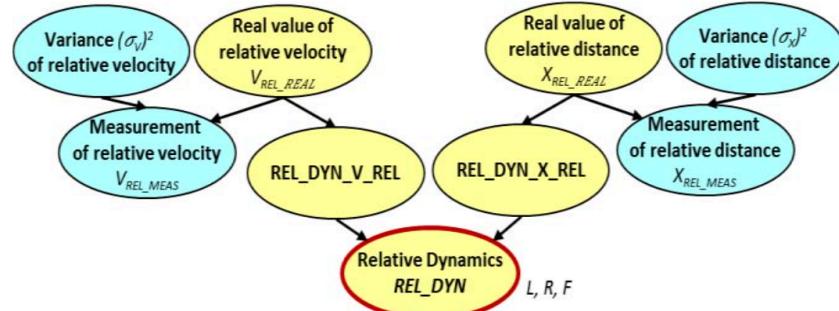
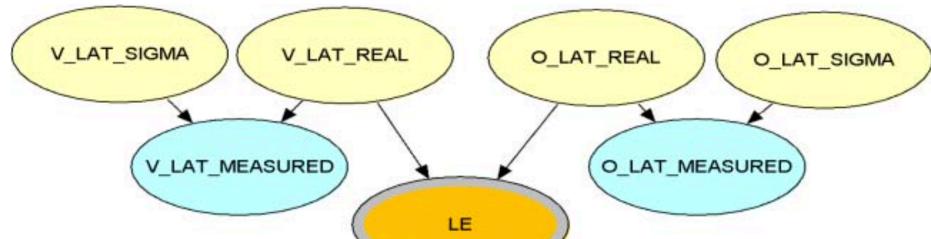
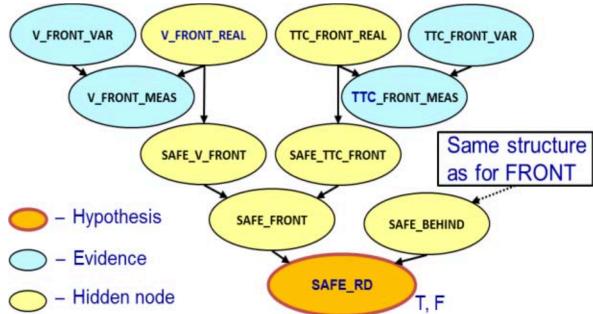


Weidl, Galia, et al. "Early Recognition of Maneuvers in Highway Traffic." *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer International Publishing, 2015.

