

# Probabilistic Modelling of Electronic Health Records

**Francisco J. R. Ruiz**

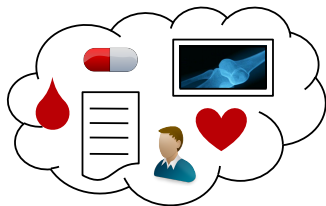
Marie Skłodowska-Curie Fellow

University of Cambridge & Columbia University

Brussels, 18 June 2019



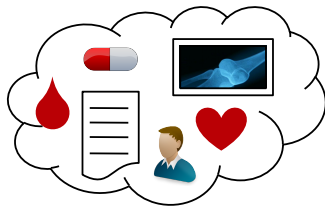
# Complex & Unstructured Datasets



Machine learning problems

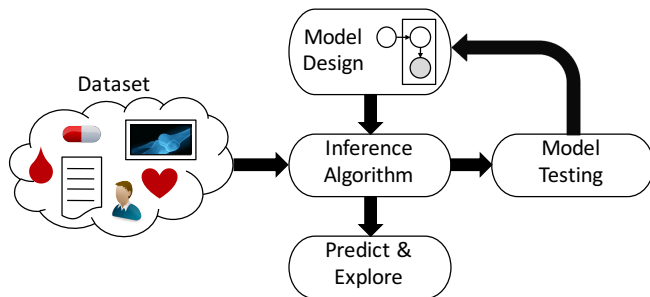
- ▶ Large unstructured datasets
- ▶ Goal: Make predictions, identify hidden patterns

# Probabilistic Machine Learning



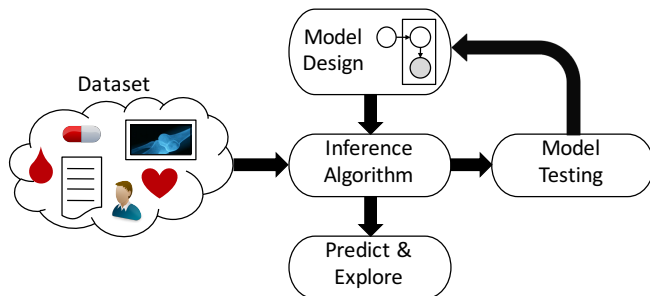
- ▶ Connect domain knowledge to data
- ▶ Uncertainty quantification
- ▶ Scalable computational tools

# Probabilistic Machine Learning



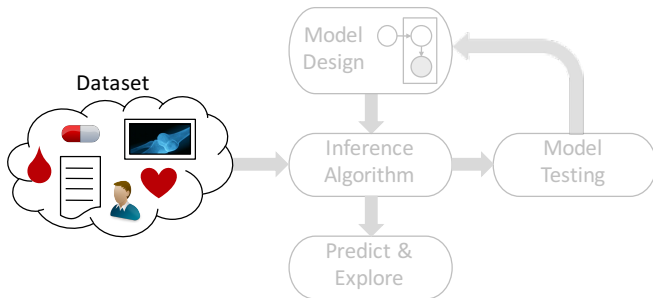
- ▶ Posit generative process with hidden and observed variables
- ▶ Given the data, reverse the process to infer hidden variables
- ▶ Use hidden structure to make predictions, explore the dataset, etc.

# Probabilistic Machine Learning



- ▶ Incorporate domain knowledge with interpretable components
- ▶ Separate assumptions from computation
- ▶ Facilitate collaboration with domain experts

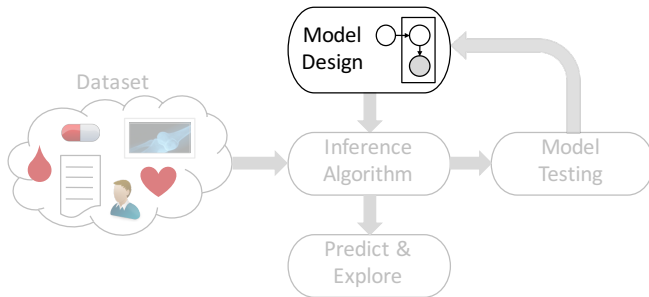
# EHR Data is Challenging to Analyze



# EHR Data is Challenging to Analyze

- ▶ Heterogeneous nature
- ▶ Noisy, missing observations
- ▶ Longitudinal
- ▶ Unobserved variables
- ▶ Large-scale datasets
- ▶ Observational
- ▶ ...

