# CPSC 330 Lecture 20: Survival analysis

Firas Moosvi

# Announcements

- Final Exam reservations are now open!
- Midterm 2 grading is underway.
- Exam Review Sessions (MT1 and MT2) will happen later this week.
- Wednesdays's class will likely end by 12 PM, I'll stick around for questions for a (long) while.

# iClicker Exercise 19.1

Select all of the following statements which are TRUE.

- a. It’s possible to use word2vec embedding representations for text classification instead of bag-of-words representation.
- b. The topic model approach we used in the last lecture, Latent Dirichlet Allocation (LDA), is an unsupervised approach.
- c. In an LDA topic model, the same word can be associated with two different topics with high probability.
- d. In an LDA topic model, a document is a mixture of multiple topics.
- e. If I train a topic model on a large collection of news articles with K = 10, I would get 10 topic labels (e.g., sports, culture, politics, finance) as output.

# Recap: Time Series Analysis

# iClicker 20.1

iClicker cloud join link: **https://join.iclicker.com/YJHS**

**Select all of the following statements which are TRUE.**

- a. We need to be careful when splitting the data when working with time series data.
- b. Cross-validation in time series can be randomly applied like in other machine learning tasks.
- c. In time series forecasting, the future value of a series can only be predicted based on its past values and cannot incorporate other variables.
- d. When we used RandomForestRegressor model on the POSIX time feature, it predicted a straight line on the test data because tree-based models are inherently unable to extrapolate (i.e., make predictions outside the - range of the training data).

# Introduction to Survival Analysis

# Group Work: Class Demo & Live Coding

For this demo, each student should click this link to create a new repo in their accounts, then clone that repo locally to follow along with the demo from today.

# iClicker 20.2

iClicker cloud join link: **https://join.iclicker.com/YJHS**

**Select all of the following statements which are TRUE.**

- a. Right censoring occurs when the endpoint of event has not been observed for all study subjects by the end of the study period.
- b. Right censoring implies that the data is missing completely at random.
- c. In the presence of right-censored data, binary classification models can be applied directly without any modifications or special considerations.
- d. If we apply the Ridge regression model to predict tenure in right censored data, we are likely to underestimate it because the tenure observed in our data is shorter than what it would be in reality.