Maneuver Recognition

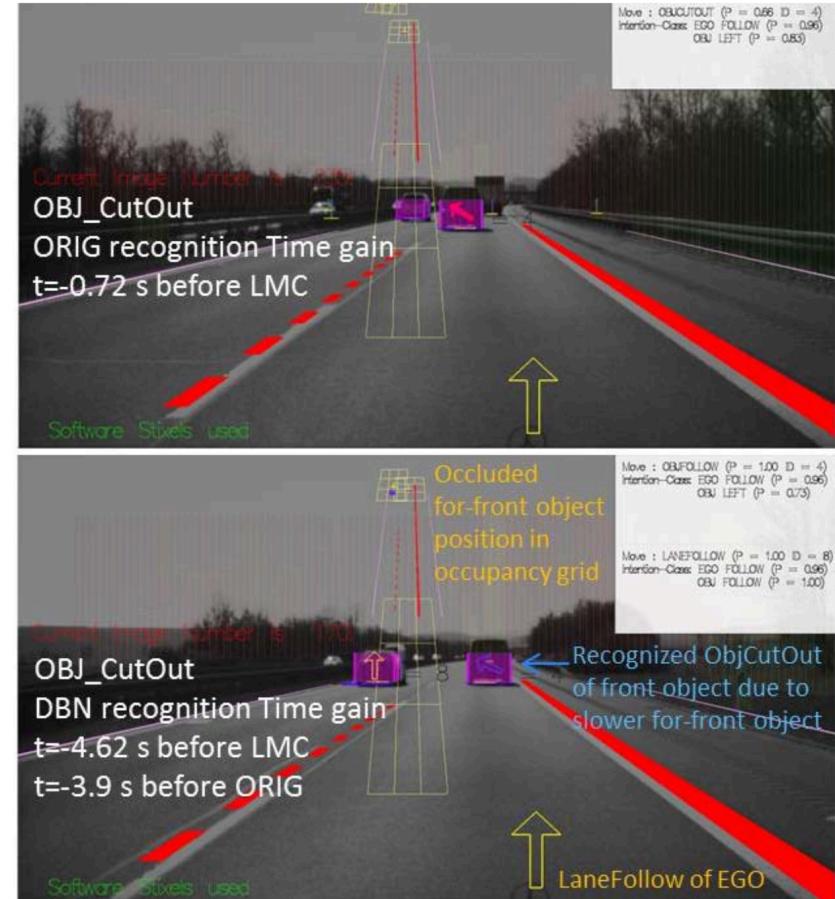
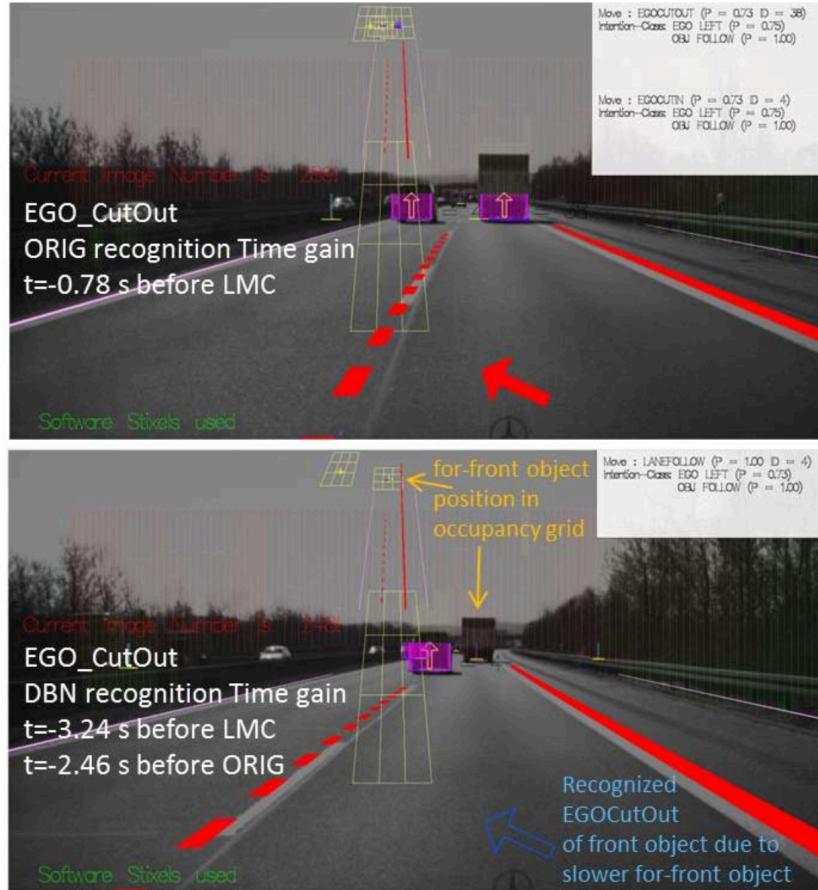
Early detection of traffic maneuvers changes for intelligent cruise control (and autonomous driving).

PROBABILISTIC MODEL



Weidl, Galia, et al. "Early Recognition of Maneuvers in Highway Traffic." *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer International Publishing, 2015.

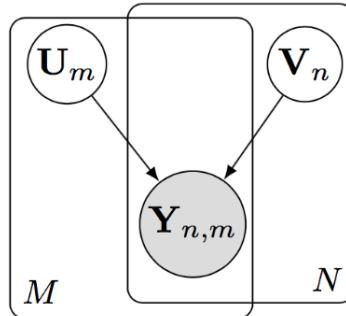
PROTOTYPE



Weidl, Galia, et al. "Early Recognition of Maneuvers in Highway Traffic." *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer International Publishing, 2015.