# Probabilistic Models



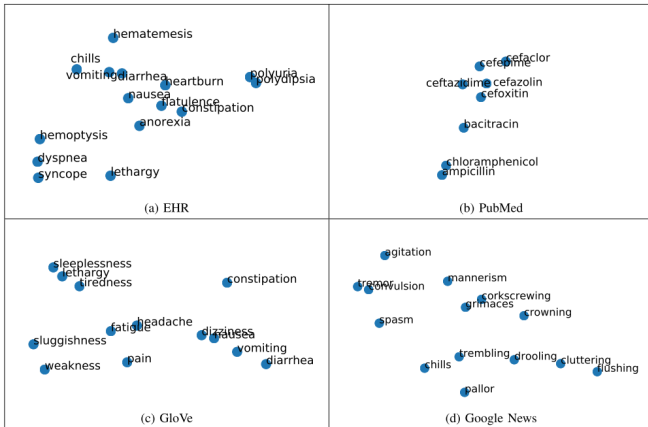


# Embedding Representations

- ▶ Find latent embedding representations of different data types
  - Medical conditions
  - Neural activity
  - Discrete or continuous data types
- ▶ Use the embeddings in a probabilistic model of heterogeneous data

# Word Embeddings

Word embeddings find distributed representations of individual words



(Image from Wang et al., 2018)

# Word Embeddings

- ▶ Word embeddings are a powerful tool for analyzing language
  - Words are placed in a low-dimensional latent space
  - Distances capture semantic similarity

*"she has had progressive difficulties with her breathing since she was admitted to hospital for respiratory failure"*

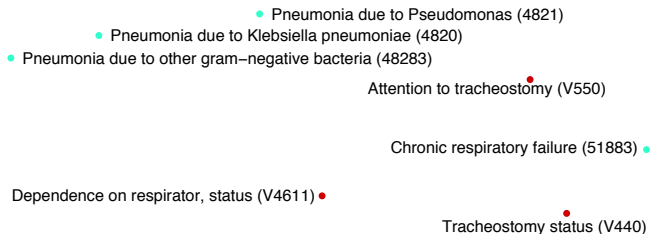
$$p(\text{breathing} \mid \text{context})$$

# Exponential Family Embeddings

- ▶ Can we find distributed features of other types of data?
  - Medical conditions
  - Neural activity
  - Discrete or continuous data types
- ▶ **Goals:**
  - Capture *data-to-data* interactions
  - Improve the *predictions* of matrix factorization
  - Obtain *features* that are useful for downstream analyses

# Exponential Family Embeddings

- ▶ Main idea: Distill the components of word embeddings
- ▶ Applications: Condition embeddings, neuron embeddings, ...
- ▶ Tools: Exponential families & Generalised linear models

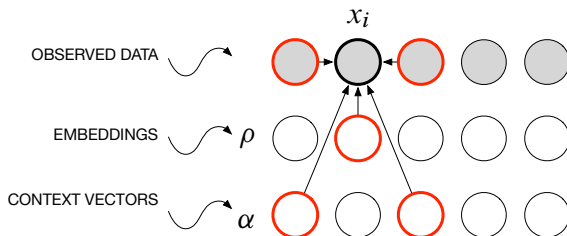


# Exponential Family Embeddings

► Observations  $x_i$ :

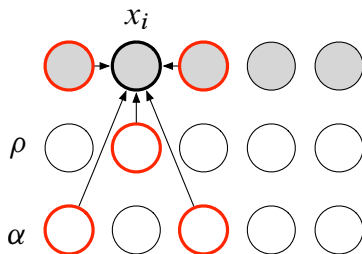
DOMAIN	INDEX	VALUE
Language	position in text $i$	word indicator
Medical conditions	condition and patient $(c, p)$	condition present
Neuroscience	neuron and time $(n, t)$	activity level

# Exponential Family Embeddings: Model Description



- ▶ Two latent vectors per data index (embedding, context)
- ▶ Model each data point conditioned on its context
- ▶ The latent variables interact in the conditional

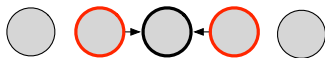
# Exponential Family Embeddings: Model Description



- ▶ Three ingredients:  
*context, conditional exponential family, embedding structure*

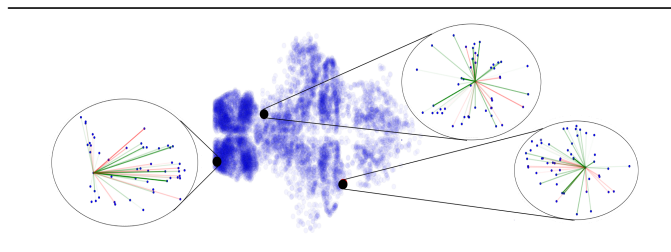


# Exponential Family Embeddings: Context



- ▶ Each data point  $i$  has a *context*  $\mathcal{C}_i$ , a set of indices
- ▶ The conditional of  $x_i$  depends on its context  $\mathcal{C}_i$

# Exponential Family Embeddings: Context



DOMAIN	DATA POINT	CONTEXT
Language	word	surrounding words
Medical conditions	condition presence	other conditions of patient
Neuroscience	neuron activity	surrounding neurons

# Exponential Family Embeddings: Exponential Family

- ▶ Use exponential families for the conditional of each data point,

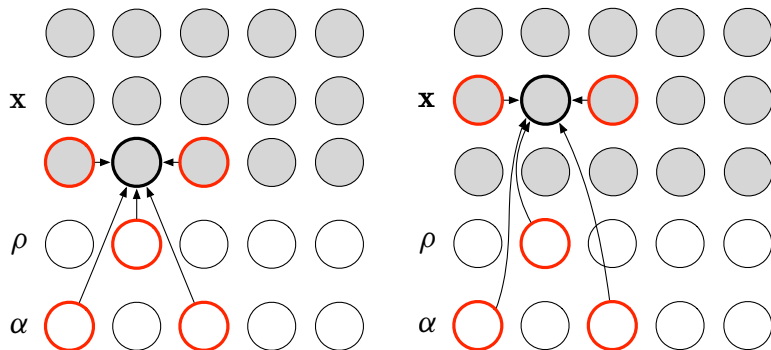
$$p(x_i | \mathbf{x}_{\mathcal{C}_i}) = \text{EXPFAM}(\eta_i(\mathbf{x}_{\mathcal{C}_i}), t(x_i))$$

- ▶ The natural parameter combines the embedding and context vectors,

$$\eta_i(\mathbf{x}_{\mathcal{C}_i}) = f_i \left( \rho[i]^\top \sum_{j \in \mathcal{C}_i} \alpha[j] x_j \right)$$

- $f_i(\cdot)$ : Link function (identity, log, ...)

# Exponential Family Embeddings: Embedding Structure



- ▶ The embedding structure determines how parameters are shared
- ▶  $\rho[i] = \rho[j]$  for  $i = (\text{pneumonia}, p)$  and  $j = (\text{pneumonia}, p')$

# Exponential Family Embeddings: Objective Function

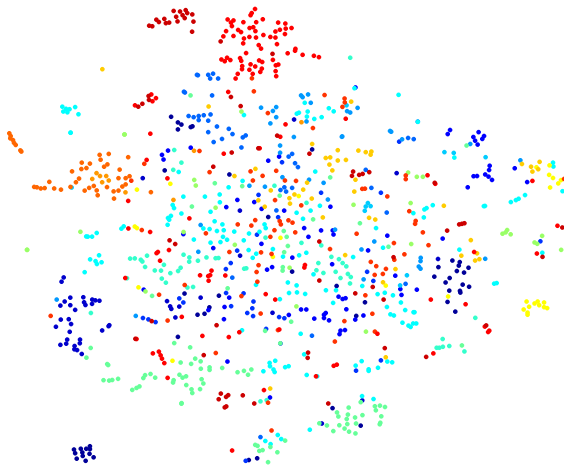
- ▶ Model each datapoint conditioned on its context
- ▶ Combine these terms in a pseudolikelihood

$$\mathcal{L}(\rho, \alpha) = \sum_i \left( \eta_i^\top t(x_i) - a(\eta_i) \right) + \mathcal{L}^{(\text{reg})}$$