Frontiers in Probabilistic Machine Learning



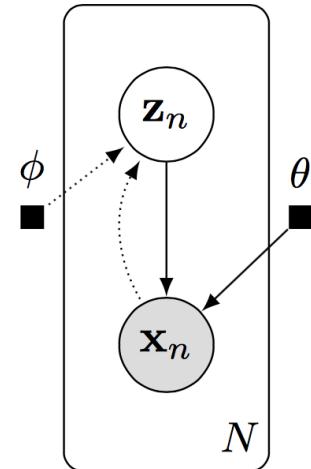
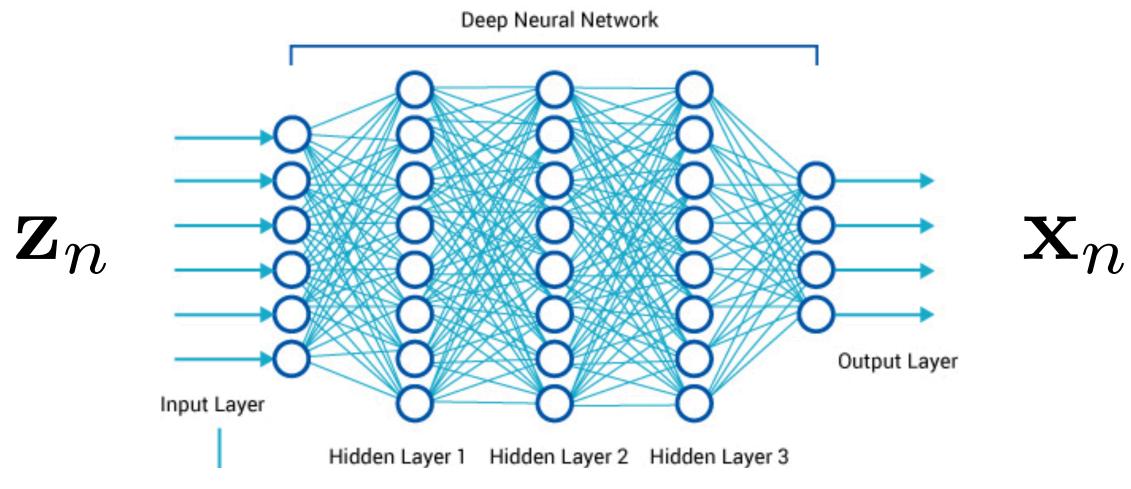


```
1 N = 10
2 M = 10
3 K = 5 # latent dimension
4
5 U = Normal(mu=tf.zeros([M, K]), sigma=tf.ones([M, K]))
6 V = Normal(mu=tf.zeros([N, K]), sigma=tf.ones([N, K]))
7 Y = Normal(mu=tf.matmul(U, V, transpose_b=True), sigma=tf.ones([N, M]))
```

Tran, Dustin, et al. "Edward: A library for probabilistic modeling, inference, and criticism." *arXiv preprint arXiv:1610.09787* (2016).

Probabilistic Programming Languages

- More powerful probabilistic modeling (e.g. Turing complete).
- Boost the productivity of data scientists.
- Expand the use of probabilistic modeling to non-experts.

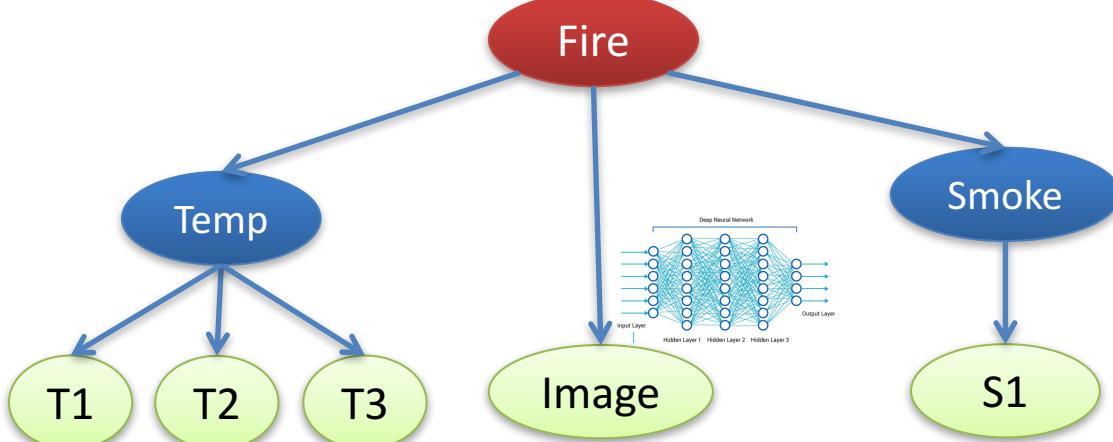


Deep Learning + Bayesian modeling

Powered by new advances in variational inference
(e.g. variational autoencoders, black-box variational inference, adversarial training, etc.).



Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).



$$p(Fire = True | t_1, t_2, t_3, s_1, image)$$

Probabilistic Programming on Tensorflow/Theano

Edward Library, PyMC3, AMIDST-II?

Thanks for your attention

www

www.amidsttoolbox.com

@

contact@amidsttoolbox.com



[@AmidstToolbox](https://twitter.com/AmidstToolbox)

AMIDST
→ TOOLBOX