- ▶ Fit the embedding by maximising  $\mathcal{L}(\rho, \alpha)$

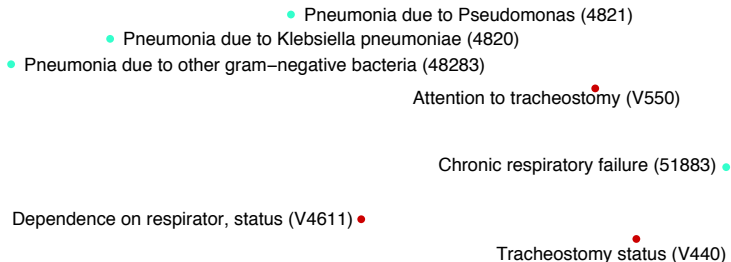
# Exponential Family Embeddings: Results

Results on MIMIC-III dataset



# Exponential Family Embeddings: Results

## Results on MIMIC-III dataset (zoom)



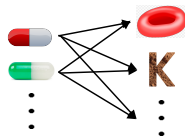
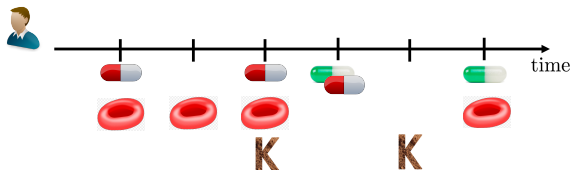
# Exponential Family Embeddings: Results and Extensions

- ▶ The fitted embeddings are interpretable
- ▶ The fitted embeddings have good predictive performance
- ▶ Extensions of EFEs:
  - Combine with more complex probabilistic models
  - Structured embeddings
  - Blend with causality ideas



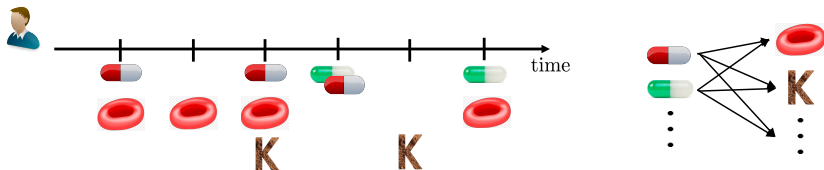
# Modelling Heterogeneous Longitudinal Data

- Data: Events in time (drugs, labs, conditions)



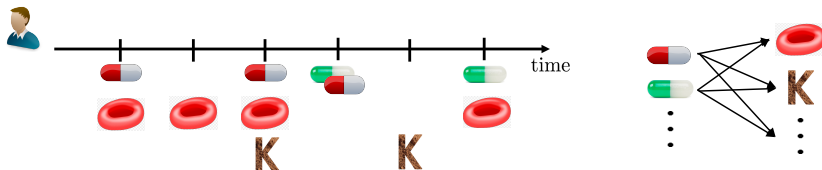
# Modelling Heterogeneous Longitudinal Data

- Effect of drugs on lab tests



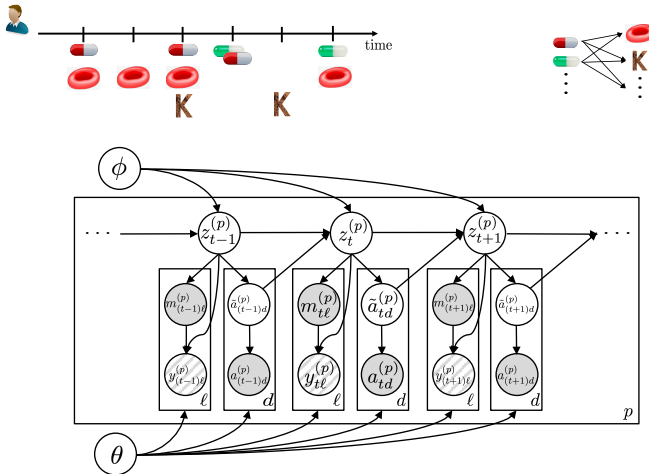
# Modelling Heterogeneous Longitudinal Data

- ▶ Learn from measurement values but also from their presence



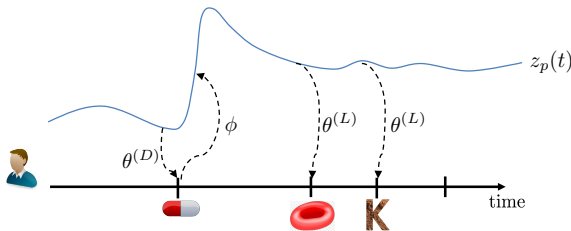
# Modelling Heterogeneous Longitudinal Data

- A first model: Linear dynamical system



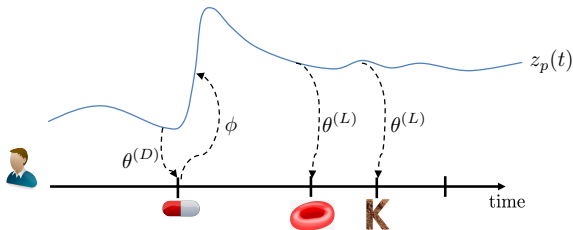
# Modelling Heterogeneous Longitudinal Data

- ▶ Continuous extension
- ▶ Extract information from time stamps and event types
- ▶ Model: temporal point processes with per-patient latent rate

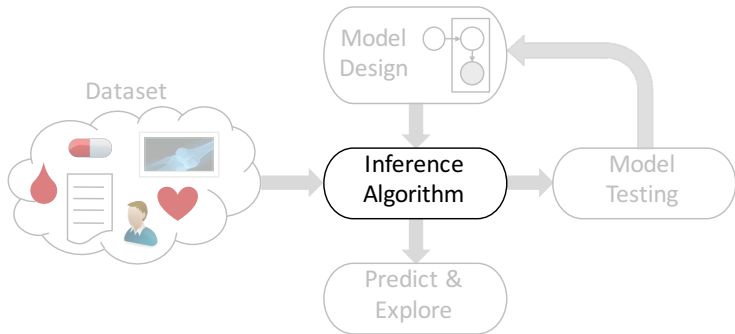


# Modelling Heterogeneous Longitudinal Data

- ▶ The embeddings appear in the likelihood
- ▶ Pre-fitted embedding, or latent variables in the model



# Scalable Inference



# Scalable Inference

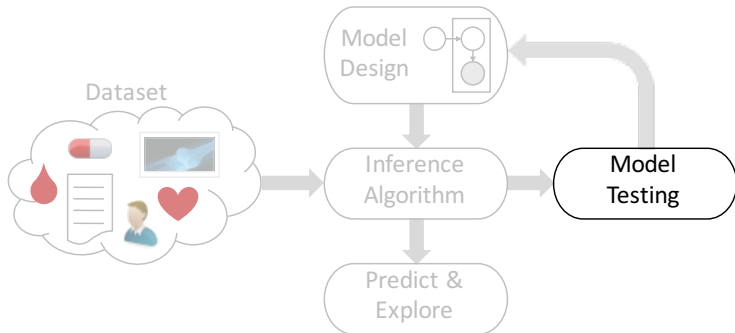
- ▶ Approximate the posterior of the latent variables given data
- ▶ Variational inference
- ▶ Scalable to:
  - Complex, non-conjugate models
  - Large numbers of observations
  - High-dimensional latent variables



# Contributions on Variational Inference

- ▶ Scale-up inference for categorical observations with many outcomes  
[Ruiz+, ICML 2018]
- ▶ Reduce variance of gradient estimators  
[Ruiz+, NeurIPS 2016] [Naesseth+, AISTATS 2017]
- ▶ More expressive variational families  
[Titsias+, AISTATS 2019] [Ruiz+, ICML 2019]

# Model Testing



# Model Testing

- ▶ Context: Probabilistic modelling for causal inference
- ▶ Problems:
  - Causal inference requires many assumptions
  - Standard Bayesian testing tools are not applicable
- ▶ Contribution: A method to check the validity of the assumptions

# Conclusions

- ▶ Probabilistic modelling is a powerful tool for analyzing EHR data
- ▶ Contributions on
  - **Modelling:** Exponential family embeddings, point process latent health model, ...
  - **Variational inference:** Scale up categoricals, reduce gradient variance, more expressive variational families
  - **Testing:** Bayesian causal model testing



EU H2020 (MSCA Actions,  
Grant Agreement 706760)